SPRINGER REFERENCE

Miodrag Lovric
Editor

International Encyclopedia of Statistical Science

Springer

A

Absolute Penalty Estimation

Ejaz S. Ahmed[1], Enayetur Raheem[2], Shakhawat Hossain[2]
[1]Professor and Department Head of Mathematics and Statistics
University of Windsor, Windsor, ON, Canada
[2]University of Windsor, Windsor, ON, Canada

In statistics, the technique of ▶least squares is used for estimating the unknown parameters in a linear regression model (see ▶Linear Regression Models). This method minimizes the sum of squared distances between the observed responses in a set of data, and the fitted responses from the regression model. Suppose we observe a collection of data $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$ on n units, where y_is are responses and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is a vector of predictors. It is convenient to write the model in matrix notation, as,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon, \tag{1}$$

where $\boldsymbol{y}$ is $n \times 1$ vector of responses, $\boldsymbol{X}$ is $n \times p$ matrix, known as the design matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the unknown parameter vector and ε is the vector of random errors. In ordinary least squares (OLS) regression, we estimate $\boldsymbol{\beta}$ by minimizing the residual sum of squares, $RSS = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, giving $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. This estimator is simple and has some good statistical properties. However, the estimator suffers from lack of uniqueness if the design matrix $\boldsymbol{X}$ is less than full rank, and if the columns of $\boldsymbol{X}$ are (nearly) collinear. To achieve better prediction and to alleviate ill conditioning problem of $\boldsymbol{X}^T\boldsymbol{X}$, Hoerl and Kernard (1970) introduced ridge regression (see ▶Ridge and Surrogate Ridge Regressions), which minimizes the RSS subject to a constraint, $\sum \beta_j^2 \le t$, in other words

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{2}$$

where $\lambda \ge 0$ is known as the complexity parameter that controls the amount of shrinkage. The larger the value of λ, the greater the amount of shrinkage. The quadratic penalty term makes $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ a linear function of $\boldsymbol{y}$. Frank and Friedman (1993) introduced bridge regression, a generalized version of penalty (or absolute penalty type) estimation, which includes ridge regression when $\gamma = 2$. For a given penalty function $\pi(\cdot)$ and regularization parameter λ, the general form can be written as

$$\phi(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\pi(\boldsymbol{\beta}),$$

where the penalty function is of the form

$$\pi(\beta) = \sum_{j=1}^{p} |\beta_j|^\gamma, \ \gamma > 0. \tag{3}$$

The penalty function in (3) bounds the L_γ norm of the parameters in the given model as $\sum_{j=1}^{m} |\beta_j|^\gamma \le t$, where t is the tuning parameter that controls the amount of shrinkage. We see that for $\gamma = 2$, we obtain ridge regression. However, if $\gamma \ne 2$, the penalty function will not be rotationally invariant. Interestingly, for $\gamma < 2$, it shrinks the coefficient toward zero, and depending on the value of λ, it sets some of them to be exactly zero. Thus, the procedure combines variable selection and shrinkage of coefficients of penalized regression. An important member of the penalized least squares (PLS) family is the L_1 penalized least squares estimator or the *lasso* [*least absolute shrinkage and selection operator*, Tibshirani (1996)]. In other words, the *absolute penalty estimator* (APE) arises when the absolute value of penalty term is considered, i.e., $\gamma = 1$ in (3). Similar to the ridge regression, the lasso estimates are obtained as

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{4}$$

The lasso shrinks the OLS estimator toward zero and depending on the value of λ, it sets some coefficients to exactly zero. Tibshirani (1996) used a quadratic programming method to solve (4) for $\hat{\boldsymbol{\beta}}^{\text{lasso}}$. Later, Efron et al. (2004) proposed least angle regression (LAR), a type of stepwise regression, with which the

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

lasso estimates can be obtained at the same computational cost as that of an ordinary least squares estimation Hastie et al. (2009). Further, the lasso estimator remains numerically feasible for dimensions m that are much higher than the sample size n. Zou and Hastie (2005) introduced a hybrid PLS regression with the so called *elastic net penalty* defined as $\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)\,|\beta_j|)$. Here the penalty function is a linear combination of the ridge regression penalty function and lasso penalty function. A different type of PLS, called *garotte* is due to Breiman (1993). Further, PLS estimation provides a generalization of both nonparametric least squares and weighted projection estimators, and a popular version of the PLS is given by Tikhonov regularization (Tikhonov 1963). Generally speaking, the ridge regression is highly efficient and stable when there are many small coefficients. The performance of lasso is superior when there are a small-to-medium number of moderate-sized coefficients. On the other hand, shrinkage estimators perform well when there are large known zero coefficients.

Ahmed et al. (2007) proposed an APE for partially linear models. Further, they reappraised the properties of shrinkage estimators based on Stein-rule estimation. There exists a whole family of estimators that are better than OLS estimators in regression models when the number of predictors is large. A partially linear regression model is defined as

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + g(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{5}$$

where $t_i \in [0,1]$ are design points, $g(\cdot)$ is an unknown real-valued function defined on $[0,1]$, and y_i, $\boldsymbol{x}$, $\boldsymbol{\beta}$, and ε_i's are as defined in the context of (1). We consider experiments where the vector of coefficients $\boldsymbol{\beta}$ in the linear part of (5) can be partitioned as $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1$ is the coefficient vector of order $p_1 \times 1$ for main effects (e.g., treatment effects, genetic effects) and $\boldsymbol{\beta}_2$ is a vector of order $p_2 \times 1$ for "nuisance" effects (e.g., age, laboratory). Our relevant hypothesis is $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. Let $\hat{\boldsymbol{\beta}}_1$ be a semiparametric least squares estimator of $\boldsymbol{\beta}_1$, and we let $\tilde{\boldsymbol{\beta}}_1$ denote the restricted semiparametric least squares estimator of $\boldsymbol{\beta}_1$. Then the semiparametric Stein-type estimator (see ►James-Stein Estimator and Semiparametric Regression Models), $\hat{\boldsymbol{\beta}}_1^S$, of $\boldsymbol{\beta}_1$ is

$$\hat{\boldsymbol{\beta}}_1^S = \tilde{\boldsymbol{\beta}}_1 + \{1 - (p_2 - 2)T^{-1}\}(\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1), \quad p_2 \geq 3 \tag{6}$$

where T is an appropriate test statistic for the H_0. A positive-rule shrinkage estimator (PSE) $\hat{\boldsymbol{\beta}}_1^{S+}$ is defined as

$$\hat{\boldsymbol{\beta}}_1^{S+} = \tilde{\boldsymbol{\beta}}_1 + \{1 - (p_2 - 2)T^{-1}\}^+(\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1), \quad p_2 \geq 3 \tag{7}$$

where $z^+ = max(0, z)$. The PSE is particularly important to control the over-shrinking inherent in $\hat{\boldsymbol{\beta}}_1^S$. The shrinkage estimators can be viewed as a competitor to the APE approach. Ahmed et al. (2007) finds that, when p_2 is relatively small with respect to p, APE performs better than the shrinkage method. On the other hand, the shrinkage method performs better when p_2 is large, which is consistent with the performance of the APE in linear models. Importantly, the shrinkage approach is free from any tuning parameters, easy to compute and calculations are not iterative. The shrinkage estimation strategy can be extended in various directions to more complex problems. It may be worth mentioning that this is one of the two areas Bradley Efron predicted for the early twenty-first century (RSS News, January 1995). Shrinkage and likelihood-based methods continue to be extremely useful tools for efficient estimation.

About the Author

The author S. Ejaz Ahmed is Professor and Head Department of Mathematics and Statistics. For biography, *see* entry ►Optimal Shrinkage Estimation.

Cross References

►Estimation
►Estimation: An Overview
►James-Stein Estimator
►Linear Regression Models
►Optimal Shrinkage Estimation
►Residuals
►Ridge and Surrogate Ridge Regressions
►Semiparametric Regression Models

References and Further Reading

Ahmed SE, Doksum KA, Hossain S, You J (2007) Shrinkage, pretest and absolute penalty estimators in partially linear models. Aust NZ J Stat 49(4):435–454

Breiman L (1993) Better subset selection using the non-negative garotte. Technical report, University of California, Berkeley

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression (with discussion). Ann Stat 32(2):407–499

Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35:109–148

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288

Tikhonov An (1963) Solution of incorrectly formulated problems and the regularization method. Soviet Math Dokl 4:1035–1038, English translation of Dokl Akad Nauk SSSR 151, 1963, 501–504

Zou H, Hastie T (2005) Regularization and variable selction via the elastic net. J R Stat Soc B 67(2):301–320

Accelerated Lifetime Testing

Francisco Louzada-Neto
Associate Professor
Universidade Federal de São Carlos, Sao Paulo, Brazil

Accelerated life tests (ALT) are efficient industrial experiments for obtaining measures of a device reliability under the usual working conditions.

A practical problem for industries of different areas is to obtain measures of a device reliability under its usual working conditions. Typically, the time and cost of such experimentation are long and expensive. The ALT are efficient for handling such situation, since the information on the device performance under the usual working conditions are obtained by considering a time and cost-reduced experimental scheme. The ALT are performed by testing items at higher stress covariate levels than the usual working conditions, such as temperature, pressure and voltage.

There is a large literature on ALT and interested readers can refer to Mann et al. (1974), Nelson (1990), Meeker and Escobar (1998) which are excellent sources for ALT. Nelson (2005a, b) provides a brief background on accelerated testing and test plans and surveys the related literature point out more than 150 related references.

A simple ALT scenario is characterized by putting k groups of n_i items each under constant and fixed stress covariate levels, X_i (hereafter stress level), for $i = 1, \ldots, k$, where $i = 1$ generally denotes the usual stress level, that is, the usual working conditions. The experiment ends after a certain pre-fixed number $r_i < n_i$ of failures, $t_{i1}, t_{i2}, \ldots, t_{ir_i}$, at each stress level, characterizing a type II censoring scheme (Lawless 2003; see also ▶Censoring Methodology). Other stress schemes, such as step (see ▶Step-Stress Accelerated Life Tests) and progressive ones, are also common in practice but will not be considered here. Examples of those more sophisticated stress schemes can be found in Nelson (1990).

The ALT models are composed by two components. One is a probabilistic component, which is represented by a lifetime distribution, such as exponential, Weibull, log-normal, log-logistic, among others. The other is a stress-response relationship (SRR), which relates the mean lifetime (or a function of this parameter) with the stress levels. Common SRRs are the power law, Eyring and Arrhenius models (Meeker and Escobar 1998) or even a general log-linear or log-non-linear SRR which encompass the formers. For sake of illustration, we shall assume an exponential distribution as the lifetime model and a general log-linear SRR. Here, the mean lifetime under the usual working conditions shall represent our device reliability measure of interesting.

Let $T > 0$ be the lifetime random variable with an exponential density

$$f(t, \lambda_i) = \lambda_i \exp\{-\lambda_i t\}, \tag{1}$$

where $\lambda_i > 0$ is an unknown parameter representing the constant failure rate for $i = 1, \ldots, k$ (number of stress levels). The mean lifetime is given by $\theta_i = 1/\lambda_i$.

The likelihood function for λ_i, under the i-th stress level X_i, is given by

$$L_i(\lambda_i) = \left(\prod_{j=1}^{r_i} f(t_{ij}, \lambda_i)\right) (S(t_{ir_i}, \lambda_i))^{n_i - r_i} = \lambda_i^{r_i} \exp\{-\lambda_i A_i\},$$

where $S(t_{ir_i}, \lambda_i)$ is the survival function at t_{ir_i} and $A_i = \sum_{j=1}^{r_i} t_{ij} + (n_i - r_i) t_{ir_i}$ denotes the total time on test for the i-th stress level.

Considering data under the k random stress levels, the likelihood function for the parameter vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ is given by

$$L(\lambda) = \prod_{i=1}^{k} \lambda_i^{r_i} \exp\{-\lambda_i A_i\}. \tag{2}$$

We consider a general log-linear SRR defined as

$$\lambda_i = \exp(-Z_i - \beta_0 - \beta_1 X_i), \tag{3}$$

where X is the covariate, $Z = g(X)$ and β_0 and β_1 are unknown parameters such that $-\infty < \beta_0, \beta_1 < \infty$.

The SRR (3) has several models as particular cases. The Arrhenius model is obtained if $Z_i = 0$, $Xi = 1/Vi$, $\beta_0 = -\alpha_1$ and $\beta_1 = \alpha_2$, where Vi denotes a level of the temperature variable. If $Z_i = 0$, $Xi = -log(Vi)$, $\beta_0 = log(\alpha)$ and $\beta_1 = \alpha_2$, where Vi denotes a level of the voltage variable we obtain the power model. Following Louzada-Neto and Pardo-Fernandéz (2001), the Eyring model is obtained if $Z_i = -\log V_i$, $Xi = 1/Vi$, $\beta_0 = -\alpha_1$ and $\beta_1 = \alpha_2$, where Vi denotes a level of the temperature variable. Interested readers can refer to Meeker and Escobar (1998) for more information about the physical models considered here.

From (2) and (3), the likelihood function for β_0 and β_1 is given by

$$L(\beta_0, \beta_1) = \prod_{i=1}^{k} \{\exp(-Z_i - \beta_0 - \beta_1 X_i)^{r_i} \exp(-\exp(-Z_i - \beta_0 - \beta_1 X_i) A_i)\}. \quad (4)$$

The maximum likelihood estimates (MLEs) of β_0 and β_1 can be obtained by direct maximization of (4), or by solving the system of nonlinear equations, $\partial \log L / \partial \theta = 0$, where $\theta' = (\beta_0, \beta_1)$. Obtaining the score function is conceptually simple and the expressions are not given explicitly. The MLEs of θ_i can be obtained, in principle, straightforwardly by considering the invariance property of the MLEs.

Large-sample inference for the parameters can be based on the MLEs and their estimated variances, obtained by inverting the expected information matrix (Cox and Hinkley 1974). For small or moderate-sized samples however we may consider simulation approaches, such as the bootstrap confidence intervals (see ►Bootstrap Methods) that are based on the empirical evidence and are therefore preferred (Davison and Hinkley 1997). Formal goodness-of-fit tests are also feasible since, from (3), we can use the likelihood ratio statistics (LRS) for testing goodness-of-fit of hypotheses such as $H_0 : \beta_1 = 0$.

Although we considered only an exponential distribution as our lifetime model, more general lifetime distributions, such as the Weibull (see ►Weibull Distribution and Generalized Weibull Distributions), log-normal, log-logistic, among others, could be considered in principle. However, the degree of difficulty in the calculations increase considerably. Also we considered only one stress covariate, however this is not critical for the overall approach to hold and the multiple covariate case can be handle straightforwardly.

A study on the effect of different reparametrizations on the accuracy of inferences for ALT is discussed in Louzada-Neto and Pardo-Fernandéz 2001). Modeling ALT with a log-non-linear SRR can be found in Perdoná et al. (2004). Modeling ALT with a threshold stress, below which the lifetime of a product can be considered to be infinity or much higher than that for which it has been developed is proposed by Tojeiro et al. (2004).

We only considered ALT in presence of constant stress loading, however non-constant stress loading, such as step stress and linearly increasing stress are provided by Miller and Nelson (1983) and Bai, Cha and Chung (1992), respectively. A comparison between constant and step stress tests is provided by Khamis (1997). A log-logistic step stress model is provided by Srivastava and Shukla (2008).

Two types of software for ALT are provided by Meeker and Escobar (2002) and ReliaSoft Corporation (2004).

About the Author

Francisco Louzada-Neto is an associate professor of Statistics at Universidade Federal de São Carlos (UFSCar), Brazil. He received his Ph.D in Statistics from University of Oxford (England). He is Director of the Centre for Hazard Studies (2004–2010, UFSCar, Brazil) and Editor in Chief of the *Brazilian Journal of Statistics* (2004–2010, Brazil). He is a past-Director for Undergraduate Studies (1992–1994, UFSCar, Brazil) and was Director for Graduate Studies in Statistics (1999–2008, UFSCar, Brazil). Louzada-Neto is single and joint author of more than 100 publications in statistical peer reviewed journals, books and book chapters, He has supervised more than 50 assistant researches, Ph.Ds, masters and undergraduates.

Cross References

►Degradation Models in Reliability and Survival Analysis
►Modeling Survival Data
►Step-Stress Accelerated Life Tests
►Survival Data

References and Further Reading

Bai DS, Cha MS, Chung SW (1992) Optimum simple ramp tests for the Weibull distribution and type-I censoring. IEEE T Reliab 41:407–413

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Khamis IH (1997) Comparison between constant- and step-stress tests for Weibull models. Int J Qual Reliab Manag 14:74–81

Lawless JF (2003) Statistical models and methods for lifetime data, 2nd end. Wiley, New York

Louzada-Neto F, Pardo-Fernandéz JC (2001) The effect of reparametrization on the accuracy of inferences for accelerated lifetime tests. J Appl Stat 28:703–711

Mann NR, Schaffer RE, Singpurwalla ND (1974) Methods for statistical analysis of reliability and life test data. Wiley, New York

Meeker WQ, Escobar LA (1998) Statistical methods for reliability data. Wiley, New York

Meeker WQ, Escobar LA (2002) SPLIDA (S-PLUS Life Data Analysis) software–graphical user interface. http://www.public.iastate.edu/~splida

Miller R, Nelson WB (1983) Optimum simple step-stress plans for accelerated life testing. IEEE T Reliab 32:59–65

Nelson W (1990) Accelerated testing – statistical models, test plans, and data analyses. Wiley, New York

Nelson W (2005a) A bibliography of accelerated test plans. IEEE T Reliab 54:194–197

Nelson W (2005b) A bibliography of accelerated test plans part II – references. IEEE T Reliab 54:370–373

Perdoná GSC, Louzada Neto F, Tojeiro CAV (2004) Bayesian modelling of log-non-linear stress-response relationships in accelerated lifetime tests. J Stat Theory Appl 3(1):5–12

Reliasoft Corporation (2004) Optimum allocations of stress levels and test units in accelerated tests. Reliab EDGE 5:10–17. http://www.reliasoft.com

Srivastava PW, Shukla R (2008) A log-logistic step-stress model. IEEE T Reliab 57:431–434

Tojeiro CAV, Louzada Neto F, Bolfarine H (2004) A Bayesian analysis for accelerated lifetime tests under an exponential power law model with threshold stress. J Appl Stat 31(6):685–691

Acceptance Sampling

M. Ivette Gomes
Professor
Universidade de Lisboa, DEIO and CEAUL, Lisboa, Portugal

Introduction

Acceptance sampling (AS) is one of the oldest statistical techniques in the area of ▶statistical quality control. It is performed out of the line production, most commonly before it, for deciding on incoming batches, but also after it, for evaluating the final product (see Duncan 1986; Stephens 2001; Pandey 2007; Montgomery 2009; and Schilling and Neubauer 2009, among others). Accepted batches go into the production line or are sold to consumers; the rejected ones are usually submitted to a *rectification process*. A *sampling plan* is defined by the *size of the sample* (samples) taken from the batch and by the associated *acceptance–rejection* criterion. The most widely used plans are given by the Military Standard tables, developed during the *World War* II, and first issued in 1950. We mention MIL STD 105E (1989) and the civil version ANSI/ASQC Z1.9 (1993) of the *American National Standards Institution* and the *American Society for Quality Control*.

At the beginning, all items and products were inspected for the identification of nonconformities. At the late 20s, Dodge and Romig (see Dodge and Romig 1959), in the Bell Laboratories, developed the area of AS, as an alternative to 100% inspection. The aim of AS is to lead producers to a decision (acceptance or rejection of a batch) and not to the estimation or improvement of the quality of a batch. Consequently, AS does not provide a direct form of *quality control*, but its indirect effects in *quality* are important: if a batch is rejected, either the supplier tries improving its production methods or the consumer (producer) looks for a better supplier, indirectly increasing quality.

Regarding the decision on the batches, we distinguish three different approaches: (1) *acceptance without inspection*, applied when the supplier is highly reliable; (2) 100% *inspection*, which is expensive and can lead to a sloppy attitude towards quality; (3) *an intermediate decision*, i.e., an *acceptance sampling program*. This increases the interest on quality and leads to the lemma: *make things right in the first place*. The type of inspection that should be applied depends on the quality of the last batches inspected. At the beginning of inspection, a so-called *normal inspection* is used, but there are two other types of inspection, a *tightened inspection* (for a history of low quality), and a *reduced inspection* (for a history of high quality). There are special and empirical switching rules between the three types of inspection, as well as for discontinuation of inspection.

Factors for Classifications of Sampling Plans

Sampling plans by attributes versus sampling plans by variables. If the item inspection leads to a binary result (conforming or nonconforming), we are dealing with *sampling by attributes*, detailed later on. If the item inspection leads to a continuous measurement X, we are *sampling by variables*. Then, we generally use sampling plans based on the sample mean and standard deviation, the so-called *variable sampling plans*. If X is normal, it is easy to compute the number of items to be collected and the criteria that leads to the rejection of the batch, with chosen risks α and β. For different *sampling plans by variables*, see Duncan (1986), among others.

Incoming versus outgoing inspection. If the batches are inspected before the product is sent to the consumer, it is called *outgoing inspection*. If the inspection is done by the consumer (producer), after they were received from the supplier, it is called *incoming inspection*.

Rectifying versus non-rectifying sampling plans. All depends on what is done with nonconforming items that were found during the inspection. When the cost of replacing faulty items with new ones, or reworking them is accounted for, the sampling plan is rectifying.

Single, double, multiple and sequential sampling plans.

- **Single sampling**. This is the most common sampling plan: we draw a random sample of n items from the batch, and count the number of nonconforming items (or the number of nonconformities, if more than one nonconformity is possible on a single item). Such a

plan is defined by n and by an associated *acceptance-rejection* criterion, usually a value c, the so-called *acceptance number*, the number of nonconforming items that cannot be exceeded. If the number of nonconforming items is greater than c, the batch is rejected; otherwise, the batch is accepted. The number r, defined as the minimum number of nonconforming items leading to the rejection of the batch, is the so-called *rejection number*. In the most simple case, as above, $r = c + 1$, but we can have $r > c + 1$.

- **Double sampling**. A *double sampling plan* is characterized by four parameters: $n_1 << n$, the size of the first sample, c_1 the acceptance number for the first sample, n_2 the size of the second sample and c_2 ($> c_1$) the acceptance number for the joint sample. The main advantage of a double sampling plan is the reduction of the total inspection and associated cost, particularly if we proceed to a *curtailment* in the second sample, i.e. we stop the inspection whenever c_2 is exceeded. Another (psychological) advantage of these plans is the way they give a second opportunity to the batch.
- **Multiple sampling**. In the *multiple plans* a predetermined number of samples are drawn before taking a decision.
- ►**Sequential sampling**. The *sequential plans* are a generalization of multiple plans. The main difference is that the number of samples is not pre-determined. If, at each step, we draw a sample of size *one*, the plan, based on Wald's test, is called *sequential item-to-item*; otherwise, it is *sequential by groups*. For a full study of multiple and sequential plans see, for instance, Duncan (1986) (see also the entry ►Sequential Sampling).

Special sampling plans. Among the great variety of special plans, we distinguish:

- **Chain sampling**. When the inspection procedures are destructive or very expensive, a small n is recommendable. We are then led to acceptance numbers equal to zero. This is dangerous for the supplier and if rectifying inspection is used, it is expensive for the consumer. In 1955, Dodge suggested a procedure alternative to this type of plans, which uses also the information of preceding batches, the so-called *chain sampling method* (see Dogdge and Romig 1959).
- **Continuous sampling plans** (CSP). There are continuous production processes, where the raw material is not naturally provided in batches. For this type of production it is common to alternate sequences of sampling inspection with 100% inspection – they are in a certain sense rectifying plans. The simplest plan of this type, the CSP-1, was suggested in 1943 by Dodge. It begins with a 100% inspection. When a pre-specified number i of consecutive nonconforming items is achieved, the plan changes into sampling inspection, with the inspection of f items, randomly selected, along the continuous production. If *one* nonconforming item is detected (the reason for the terminology CSP-1), 100% inspection comes again, and the nonconforming item is replaced. For properties of this plan and its generalizations see Duncan (1986).

A Few Characteristics of a Sampling Plan

OCC. The *operational characteristic curve* (OCC) is $P_a \equiv P_a(p) = \mathbb{P}(\text{acceptance of the batch} \mid p)$, where p is the probability of a nonconforming item in the batch.

AQL and LTPD (or RQL). The sampling plans are built taken into account the wishes of both the supplier and the consumer, defining two quality levels for the judgment of the batches: the *acceptance quality level* (AQL), the worst operating quality of the process which leads to a high probability of acceptance of the batch, usually 95% – for the protection of the supplier regarding high quality batches, and the *lot tolerance percent defective* (LTPD) or *rejectable quality level* (RQL), the quality level below which an item cannot be considered acceptable. This leads to a small acceptance of the batch, usually 10% – for the protection of the consumer against low quality batches. There exist two types of decision, acceptance or rejection of the batch, and two types of risks, to reject a "good" (high quality) batch, and to accept a "bad" (low quality) batch. The probabilities of occurrence of these risks are the so-called *supplier risk* and *consumer risk*, respectively. In a *single sampling plan*, the *supplier risk* is $\alpha = 1 - P_a(\text{AQL})$ and the *consumer risk* is $\beta = P_a(\text{LTPD})$. The sampling plans should take into account the specifications AQL and LTPD, i.e. we are supposed to find a single plan with an OCC that passes through the points (AQL, 1-α) and (LTPD, β). The construction of double plans which protect both the supplier and the consumer are much more difficult, and it is no longer sufficient to provide indication on two points of the OCC. There exist the so-called *Grubbs' tables* (see Montgomery 2009) providing (c_1, c_2, n_1, n_2), for $n_2 = 2n_1$, as an example, $\alpha = 0.05$, $\beta = 0.10$ and several rates RQL/AQL.

AOQ, AOQL and ATI. If there is a *rectifying inspection program* – a corrective program, based on a 100% inspection and replacement of nonconforming by conforming items, after the rejection of a batch by an AS plan –, the most relevant *characteristics* are the *average outgoing quality* (AOQ), $\text{AOQ}(p) = p\,(1 - n/N)\,P_a$, which attains

a maximum at the so-called *average output quality limit* (AOQL), the worst average quality of a product after a rectifying inspection program, as well as the *average total inspection* (ATI), the amount of items subject to inspection, equal to n if there is no rectification, but given by $\mathrm{ATI}(p) = nP_a + N(1 - P_a)$, otherwise.

Acknowledgments

Research partially supported by FCT/OE, POCI 2010 and PTDC/FEDER.

About the Author

For biography of M. Ivette Gomes *see* the entry ►Statistical Quality Control.

Cross References

►Industrial Statistics
►Sequential Sampling
►Statistical Quality Control
►Statistical Quality Control: Recent Advances

References and Further Reading

Dodge HF, Romig HG (1959) Sampling inspection tables, single and double sampling, 2nd edn. Wiley, New York
Duncan AJ (1986) Quality control and industrial statistics, 5th edn. Irwin, Homehood
Montgomery DC (2009) Statistical quality control: a modern introduction, 6th edn. Wiley, Hoboken, NJ
Pandey BN (2007) Statistical techniques in life-testing, reliability, sampling theory and quality control. Narosa, New Delhi
Schilling EG, Neubauer DV (2009) Acceptance sampling in quality control, 2nd edn. Chapman and Hall/CRC, New York
Stephens KS (2001) The handbook of applied acceptance sampling: plans, principles, and procedures. ASQ Quality, Milwaukee

Actuarial Methods

Vassiliy Simchera
Director
Rosstat's Statistical Research Institute, Moscow, Russia

A specific (and relatively new) type of financial calculations are actuarial operations, which represent a special (in majority of countries they are usually licensed) sphere of activity related to identifications of risks outcomes and market assessment of future (temporary) borrowed current assets and liabilities costs for their redemption.

The broad range of existing and applicable actuarial calculations require use of various methods and inevitably predetermines a necessity of their alteration depending on concrete cases of comparison analysis and selection of most efficient of them.

The condition of success is a typology of actuarial calculations methods, based on existing typology fields and objects of their applications, as well as knowledge of rule for selection of most efficient methods, which would provide selection of target results with minimum costs or high accuracy.

Regarding the continuous character of financial transactions, the actuarial calculations are carried out permanently. The aim of actuarial calculations in every particular case is probabilistic determination of profit sharing (transaction return) either in the form of financial liabilities (interest, margin, agio, etc.) or as commission charges (such as royalty).

The subject of actuarial calculations can be distinguished in the narrow and in the broad senses.

The given subject in the broad sense covers financial and actuarial accounts, budgeting, balance, audit, assessment of financial conditions and financial provision for all categories and types of borrowing institutions, basis for their preferential financial decisions and transactions, conditions and results of work for different financial and credit institutions; financial management of cash flows, resources, indicators, mechanisms, instruments, as well as financial analysis and audit of financial activity of companies, countries, nations their groups and unions, including national system of financial account, financial control, engineering, and forecast. In other words, the subject of actuarial calculations is a process of determination of any expenditures and incomes from any type of transactions in the shortest way.

In the narrow sense it is a process of determination, in the same way, of future liabilities and their comparison with present assets in order to estimate their sufficiency, deficit of surplus.

We can define general and efficient actuarial calculations, the principals of which are given below.

Efficient actuarial calculations imply calculations of any derivative indicators, which are carried out through conjugation (comparison) of two or more dissimilar initial indicators, the results of which are presented as different relative numbers (coefficients, norms, percents, shares, indices, rates, tariffs, etc.), characterizing differential (effect) of anticipatory increment of one indicator in comparison with another one.

In some cases similar values are called gradients, derivatives (of different orders), elasticity coefficients, or

anticipatory coefficients and can be determined by reference to more complex statistical and mathematical methods including geometrical, differential, integral, and correlation and regression multivariate calculations.

Herewith in case of application of nominal comparison scales for two or more simple values (so called scale of simple interests, which are calculated and represented in terms of current prices) they are determined and operated as it was mentioned by current nominal financial indicators, but in case of real scales application, i.e. scales of so called compound interests, they are calculated and represented in terms of future or current prices, that is real efficient financial indicators.

In case of insurance scheme the calculation of efficient financial indicators signify the special type of financial calculations i.e. actuarial calculations, which imply additional profit (discounts) or demanding compensation of loss (loss, damage or loss of profit) in connection with occurrence of contingency and risks (risk of legislation alteration, exchange rates, devaluation or revaluation, inflation or deflation, changes in efficiency coefficients).

Actuarial calculations represent special branch of activity (usually licensed activity) dealing with market assessment of compliance of current assets of insurance, joint-stock, investment, pension, credit and other financial companies (i.e. companies engaged in credit relations) with future liabilities to the repayment of credit in order to prevent insolvency of a debtor and to provide efficient protection for investors-creditors.

Actuarial calculations assume the comparison of assets (ways of use or allocation of obtained funds) with liabilities (sources of gained funds) for borrowing companies of all types and forms, which are carried out in aggregate by particular items of their expenses under circumstances of mutual risks in order to expose the degree of compliance or incompliance (surplus or deficit) of borrowed assets with future liabilities in term of repayment, in other words to check the solvency of borrowing companies.

Borrowing companies – insurance, stock, broker and auditor firms, banks, mutual, pension, and other specialized investment funds whose accounts payable two or more times exceeds their own assets and appear to be a source of high risk, which in turn affects interests of broad groups of business society as well as population – are considered as companies that are subjects to obligatory insurance and actuarial assessment.

Actuarial calculations assume the construction of balances for future assets and liabilities, probabilistic assessment of future liabilities repayment (debts) at the expense of disposable assets with regard to risks of changes of their amount on hand and market prices. The procedures of documentary adoption, which include construction of actuarial balances and preparation of actuarial reports and conclusions, are called actuarial estimation; the organizations that are carrying out such procedures are called actuarial organizations.

Hence, there is a necessity to learn the organization and technique of actuarial methods (estimations) in aggregate; as well as to introduce the knowledge of actuarial subjects to any expert who is involved in direct actuarial estimations of future assets and liabilities costs of various funds, credit, insurance, and similarly financial companies. This is true for assets and liabilities of any country.

The knowledge of these actuarial assessments and practical use is a significant reserve for increasing not only efficiency but (more important today) legitimate, transparent, and protected futures for both borrowing and lending companies.

Key Terms

Actuary (actuarius – Latin) – profession, appraiser of risks, certified expert on assessment of documentary insurance (and wider – financial) risks; in insurance – insurer; in realty agencies – appraiser; in accounting – auditor; in financial markets – broker (or bookmaker); in the past registrar and holder of insurance documents; in England – adjuster or underwriter.

Actuarial transactions – special field of activity related to determination of insurance outcomes in circumstances of uncertainty that require knowledge of probability theory and actuarial statistics methods and mathematics, including modern computer programs.

Actuarial assessment – type of practical activity, licensed in the majority of countries, related to preparation of actuarial balances, market assessment of current and future costs of assets and liabilities of insurer (in case of pension insurance assets and liabilities of non-governmental pension funds, insurances companies and specialized mutual trust funds); completed with preparation of actuarial report according to standard methodologies and procedures approved, as a rule in conventional (sometimes in legislative) order.

Actuarial estimations – documentary estimations of chance outcomes (betting) of any risk (gambling) actions (games) with participation of two or more parties with fixed (registered) rates of repayment of insurance premium and compensations premium for possible losses. They differ by criteria of complexity – that is elementary (simple or initial) and complex. The most widespread cases of elementary actuarial estimations are bookmaker estimations of profit and loss from different types of gambling including playing cards, lottery, and casinos, as well as risk

taking on modern stock exchange, foreign exchange markets, commodity exchanges, etc. The complex estimations assume determination of profit from second and consequent derived risks (outcomes over outcomes, insurance over insurance, repayment on repayment, transactions with derivatives, etc.). All of these estimations are carried out with the help of various method of high mathematics (first of all, numeric methods of probability theory and mathematical statistics). They are also often represented as methods of high actuarial estimations.

Generally due to ignorance about such estimations, current world debt (in 2008 approximately 700 trillion USD, including 300 trillion USD in the USA) has drastically exceeded real assets, which account for about 65 trillion USD, which is actually causing the enormous financial crisis everywhere in the world.

Usually such estimations are being undertaken towards future insurance operations, profits and losses, and that is why they are classified as strictly approximate and represented in categories of probabilistic expectations.

The fundamental methods of actuarial estimations are the following: methods for valuing investments, selecting portfolios, pricing insurance contracts, estimating reserves, valuing portfolios, controlling pension scheme, finances, asset management, time delays and underwriting cycle, stochastic approach to life insurance mathematics, pension funding and feed back, multiple state and disability insurance, and methods of actuarial balances.

The most popular range of application for actuarial methods are: 1) investments, (actuarial estimations) of investments assets and liabilities, internal and external, real and portfolio types their mathematical methods and models, investments risks and management; 2) life insurance (various types and methods, insurance bonuses, insurance companies and risks, role of the actuarial methods in management of insurance companies and reduction of insurance risks); 3) general insurance (insurance schemes, premium rating, reinsurance, reserving); 4) actuarial provision of pension insurance (pension investments – investment policy, actuarial databases, meeting the cost, actuarial researches).

Scientist who have greatly contributed to actuarial practices: William Morgan, Jacob Bernoulli, A. A. Markov, V. Y. Bunyakovsky, M. E. Atkinson, M. H. Amsler, B. Benjamin, G. Clark, C. Haberman, S. M. Hoem, W. F. Scott, and H. R. Watson.

World's famous actuary's schools and institutes: The Institute of Actuaries in London, Faculty of Actuaries in Edinburgh (on 25 May 2010, following a ballot of Fellows of both institutions, it was announced that the Institute and Faculty would merge to form one body – the "Institute and Faculty of Actuaries"), Charted Insurance Institute, International Association of Actuaries, International Forum of Actuaries Associations, International Congress of Actuaries, and Groupe Consultatif Actuariel Européen.

About the Author

Professor Vassiliy M. Simchera received his PhD at the age of 24 and his Doctor's degree when he was 35. He has been Vice-president of the Russian Academy of Economical Sciences (RAES), Chairman of the Academic Council and Counsel of PhDs dissertations of RAES, Director of Russian State Scientific and Research Statistical Institute of Rosstat (Moscow, from 2000). He was also Head of Chair of statistics in the All-Russian Distant Financial and Statistical Institute (1983–2000), Director of Computer Statistics Department in the State Committee on statistics and techniques of the USSR (1973–1983), and Head of Section of Statistical Researches in the Science Academy of the USSR (1965–1973). He has supervised 8 Doctors and over 50 PhD's. He has (co-) authored over 50 books and 350 articles, including the following books: *Encyclopedia of Statistical Publications* (2001, 991 p., in co-authorship), *Financial and Actuarial Calculations* (2002), *Organization of State Statistics in Russian Federation* (2004) and *Development of Russia's Economy for 100 Years*, 1900–2005 (2006). Professor Simchera was founder and executive director (1987–1991) of Russian Statistical Association, member of various domestic and foreign academies, as well as scientific councils and societies. He has received numerous honors and awards for his work, including Honored Scientist of Russian Federation (2001) (Decree of the President of the Russian Federation) and Saint Nicolay Chudotvoretz honor of III degree (2006). He is a full member of the International Statistical Institute (from 2001).

Cross References

►Careers in Statistics
►Insurance, Statistics in
►Kaplan-Meier Estimator
►Life Table
►Population Projections
►Probability, History of
►Quantitative Risk Management
►Risk Analysis
►Statistical Aspects of Hurricane Modeling and Forecasting
►Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts
►Survival Data

References and Further Reading

Benjamin B, Pollard JH (1980) The analysis of mortality and other actuarial statistics, 2nd edn. Heinemann, London

Black K, Skipper HD (1987) Life insurance. Prentice Hall, Englewood Cliffs, New Jersey

Booth P, Chadburn R, Cooper D, Haberman S and James D (1999) Modern actuarial theory and practice. Chapman and Hall/CHC, London, New York

Simchera VM (2003) Introduction to financial and actuarial calculations. Financy and Statistika Publishing House, Moscow

Teugels JL, Sundt B (2004) The encyclopedia of actuarial science, 3 vols. Wiley, Hoboken, NJ

Transactions of International Congress of Actuaries, vol. 1–10; J Inst Actuar, vol. 1–150

Adaptive Linear Regression

Jana Jurečková
Professor
Charles University in Prague, Prague, Czech Republic

Consider a set of data consisting of n observations of a response variable Y and of vector of p explanatory variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$. Their relationship is described by the *linear regression model* (see ►Linear Regression Models)

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + e.$$

In terms of the observed data, the model is

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i, \quad i = 1, 2, \ldots, n.$$

The variables $e_1, \ldots, e_n$ are unobservable model errors, which are assumed being independent and identically distributed random variables with a distribution function F and density f. The density is unknown, we only assume that it is symmetric around 0. The vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is an unknown parameter, and the problem of interest is to estimate $\boldsymbol{\beta}$ based on observations $Y_1, \ldots, Y_n$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, $i = 1, \ldots, n$.

Besides the classical ►least squares estimator, there exists a big variety of *robust estimators* of $\boldsymbol{\beta}$. Some are distributionally robust (less sensitive to deviations from the assumed shape of f), others are resistant to the leverage points in the design matrix and have a high breakdown point [introduced originally by Hampel (1968), the finite sample version is studied in Donoho and Huber (1983)].

The last 40 years brought a host of statistical procedures, many of them enjoying excellent properties and being equipped with a computational software (see ►Computational Statistics and ►Statistical Software: An Overview). On the other hand, this progress has put an applied statistician into a difficult situation: If one needs to fit the data with a regression hyperplane, he (she) is hesitating which procedure to use. If there is more information on the model, then the estimation procedure can be chosen accordingly. If the data are automatically collected by a computer and the statistician is not able to make any diagnostics, then he (she) might use one of the high breakdown-point estimators. However, many decline this idea due to the difficult computation. Then, at the end, the statistician can prefer the simplicity to the optimality and uses either the classical least squares (LS), LAD-method or other reasonably simple method.

Instead of to fix ourselves on one fixed method, one can try to combine two convenient estimation methods, and in this way diminish eventual shortages of both. Taylor (1973) suggested to combine the LAD (minimizing the L_1 norm) and the least squares (minimizing the L_2 norm) methods. Arthanari and Dodge (1981) considered a convex combination of LAD- and LS-methods. Simulation study by Dodge and Lindstrom (1981) showed that this procedure is robust to small deviations from the normal distribution (see ►Normal Distribution, Univariate). Dodge (1984) extended this method to a convex combination of LAD and Huber's M-estimation methods (see ►Robust Statistics and Robust Statistical Methods). Dodge and Jurečková (1987) observed that the convex combination of two methods could be adapted in such a way that the resulted estimator has the minimal asymptotic variance in the class of estimators of a similar kind, no matter what is the unknown distribution. The first numerical study of this procedure was made by Dodge et al. (1991). Dodge and Jurečková (1988, 1991) then extended the adaptive procedure to the combinations of LAD- with M-estimation and with the trimmed least squares estimation. The results and examples are summarized in monograph of Dodge and Jurečková (2000), where are many references added.

Let us describe the general idea, leading to a construction of an adaptive convex combination of two estimation methods: We consider a family of symmetric densities indexed by an suitable measure of scale s :

$$\mathcal{F} = \left\{ f : f(z) = s^{-1} f_0(z/s),\ s > 0 \right\}.$$

The shape of f_0 is generally unknown; it only satisfies some regularity conditions and the unit element $f_0 \in \mathcal{F}$ has the scale $s_0 = 1$. We take $s = 1/f(0)$ when we combine L_1-estimator with other class of estimators.

The scale characteristic s is estimated by a consistent estimator $\hat{s}_n$ based on $Y_1, \ldots, Y_n$, which is regression-invariant and scale-equivariant, i.e.,

(a) $\hat{s}_n(\mathbf{Y}) \xrightarrow{p} s$ as $n \to \infty$
(b) $\hat{s}_n(\mathbf{Y} + \mathbf{Xb}) = \hat{s}_n(\mathbf{Y})$ for any $\mathbf{b} \in \mathbb{R}^p$ (regression-invariance)
(c) $\hat{s}_n(c\mathbf{Y}) = c\hat{s}_n(\mathbf{Y})$ for $c > 0$ (scale-equivariance).

Such estimator based on the regression quantiles was constructed e.g., by Dodge and Jurečková (1995). Other estimators are described in the monograph by Koenker (2005).

The adaptive estimator $\mathbf{T}_n(\delta)$ of $\boldsymbol{\beta}$ is defined as a solution of the minimization problem

$$\sum_{i=1}^{n} \rho\left(\frac{Y_i - \mathbf{x}_i^\top \mathbf{t}}{\hat{s}_n}\right) := \min$$

with respect to $\mathbf{t} \in \mathbb{R}^p$, where

$$\rho(z) = \delta\rho_1(z) + (1-\delta)\rho_2(z) \tag{1}$$

with a suitable fixed δ, $0 \le \delta \le 1$, where $\rho_1(z)$ and $\rho_2(z)$ are symmetric (convex) discrepancy functions defining the respective estimators. For instance, $\rho_1(z) = |z|$ and $\rho_2(z) = z^2$ if we want to combine LAD and LS estimators. Then $\sqrt{n}(\mathbf{T}_n(\delta) - \boldsymbol{\beta})$ has an asymptotically normal distribution (see ►Asymptotic Normality) $\mathcal{N}_p(\mathbf{0}, \mathbf{Q}^{-1}\sigma^2(\delta,\rho,f))$ with the variance dependent on δ, ρ and f, where

$$\mathbf{Q} = \lim_{n\to\infty} \mathbf{Q}_n, \quad \mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top.$$

Using $\delta = \delta_0$ which minimizes $\sigma^2(\delta,\rho,f)$ with respect to δ, $0 \le \delta \le 1$, we get an estimator $\mathbf{T}_n(\delta_0)$ minimizing the asymptotic variance for a fixed distribution shape. Typically, $\sigma^2(\delta,\rho,f)$ depends on f only through two moments of f_0. However, these moments should be estimated on the data.

Let us illustrate the procedure on the combination of the least squares and the L_1 procedures. Set

$$\sigma^2 = \int z^2 f(z)dz, \quad \sigma_0^2 = \int z^2 f_0(z)dz \tag{2}$$
$$E_1^0 = \int |z| f_0(z)dz, \quad E_1 = \int |z| f(z)dz.$$

Then

$$\sigma^2 = \int z^2 f(z)dz = s^2\sigma_0^2, \qquad E_1 = \int |z| f(z)dz = sE_1^0$$

and the corresponding asymptotic variance of $\mathbf{T}_n(\delta)$ is

$$\sigma^2(\delta,f,s) = \frac{s^2}{4}\{4(1-\delta)^2\sigma_0^2 + 4\delta(1-\delta)E_1^0 + \delta^2\}. \tag{3}$$

If we know all moments in (2), we minimize the variance (3) with respect to δ, under the restriction $0 \le \delta \le 1$. It is minimized for $\delta = \delta_0$ where

$$\delta_0 = \begin{cases} 0 & \text{if } 2\sigma_0^2 \le E_1^0 < 1/2 \\ \dfrac{4\sigma_0^2 - 2E_1^0}{4\sigma_0^2 - 4E_1^0 + 1} & \text{if } E_1^0 < 1/2 \text{ and } E_1^0 < 2\sigma_0^2 \\ 1 & \text{if } 1/2 \le E_1^0 < 2\sigma_0^2. \end{cases}$$

The estimator $\mathbf{T}_n(\delta_0)$ of $\boldsymbol{\beta}$ is a solution of the minimization

$$\sum_{i=1}^{n} \rho((Y_i - \mathbf{x}_i^\top \mathbf{t})/\hat{s}_n) := \min, \quad \mathbf{t} \in \mathbb{R}^p,$$
$$\rho(z) = (1-\delta_0)z^2 + \delta_0|z|, \; z \in \mathbb{R}^1. \tag{4}$$

But δ_0 is unknown, because the entities in (2) depend on the unknown distribution f. Hence, we should replace δ_0 by an appropriate estimator based on $\mathbf{Y}$. We shall proceed in the following way:

First estimate $E_1^0 = E_1/s = f(0)\int_{\mathbb{R}} |z| f(z)dz$ by

$$\widehat{E}_1^0 = \hat{s}_n^{-1}(n-p)^{-1}\sum_{i=1}^{n}\left|Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_n\left(\frac{1}{2}\right)\right| \tag{5}$$

where $\widehat{\boldsymbol{\beta}}_n(1/2)$ is the LAD-estimator of $\boldsymbol{\beta}$. The choice of optimal $\widehat{\delta}_{0n}$ is then based on the following decision procedure (Table 1).

It can be proved that $\widehat{\delta}_{0n} \xrightarrow{p} \delta_0$ as $n \to \infty$ and $\mathbf{T}_n(\widehat{\delta}_{0n})$ is a consistent estimator of $\boldsymbol{\beta}$ and is asymptotically normally distributed with the minimum possible variance.

Adaptive Linear Regression. Table 1 Decision procedure

Compute $\widehat{E}_1^0$ as in (5).

(1) If $\widehat{E}_{1n}^0 < 1/2$, calculate

$$\widehat{\sigma}_{0n}^2 = \frac{1}{\hat{s}_n^2(n-p)}\sum_{i=1}^{n}\left(Y_i - \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}(1/2)\right)^2$$

and go to (2). If not, go to (4).

(2) If $\widehat{E}_{1n}^0 \ge 2\widehat{\sigma}_{0n}^2$, put $\widehat{\delta}_{0n} = 0$. Then $\mathbf{T}_n$ is the ordinary LS estimator of $\boldsymbol{\beta}$. If not, go to (3).

(3) If $\widehat{E}_{1n}^0 < 2\widehat{\sigma}_{0n}^2$, calculate

$$\widehat{\delta}_{0n} = \frac{4\widehat{\sigma}_{0n}^2 - 2\widehat{E}_{1n}^0}{4\widehat{\sigma}_{0n}^2 - 4\widehat{E}_{1n}^0 + 1}$$

and perform the minimization (4) with the function ρ equal to

$$(1-\widehat{\delta}_{0n})\sum_{i=1}^{n}\left(\frac{Y_i - \mathbf{x}_i'\mathbf{t}}{\hat{s}_n}\right)^2 + \widehat{\delta}_{0n}\sum_{i=1}^{n}\left|\frac{Y_i - \mathbf{x}_i^\top\mathbf{t}}{\hat{s}_n}\right|.$$

(4) Put $\widehat{\delta}_{0n} = 1$; then $\mathbf{T}_n$ is the LAD-estimate of $\boldsymbol{\beta}$.

Many numerical examp les based on real data can be find in the monograph Dodge and Jurečková (2000).

Acknowledgments

The research was supported by the Czech Republic Grant 201/09/0133 and by Research Projects MSM 0021620839 and LC 06024.

About the Author

Jana Jurečková was born on September 20, 1940 in Prague, Czechoslovakia. She has her Ph.D. in Statistics from Czechoslovak Academy of Sciences in 1967; some twenty years later, she was awarded the DrSc from Charles University. Her dissertation, under the able supervision of late Jaroslav Hajek, related to "uniform asymptotic linearity of rank statistics" and this central theme led to significant developments in nonparametrics, robust statistics, time series, and other related fields. She has extensively collaborated with other leading statisticians in Russia, USA, Canada, Australia, Germany, Belgium, and of course, Czech Republic, among other places. A (co-)author of several advanced monographs and texts in Statistics, Jana has earned excellent international reputation for her scholarly work, her professional accomplishment and her devotion to academic teaching and councelling. She has been with the Mathematics and Physics faculty at Charles University, Prague, since 1967, where she earned the Full Professor's rank in 1992. She has over 100 publications in the leading international journals in statistics and probability, and she has supervised a number of Ph.D. students, some of them have acquired international reputation on their own. (Communicated by P. K. Sen.)

Cross References

►Robust Regression Estimation in Generalized Linear Models
►Robust Statistical Methods
►Robust Statistics

References and Further Reading

Arthanari TS, Dodge Y (1981) Mathematical programming in statistics. Wily, Interscience Division, New York; (1993) Wiley Classic Library

Dodge Y (1984) Robust estimation of regression coefficient by minimizing a convex combination of least squares and least absolute deviations. Comp Stat Quart 1:139–153

Dodge Y, Jurečková J (1987) Adaptive combination of least squares and least absolute deviations estimators. In: Dodge Y (ed) Statistical data analysis based on L_1 – norm and related methods. North-Holland, Amsterdam, pp 275–284

Dodge Y, Jurečková J (1988) Adaptive combination of M-estimator and L_1 – estimator in the linear model. In: Dodge Y, Fedorov VV, Wynn HP (eds) Optimal design and analysis of experiments. North-Holland, Amsterdam, pp 167–176

Dodge Y, Jurečková J (1991) Flexible L-estimation in the linear model. Comp Stat Data Anal 12:211–220

Dodge Y, Jurečková J (1995) Estimation of quantile density function based on regression quantiles. Stat Probab Lett 23: 73–78

Dodge Y, Jurečková J (2000) Adaptive regression. Springer, New York. ISBN 0-387-98965-X

Dodge Y, Lindstrom FT (1981) An alternative to least squares estimations when dealing with contaminated data. Technical report No 79, Oregon State University, Corvallis

Dodge Y, Antoch J, Jurečková J (1991) Adaptive combination of least squares and least absolute deviation estimators. Comp State Data Anal 12:87–99

Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum KA, Hodges JL (eds) A festschrift for Erich Lehmann. Wadsworth, Belmont, California

Hampel FR (1968) Contributions to the theory of robust estimation. PhD Thesis, University of California, Berkely

Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge. ISBN 0-521-84573-4

Taylor LD (1973) Estimation by minimizing the sum of absolute errors. In: Zarembka P (ed) Frontiers in econometrics. Academic, New York, pp 189–190

Adaptive Methods

Saïd El Melhaoui
Professor Assistant
Université Mohammed Premier, Oujda, Morocco

Introduction

Statistical procedures, the efficiencies of which are optimal and invariant with regard to the knowledge or not of certain features of the data, are called adaptive statistical methods.

Such procedures should be used when one suspects that the usual inference assumptions, for example, the normality of the error's distribution, may not be met. Indeed, traditional methods have a serious defect. If the distribution of the error is non-normal, the power of classical tests, as *pseudo-Gaussian tests*, can be much less than the optimal power. So, the variance of the classical least squares estimator is much bigger than the smallest possible variance.

What Is Adaptivity?

The adaptive methods deal with the problem of estimating and testing hypotheses about a parameter of interest θ in the presence of nuisance parameter ν. The fact that ν remains unspecified induces, in general, a loss of efficiency

with the situation where v is exactly specified. *Adaptivity* occurs when the loss of efficiency is null, i.e., when we can estimate (testing hypotheses about) θ as when not knowing v as well as when knowing v. The method used in this respect is called *adaptive*.

Adaptivity is a property of the model under study, the best known of which is undoubtedly the symmetric location model; see Stone (1975). However, under a totally unspecified density, possibly non-symmetric, the mean can not be adaptively estimated.

Approaches to Adaptive Inference

Approaches to adaptive inference mainly belong to one of two types: either to estimate the unknown parameters v in some way, or to use the data itself to determine which statistical procedure is the most appropriate to these data. These two approaches are the starting points of two rather distinct strands of the statistical literature. *Nonparametric adaptive inference*, on one hand, where v is estimated from the sample, and on the other hand, *data-driven methods*, where the shape of v is identified via a selection statistic to distinguish the effective statistical procedure suitable at the current data.

Nonparametric Methods

The first approach is often used for the *semiparametric model*, where θ is a Euclidean parameter and the nuisance parameter is an infinite dimensional parameter f - often, the unspecified density of some white noise underlying the data generating process.

Stein (1956) introduced the notion of adaptation and gave a simple necessary condition for adaptation in semiparametric models. A comprehensive account of adaptive inference can be found in the monograph by Bickel et al. (1993) for semiparametric models with independent observations. Adaptive inference for dependent data have been studied in a series of papers, e.g., Kreiss (1987), Drost et al. (1997), and Koul and Schick (1997). The current state of the art is summarized in Grenwood et al. (2004).

The basic idea in this literature is to estimate the underlying f using a portion of the sample, and to reduce locally and asymptotically the semiparametric problem to a simpler parametric one, through the so-called "*least favorable parametric submodel*" argument. In general, the resulting computations are non-trivial.

An alternative technique is the use of *adaptive rank based statistics*. Hallin and Werker (2003) proposed a sufficient condition for adaptivity; that is, adaptivity occurs if a parametrically efficient method based on rank statistics can be derived. Then, it suffices, to substitute f in the rank statistics by an estimate $\hat{f}$ measurable on the ►*order statistics*. Some results in this direction have been obtained by Hájek (1962), Beran (1974), and Allal and El Melhaoui (2006).

Finally, these nonparametric adaptive methods, when they exist, are robust in efficiency: they cannot be outperformed by any non-adaptive method. However, these methods have not been widely used in practice, because the estimation of density, typically, requires a large number of observations.

Data-Driven Methods

The second strand of literature addresses the same problem of constructing adaptive inference, and consists of the use of the data to determine which statistical procedure should be used and then using the data again to carry out the procedure.

The was first proposed by Randles and Hogg (1973). Hogg et al. (1975) used the measure of symmetry and tailweight as selection statistics in and adaptive two-sample test. If the selection fell into one of the regions defined by the adaptive procedure, then a certain set of rank scores was selected, whereas if the selection statistic fell into a different region, then different rank scores would be used in the test. Hogg and Lenth (1984) proposed an adaptive estimator of the mean of symmetric distribution. They used selection statistics to determine if a mean, a 25% trimmed mean, or median should be used as an estimate of the mean of population. O'Gorman (2002) proposed an adaptive procedure that performs the commonly used tests of significance, including the two-sample test, a test for a slope in linear regression, and a test for interaction in two-way factorial design. A comprehensive account of this approach can be found in the monograph by O'Gorman (2004).

The advantage of the data-driven methods is that if an adaptive method is properly constructed, it automatically downweight outliers and could easily be applied in practice. However, and contrary to the nonparametric approach, the adaptive data-driven method is the best among the existing procedures, but not the best that can be built. As a consequence, the method so built is not definitively optimal.

Cross References

►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Robust Inference
►Robust Statistical Methods
►Robust Statistics

References and Further Reading

Allal J, El Melhaoui S (2006) Tests de rangs adaptatifs pour les modèles de régression linéaires avec erreurs ARMA. Annales des Sciences Mathématiques du Québec 30:29–54

Beran R (1974) Asymptotically efficient adaptive rank estimates in location models. Annals of Statistics 2:63–74

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1993) Efficient and adaptive estimation for semiparametric models. Johns Hopkins University Press, Baltimore, New York

Drost FC, Klaassen CAJ, Ritov Y, Werker BJM (1997) Adaptive estimation in time-series models. Ann Math Stat 29: 786–818

Greenwood PE, Muller UU, Wefelmeyer W (2004) An introduction to efficient estimation for semiparametric time series. In: Nikulin MS, Balakrishnan N, Mesbah M, Limnios N (eds) Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life. Statistics for Industry and Technology, Birkhäuser, Boston, pp. 253–269

Hájek J (1962) Asymptotically most powerful rank-order tests. Ann Math Stat 33:1124–1147

Hallin M, Werker BJM (2003) Semiparametric Efficiency Distribution-Freeness, and Invariance. Bernoulli 9:137–165

Hogg RV, Fisher DM, Randles RH (1975) A two simple adaptive distribution-free tests. J Am Stat Assoc 70:656–661

Hogg RV, Lenth RV (1984) A review of some adaptive statistical techniques. Commun Stat – Theory Methods 13:1551–1579

Koul HL, Schick A (1997) Efficient estimation in nonlinear autoregressive time-series models. Bernoulli 3:247–277

Kreiss JP (1987) On adaptive estimation in stationary ARMA processes. Ann Stat 15:112–133

O'Gorman TW (2002) An adaptive test of significance for a subset of regression coefficients. Stat Med 21:3527–3542

O'Gorman TW (2004) Applied adaptive statistical methods: tests of significance and confidence intervals. Society for Industrial and Applied Mathematics, Philadelphia

Randles RH, Hogg RV (1973) Adaptive distribution-free tests. Commun Stat 2:337–356

Stein C (1956) Efficient nonparametric testing and estimation. In: Proceedings of Third Berkeley Symposium on Mathametical Statistics and Probability, University of California Press, Berkeley, vol 1, pp. 187–195

Stone CJ (1975) Adaptive maximum likelihood estimators of a location parameter. Ann Stat 3:267–284

Adaptive Sampling

George A. F. Seber[1], Mohammad Salehi M.[2]
[1]Emeritus Professor of Statistics
Auckland University, Auckland, New Zealand
[2]Professor
Isfahan University of Technology, Isfahan, Iran

Adaptive sampling is particularly useful for sampling populations that are sparse but clustered. For example, fish can form large, widely scattered schools with few fish in between. Applying standard sampling methods such as simple random sampling (SRS, see ▶Simple Random Sample) to get a sample of plots from such a population could yield little information, with most of the plots being empty. The idea can be simply described follows. We go fishing in a lake using a boat and, assuming complete ignorance about the population, we select a location at random and fish. If we don't catch anything we select another location at random and try again. If we do catch something we fish in a specific neighborhood of that location and keep expanding the neighborhood until we catch no more fish. We then move on to a second location. This process continues until we have, for example, fished at a fixed number of locations or until our total catch has exceeded a certain number of fish. This kind of technique where the sampling is adapted to what turns up at each stage has been applied to a variety of diverse populations such as marine life, birds, mineral deposits, animal habitats, forests, and rare infectious diseases, and to pollution studies.

We now break down this process into components and introduce some general notation. Our initial focus will be on adaptive ▶cluster sampling, the most popular of the adaptive methods developed by Steven Thompson in the 1990s. Suppose we have a population of N plots and let y_i be a variable that we measure on the ith plot ($i = 1, 2, \ldots, N$). This variable can be continuous (e.g., level of pollution or biomass), discrete (e.g., number of animals or plants), or even just an indicator variable taking the value 1 for presence and zero for absence. Our aim is to estimate some function of the population y values such as, for example, the population total $\tau = \sum_{i=1}^{N} y_i$, the population mean $\mu = \tau/N$, or the population density $D = \tau/A$, where A is the population area.

The next step is to determine the nature of the neighborhood of each initially chosen plot. For example, we could choose all the adjacent units with a common boundary which, together with unit i, form a "cross" Neighborhoods can be defined to have a variety of patterns and the units in a neighborhood do not have to be contiguous (next to each other). We then specify a condition C such as $y_i > c$ which determines when we sample the neighborhood of the ith plot; typically $c = 0$ if y is a count. If C for the ith plot or unit is satisfied, we sample all the units in the neighborhood and if the rule is satisfied for any of those units we sample their neighborhoods as well, and so on, thus leading to a cluster of units. This cluster has the property that all the units on its "boundary" (called "edge units") do not satisfy C. Because of a dual role played by the edge units, the underlying theory is based on the concept of a network, which is a cluster minus its edge units.

It should be noted that if the initial unit selected is any one of the units in the cluster except an edge unit, then

all the units in the cluster end up being sampled. Clearly, if the unit is chosen at random, the probability of selecting the cluster will depend on the size of the cluster. For this reason adaptive cluster sampling can be described as unequal probability cluster sampling – a form of biased sampling.

The final step is to decide how we choose both the size and the method of selecting the initial sample size. Focusing on the second of these for the moment, one simple approach would be to use SRS to get a sample of size n_1, say. If a unit selected in the initial sample does not satisfy C, then there is no augmentation and we have a cluster of size one. We note that even if the units in the initial sample are distinct, as in SRS, repeats can occur in the final sample as clusters may overlap on their edge units or even coincide. For example, if two non-edge units in the same cluster are selected in the initial sample, then that whole cluster occurs twice in the final sample. The final sample then consists of n_1 (not necessarily distinct) clusters, one for each unit selected in the initial sample. We finally end up with a total of n units, which is random, and some units may be repeated.

There are many modifications of the above scheme depending on the nature of the population and we mention just a few. For example, the initial sample may be selected by sampling with replacement, or by using a form of systematic sampling (with a random start) or by using unequal probability sampling, as in sampling a tree with probability proportional to its basal area. Larger initial sampling units other than single plots can be used, for example a strip transect (primary unit) commonly used in both aerial and ship surveys of animals and marine mammals. Other shaped primary units can also be used and units in the primary unit need not be contiguous. If the population is divided into strata, then adaptive cluster sampling can be applied within each stratum, and the individual estimates combined. How they are combined depends on whether clusters are allowed to cross stratum boundaries or not. If instead of strata, we simply have a number of same-size primary units and choose a sample of primary units at random, and then apply the adaptive sampling within each of the chosen primary units, we have two-stage sampling with its appropriate theory.

In some situations, the choice of c in condition C is problematical as, with a wrong choice, we may end up with a feast or famine of plots. Thompson suggested using the data themselves, in fact the ►order statistics for the y_i values in the initial sample. Sometimes animals are not always detected and the theory has been modified to allow for incomplete detectability. If we replace y_i by a vector, then the scheme can be modified to allow for multivariate data.

We now turn our attention to sample sizes. Several ways of controlling sample sizes have been developed. For example, to avoid duplication we can remove a network once it has been selected by sampling networks without replacement. Sequential methods can also be used, such as selecting the initial sample sequentially until n exceeds some value. In fact Salehi, in collaboration with various other authors has developed a number of methods using both inverse and sequential schemes. One critical question remains: How can we use a pilot survey to design an experiment with a given efficiency or expected cost? One solution has been provided using the two-stage sampling method mentioned above (Salehi and Seber 1997).

We have not said anything about actual estimates as this would take several pages. However, a number of estimates associated with the authors Horvitz-Thompson (see ►Horvitz–Thompson Estimator), Hansen-Hurwitz, and Murthy have all been adapted to provide unbiased estimates for virtually all the above schemes and modifications. Salehi (1999) has also used the famous ►Rao-Blackwell theorem to provide more efficient unbiased estimates in a number of cases. The mentioned estimators based on small samples under adaptive cluster sampling often have highly skewed distributions. In such situations, confidence intervals (see ►Confidence Interval) based on traditional normal approximation can lead to unsatisfactory results, with poor coverage properties; for another solution see Salehi et al. (2009a).

As you can see, the topic is rich in applications and modifications and we have only told part of the story! For example, there is a related topic called adaptive allocation that has been used in fisheries; for a short review of adaptive allocation designs see Salehi et al. (2009b). Extensive references to the above are Thompson and Seber (1996) and Seber and Salehi (2004).

About the Author

Professor Seber was appointed to the foundation Chair in Statistics and Head of a newly created Statistics Unit within the Mathematics Department at the University of Auckland in 1973. He was involved in forming a separate Department of Statistics in 1994. He was awarded the Hector Medal by the Royal Society of New Zealand for fundamental contributions to statistical theory, for the development of the statistics profession in New Zealand, and for the advancement of statistics education through his teaching and writing (1999). He has authored or coauthored ten books as well as several second editions, and numerous research papers. However, despite the breadth of his contribution from linear models, multivariate statistics, linear regression, non-linear models, to adaptive sampling, he is perhaps still best known internationally for his research

on the estimation of animal abundance. He is the author of the internationally recognized text *Estimation of Animal Abundance and Related Parameters* (Wiley, 2nd edit., 1994; paperback reprint, Blackburn, 2002). The third conference on Statistics in Ecology and Environmental Monitoring was held in Dunedin (1999) "to mark and recapture the contribution of Professor George Seber to Statistical Ecology."

Cross References

►Cluster Sampling
►Empirical Likelihood Approach to Inference from Sample Survey Data
►Statistical Ecology

References and Further Reading

Salehi MM (1999) Rao-Blackwell versions of the Horvitz-Thompson and Hansen-Hurwitz in adaptive cluster sampling. J Environ Ecol Stat 6:183–195
Salehi MM, Seber GAF (1997) Two stage adaptive cluster sampling. Biometrics 53:959–970
Salehi MM, Mohammadi M, Rao JNK, Berger YG (2010a) Empirical Likelihood confidence intervals for adaptive cluster sampling. J Environ Ecol Stat 17:111–123
Salehi MM, Moradi M, Brown JA, Smith DR (2010b) Efficient estimators for adaptive two-stage sequential sampling. J Stat Comput Sim, DOI: 10.1080/00949650903005664
Seber GAF, Salehi MM (2004) Adaptive sampling. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 1, 2nd edn. Wiley, New York
Thompson SK, Seber GAF (1996) Adaptive sampling. Wiley, New York

Advantages of Bayesian Structuring: Estimating Ranks and Histograms

Thomas A. Louis
Professsor
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

Introduction

Methods developed using the Bayesian formalism can be very effective in addressing both Bayesian and frequentist goals. These advantages are conferred by full probability modeling are most apparent in the context of ►nonlinear models or in addressing non-standard goals. Once the likelihood and the prior have been specified and data observed, ►Bayes' Theorem maps the prior distribution into the posterior. Then, inferences are computed from the posterior, possibly guided by a ►loss function. This last step allows proper processing for complicated, non-intuitive goals. In this context, we show how the Bayesian approach is effective in estimating ►ranks and CDFs (histograms). We give the basic ideas; see Lin et al. (2006, 2008); Paddock et al. (2006) and the references thereof for full details and generalizations.

Importantly, as Carlin and Louis (2009) and many authors caution, the Bayesian approach is not a panacea. Indeed, the requirements for an effective procedure are more demanding than those for a frequentist approach. However, the benefits are many and generally worth the effort, especially now that ►Markov Chain Monte Carlo (MCMC) and other computing innovations are available.

A Basic Hierarchical Model

Consider a basic, compound sampling model with parameters of interest $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and data $\mathbf{Y} = (Y_1, \ldots, Y_K)$. The θ_k are *iid* and conditional on the θs, the Y_k are independent.

$$\begin{aligned} \theta_k &\overset{iid}{\sim} G(\cdot) \\ Y_k|\theta_k &\overset{indep}{\sim} f_k(Y_k|\theta_k) \end{aligned} \tag{1}$$

in practice, the θ_k might be the true differential expression of the *kth* gene, the true standardized mortality ratio for the *kth* dialysis clinic, or the true, underlying region-specific disease rate. Generalizations of (1) include adding a third stage to represent uncertainty in the prior, a regression model in the prior, or a priori association among the θs.

Assume that the θ_k and $\boldsymbol{\eta}$ are continuous random variables. Then, their posterior distribution is,

$$g(\boldsymbol{\theta} \mid \mathbf{Y}) = \prod_1^K g(\theta_k \mid Y_k) \tag{2}$$

$$g(\theta_k \mid Y_k) = \frac{f_k(Y_k \mid \theta_k) g(\theta_k)}{\int f_k(Y_k \mid s) g(s) ds} = \frac{f_k(Y_k \mid \theta_k) g(\theta_k)}{f_G(Y_k)}$$

Ranking

The ranking goal nicely shows the beauty of Bayesian structuring. Following Shen and Louis (1998), if the θ_k were directly observed, then their ranks (R_k) and percentiles (P_k) are:

$$R_k(\boldsymbol{\theta}) = \text{rank}(\theta_k) = \sum_{j=1}^K I_{\{\theta_k \geq \theta_j\}}; \; P_k(\boldsymbol{\theta}) = R_k(\boldsymbol{\theta})/(K+1). \tag{3}$$

The smallest θ has rank 1 and the largest has rank K. Note that the ranks are monotone transform invariant (e.g., ranking the logs of parameters produces the original ranks) and estimated ranks should preserve this invariance. In practice, we don't get to observe the θ_k, but can use their posterior distribution (2) to make inferences. For example, minimizing posterior squared-error loss *for the ranks* produces,

$$\bar{R}_k(\mathbf{Y}) = E_{\boldsymbol{\theta}|\mathbf{Y}}[R_k(\boldsymbol{\theta}) \mid \mathbf{Y}] = \sum_{j=1}^{K} \text{pr}(\theta_k \geq \theta_j \mid \mathbf{Y}). \qquad (4)$$

The $\bar{R}_k$ are shrunk towards the mid-rank, $(K+1)/2$, and generally are not integers. Optimal integer ranks result from ranking the $\bar{R}_k$, producing,

$$\hat{R}_k(\mathbf{Y}) = \text{rank}(\bar{R}_k(\mathbf{Y})); \; \hat{P}_k = \hat{R}_k/(K+1). \qquad (5)$$

Unless the posterior distributions of the θ_k are stochastically ordered, ranks based on maximum likelihood estimates or those based on hypothesis test statistics perform poorly. For example, if all θ_k are equal, MLEs with relatively high variance will tend to be ranked at the extremes; if Z-scores testing the hypothesis that a θ_k is equal to the typical value are used, then the units with relatively small variance will tend to be at the extremes. Optimal ranks compromise between these two extremes, a compromise best structured by minimizing posterior expected loss in the Bayesian context.

Example: The basic Gaussian-Gaussian model

We specialize (1) to the model with a Gaussian prior and Gaussian sampling distributions, with possibly different sampling variances. Without loss of generality assume that the prior mean is $\mu = 0$ and the prior variance is $\tau^2 = 1$. We have,

$$\begin{aligned} \theta_k \; &iid \; N(0,1), \\ Y_k|\theta_k &\sim N(\theta_k, \sigma_k^2) \\ \theta_k \mid Y_k \; &ind \; N\left(\theta_k^{pm}, (1-B_k)\sigma_k^2\right) \\ \theta_k^{pm} &= (1-B_k)Y_k; \; B_k = \sigma_k^2/(\sigma_k^2+1). \end{aligned}$$

The σ_k^2 are an ordered, geometric sequence with ratio of the largest σ^2 to the smallest $rls = \sigma_K^2/\sigma_1^2$ and ▶geometric mean $gmv = GM(\sigma_1^2, \ldots, \sigma_K^2)$. When $rls = 1$, the σ_k^2 are all equal. The quantity gmv measures the typical sampling variance and here we consider only $gmv = 1$.

Table 1 documents *SEL* performance for $\hat{P}_k$ (the optimal approach), Y_k (the MLE), ranked θ_k^{pm} and ranked $\exp\left\{\theta_k^{pm} + \frac{(1-B_k)\sigma_k^2}{2}\right\}$ (the posterior mean of e^{θ_k}). We present this last to assess performance for a monotone, non-linear transform of the target parameters. For $rls = 1$, the posterior distributions are stochastically ordered and the four sets of percentiles are identical, as is their performance. As rls increases, performance of Y_k-derived percentiles degrades, those based on the θ_k^{pm} are quite competitive with $\hat{P}_k$, but performance for percentiles based on the posterior mean of e^{θ_k} degrades as rls increases. Results show that though the posterior mean can perform well, in general it is not competitive with the optimal approach.

Advantages of Bayesian Structuring: Estimating Ranks and Histograms. Table 1 Simulated preposterior 10,000 × *SEL* for $gmv = 1$. As a baseline for comparison, if the data provided no information on the θ_k ($gmv = \infty$), all entries would be 833. If the data provided perfect information ($gmv = 0$), all entries would be 0

	Percentiles based on			
rls	$\hat{P}_k$	θ_k^{pm}	$\exp\left\{\theta_k^{pm} + \frac{(1-B_k)\sigma_k^2}{2}\right\}$	Y_k
1	516	516	516	516
25	517	517	534	582
100	522	525	547	644

Estimating the CDF or Histogram

Similar advantages of the Bayesian approach apply to estimating the empirical distribution function (EDF) of the θ_k,

$$G_K(t \mid \boldsymbol{\theta}) = K^{-1} \sum I_{\{\theta_k \leq t\}}.$$

As shown by Shen and Louis (1998), The optimal SEL estimate is

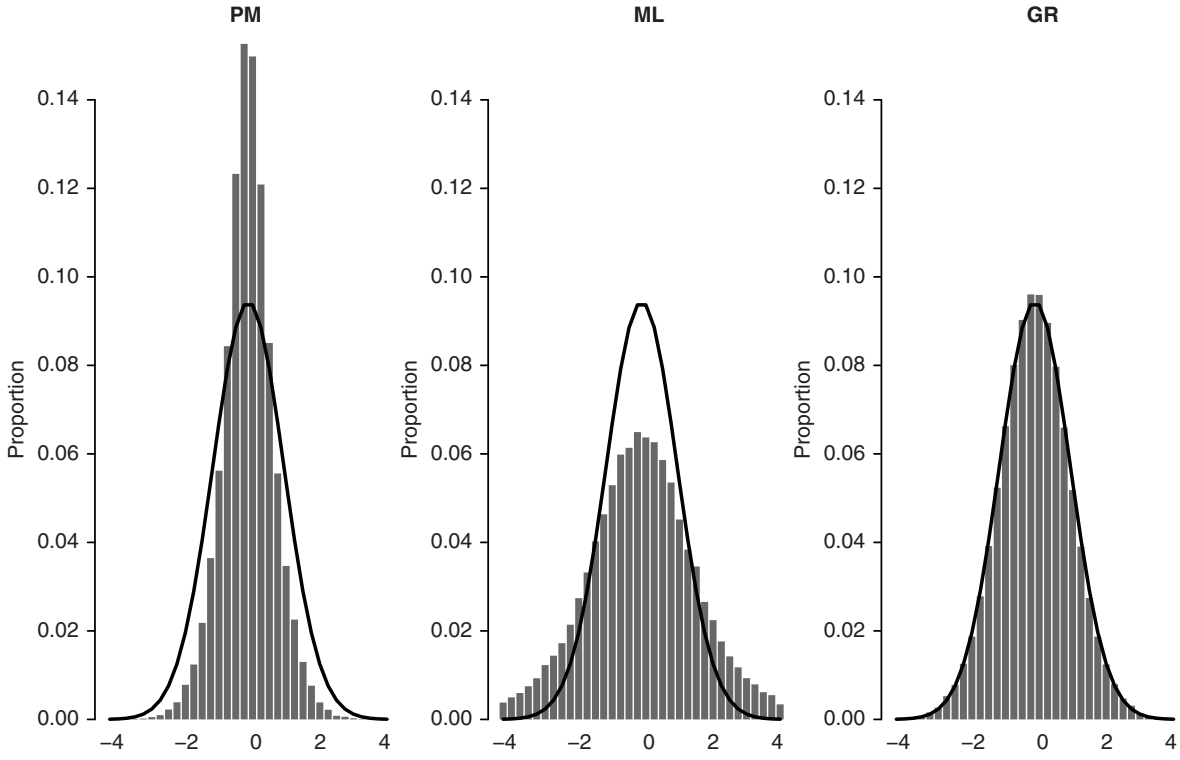
$$\bar{G}_K(t|\mathbf{Y}) = E[G_K(t \mid \boldsymbol{\theta})|\mathbf{Y}] = K^{-1} \sum Pr(\theta_k \leq t|\mathbf{Y}).$$

The optimal discrete distribution estimate with at most K mass points is $\hat{G}_K$, with mass K^{-1} at

$$\hat{U}_j = \bar{G}_K^{-1}\left(\frac{2j-1}{2K}\middle|\mathbf{Y}\right), \qquad j = 1, \ldots, K$$

The EDF is easy to compute from MCMC sampling. After burn-in, pool all θs, order them and set U_j equal to the $(2j-1)th$ order statistic.

Bayesian structuring to estimate G_K pays big dividends. As shown in Fig. 1, for the basic Gaussian model it produces the correct spread, whereas the histogram of the θ_k^{pm} (the posterior means) is under-dispersed and that of the Y_k (the MLEs) is over dispersed. More generally, when the true EDF is asymmetric or multi-modal,



Advantages of Bayesian Structuring: Estimating Ranks and Histograms. Fig. 1 Histogram estimates using θ^{pm}, ML, and $\bar{G}_K$ for the basic Gaussian/Gaussian model. $GM(\{\sigma_k^2\}) = 1$, $rls = 100$

the Bayesian approach also produces the correct shape Paddock et al. (2006).

Discussion

The foregoing are but two examples of the effectiveness of Bayesian structuring. Many more are available in the cited references and in other literature. In closing, we reiterate that the Bayesian approach needs to be used with care; there is nothing automatic about realizing its benefits.

Acknowledgments

Research supported by NIH/NIDDK Grant 5R01DK061662.

About the Author

Dr. Thomas Louis is Professor of Biostatistics, Johns Hopkins Bloomberg School of Public Health. He was President, International Biometric Society (IBS), Eastern North American Region (1992) and President, International Biometric Society (2006–2007). He is a Fellow of the American Statistical Association (1988), American Association for the Advancement of Science (1996), and Elected member, International Statistical Institute (1985). He was Editor, *JASA Applications and Case Studies* (2001–2003), Currently he is Co-editor, *Biometrics* (2009–2011). He is principal or co-advisor for 65 doctoral students and more than 40 masters students. He has delivered more than 450 invited presentations. Professor Louis has (co-)authored about 170 refereed papers and books, including *Bayesian Methods for Data Analysis* (with B.P. Carlin, Chapman & Hall/CRC, 3rd edition, 2009).

Cross References

►Bayes' Theorem
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Prior Bayes: Rubin's View of Statistics

References and Further Reading

Carlin BP, Louis TA (2009) Bayesian methods for data analysis, 3rd edn. Chapman and Hall/CRC, Boca Raton

Lin R, Louis TA, Paddock SM, Ridgeway G (2006) Loss function based ranking in two-stage, hierarchical models. Bayesian Anal 1:915–946

Lin R, Louis TA, Paddock SM, Ridgeway G (2009) Ranking of USRDS, provider-specific SMRs from 1998–2001. Health Serv Out Res Methodol 8:22–48

Paddock S, Ridgeway G, Lin R, Louis TA (2006) Flexible distributions for triple-goal estimates in two-stage hierarchical models. Comput Stat Data An 50(11):3243–3262
Shen W, Louis TA (1998) Triple-goal estimates in two-stage, hierarchical models. J Roy Stat Soc B 60:455–471

African Population Censuses

James P. M. Ntozi
Professor of Demographic Statistics
Makerere University, Kampala, Uganda

Definition

A Population ►census is the total process of collecting, compiling, evaluating, analyzing and disseminating demographic, economic and social data related to a specified time, to all persons in a country or a well defined part of a country.

History of Population Censuses

Population censuses are as old as human history. There are records of census enumerations as early as in 4000 BC in Babylonia, in 3000 BC in China and in 2500 BC in Egypt. The Roman Empire conducted population censuses and one of the most remembered censuses was the one held around AD 1 when Jesus Christ was born as his parents had moved from Nazareth to Bethlehem for the purpose of being counted. However, modern censuses did not start taking place until one was held in Quebec, Canada in 1666. This was followed by one in Sweden in 1749, USA in 1790, UK in 1801 and India 1871.

African Population Censuses

In the absence of complete civil registration systems in Africa, population censuses provide one of the best sources of socioeconomic and demographic information for the continent. Like in other parts of the world, censuses in Africa started as headcounts and assemblies until after the Second World War. The British were the first to introduce modern censuses in their colonial territories in west, east and southern Africa. For example in East Africa, the first modern census was conducted in 1948 in what was being referred to as British East Africa consisting of Kenya and Uganda. This was followed by censuses in 1957 in Tanzania, in 1959 in Uganda and 1962 in Kenya to prepare the countries for their political independence in 1961, 1962 and 1963, respectively. Other censuses have followed in these three countries. Similarly, the British West African countries of Ghana, Gambia, Nigeria and Sierra Leone were held in 1950s, 1960s and 1970s. In Southern Africa, similar censuses were held in Botswana, Lesotho, Malawi, Swaziland, Zambia and Zimbabwe in 1960s and 1970s, long before the Francophone and Lusophone countries did so. It was not until in 1970s and 1980s that the Francophone and Lusophone African countries started doing censuses instead of sample surveys which they preferred.

To help African countries do population censuses, United Nations set up an African census programme in late 1960s. Out of 41 countries, 22 participated in the programme. This programme closed in 1977 and was succeeded by the Regional Advisory Services in the demographic statistics set up as a section of Statistics Division at the United Nations Economic Commission for Africa. This section supported many African countries in conducting the 1980 and 1990 rounds of censuses. The section was superseded by the UNFPA sub-regional country support teams stationed in Addis Ababa, Cairo, Dakar and Harare. Each of these teams had census experts to give advisory services to countries in the 2000 round of censuses. These teams have now been reduced to three teams stationed in Pretoria, Cairo and Dakar and are currently supporting the African countries in population censuses.

There were working group committees on census on each round of censuses to work on the content of census ►questionnaire. For instance, in the 1980 round of censuses the working group recommended that the census questionnaire should have geographic characteristics, demographic characteristics, economic characteristics, community level variables and housing characteristics. In 1990 round of censuses, questions on the disabled persons were recommended to be added to the 1980 round questions. Later in the 2000 round of censuses, questions on economic establishments, agricultural sector and deaths in households were added. In the current round of 2010 censuses, the questions on disability were sharpened to capture the data better. New questions being asked include those on child labour, age at first marriage, ownership of mobile phone, ownership of email address, access to internet, distance to police post, access to salt in household, most commonly spoken language in household and cause of death in household.

In the 1960 and 1970s round of censuses, Post enumeration surveys (PES) to check on the quality of the censuses were attempted in Ghana. However, the experience with and results from PES were not encouraging, which discouraged most of the African countries from conducting them. Recently, the Post enumeration surveys have been revived and conducted in several African

countries like South Africa, Tanzania and Uganda. The challenges of PES have included: poor cartographic work, neglecting operational independence, inadequate funding, fatigue after the census, matching alternative names, lack of qualified personnel, useless questions in PES, probability sample design and selection, field reconciliation, lack of unique physical addresses in Africa and neglect of pretest of PES.

The achievements of the African censuses include supplying the needed sub-national data to the decentralized units for decision making processes, generating data for monitoring poverty reduction programmes, providing information for measuring indicators of most MDGs, using the data for measuring the achievement of indicators of International Conference on Population and Development (ICP), meeting the demand for data for emerging issues of socioeconomic concerns, accumulating experience in the region of census operations and capacity building at census and national statistical offices.

However, there are still several limitations associated with the African censuses. These have included inadequate participation of the population of the region; only 57% of the African population was counted in the 2000 round of censuses, which was much below to what happened in other regions: Oceania – 100%, Europe and North America – 99%, Asia – 97%, South America – 80% and the world – 91%. Other shortcomings were weak organizational and managerial skills, inadequate funding, non-conducive political environment, civil conflicts, weak technical expertise at NSOs and lack of data for gender indicators.

About the Author

Dr. James P. M. Ntozi is a Professor of demographic statistics at the Institute of Statistics, Makerere University, Kampala, Uganda. He is a founder and Past president of Uganda Statistical Society and Population Association of Uganda. He was a Council member of the International Statistical Institute and Union for African Population Studies, currently a Fellow and Chartered Statistician of the Royal Statistical Society and Council member of the Uganda National Academy of Sciences. He has authored, coauthored, and presented over 100 scientific papers as well as 6 books on fertility and censuses in Africa. He was an Editor of *African Population Studies*, co-edited 4 books, and is currently on the editorial board of *African Statistical Journal* and the *Journal of African Health Sciences*. He has received awards from Population Association of America, Uganda Statistical Society, Makerere University, Bishop Stuart University, Uganda and Ankole Diocese, Church of Uganda. James has been involved in planning and implementation of past Uganda censuses of population and housing of 1980, 1991, and 2002. He is currently helping the Liberian Statistical office to analyze the 2008 census data. Professor Ntozi is a past Director of the Institute of Statistics and Applied Economics, a regional statistical training center based at Makerere University, Uganda, and responsible for training many leaders in statistics and demography in sub-Saharan Africa for over 40 years. His other professional achievements have been research and consultancies in fertility, HIV/AIDS, Human Development Reports, and strategic planning.

Cross References

►Census
►Economic Statistics
►Integrated Statistical Databases
►Population Projections
►Promoting, Fostering and Development of Statistics in Developing Countries
►Role of Statistics: Developing Country Perspective
►Selection of Appropriate Statistical Methods in Developing Countries

References and Further Reading

Onsembe JO (2009) Postenumeration surveys in Africa. Paper presented at the 57th ISI session, Durban, South Africa
Onsembe JO, Ntozi JPM (2006) The 2000 round of censuses in Africa: achievements and challenges. Afr Stat J 3, November 6

Aggregation Schemes

Devendra Chhetry
President of the Nepal Statistical Association (NEPSA),
Professor and Head
Tribhuvan University, Kathmandu, Nepal

Given a data vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and a weight vector $\mathbf{w} = (w_1, w_2, \ldots, w_n)$, there exist three aggregation schemes in the area of statistics that, under certain assumptions, generate three well-known measures of location: arithmetic mean (*AM*), ►geometric mean (*GM*), and ►harmonic mean (*HM*), where it is implicitly understood that the data vector $\mathbf{x}$ contains values of a single variable. Among all these three measures, *AM* is more frequently used in statistics for some theoretical reasons. It is well known that $AM \geq GM \geq HM$ where equality holds only when all components of $\mathbf{x}$ are equal.

In recent years, some of these three and a new aggregation scheme are being practiced in the aggregation of development or deprivation indicators by extending the definition of data vector to a vector of indicators, in the sense that it contains measurements of development or deprivation of several sub-population groups or measurements of several dimensions of development or deprivation. The measurements of development or deprivation are either available in the form of percentages or need to be transformed in the form of unit free indices. Physical Quality of Life Index (Morris 1979), Human Development Index (UNDP 1991), Gender-related Development Index (UNDP 1995), Gender Empowerment Measure (UNDP 1995), and Human Poverty Index (UNDP 1997) are some of the aggregated indices of several dimensions of development or deprivation.

In developing countries, aggregation of development or deprivation indicators is a challenging task, mainly due to two reasons. First, indicators usually display large variations or inequalities in the achievement of development or in the reduction of deprivation across the sub-populations or across the dimensions of development or deprivation within a region. Second, during the process of aggregation it is desired to incorporate the public aversion to social inequalities or, equivalently, public preference for social equalities. Public aversion to social inequalities is essential for development workers or planners of developing countries for bringing marginalized sub-populations into the mainstream by monitoring and evaluation of the development works. Motivated by this problem, Anand and Sen (UNDP 1995) introduced the notion of the gender-equality sensitive indicator (GESI).

In societies of equal proportion of female and male population, for example, the AM of 60 and 30 percent of male and female literacy rate is the same as that of 50 and 40 percent, showing that AM fails to incorporate the public aversion to gender inequality due to the *AM's built-in problem of perfect substitutability*, in the sense that a 10 percentage point decrease in female literacy rate in the former society as compared to the latter one is substituted by the 10 percentage point increase in male literacy rate. The *GM* or *HM*, however, incorporates the public aversion to gender inequality because they do not posses the perfect substitutability property. Instead of *AM*, Anand and Sen used *HM* in the construction of GESI.

In the above example consider that society perceives the social problem from the perspective of deprivation; that is, instead of gender-disaggregated literacy rates society considers gender-disaggregated illiteracy rates. Arguing as before, it immediately follows that *AM* fails to incorporate the public aversion to gender inequality. It also follows that neither *GM* nor *HM* incorporates the public aversion to gender inequality. A new aggregation scheme is required for aggregating indicators of deprivation.

So far, currently practiced aggregation schemes are accommodated within a slightly modified version of the following single mathematical function due to Hardy et al. (1952) under the assumption that components of $\mathbf{x}$ and $\mathbf{w}$ are positive and the sum of the components of $\mathbf{w}$ is unity.

$$\mu(\mathbf{x}, \mathbf{w}, r) = \begin{cases} \left(\sum_{i=1}^{n} w_i {x_i}^r \right)^{1/r} & \text{if } r \neq 0, \\ \prod_{i=1}^{n} {x_i}^{w_i} & \text{if } r = 0. \end{cases} \tag{1}$$

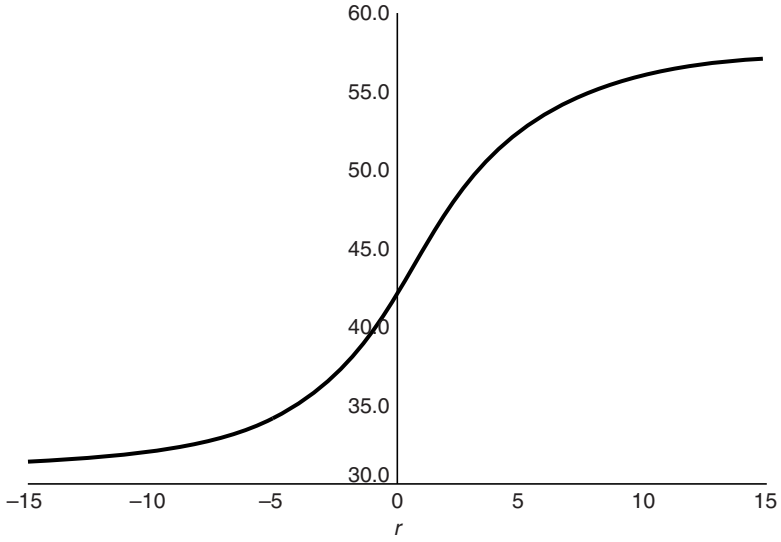
For fixed $\mathbf{x}$ and $\mathbf{w}$, the function (1) is defined for all real numbers, implying that the function (1) yields an infinite number of aggregation schemes. In particular, it yields *AM* when $r = 1$, *HM* when $r = -1$, and obviously *GM* when $r = 0$, and a new aggregation scheme suggested by Anand and Sen in constructing Human Poverty Index when $n = 3$, $w_1 = w_2 = w_3 = 1/3$ and $r = 3$ (UNDP 1997). It is well known that the values of the function are bounded between $x_{(1)}$ and $x_{(n)}$, where $x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$ and $x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$, and the function is strictly increasing with respect to r if all the components of data vector are not equal (see Fig. 1 when $w_1 = w_2 = 0.5$, $x_1 = 60\%$ and $x_2 = 30\%$).

The first two partial derivatives of the function with respect to the k^{th} component of the vector $\mathbf{x}$ yield the following results where $g(\mathbf{x}, \mathbf{w})$ is *GM*.

$$\frac{\partial \mu(\mathbf{x}, \mathbf{w}, r)}{\partial x_k} = \begin{cases} w_k \left(\frac{x_k}{\mu(\mathbf{x}, \mathbf{w}, r)} \right)^{r-1} & \text{if } r \neq 0, \\ w_k \, g(\mathbf{x}, \mathbf{w}) x_k^{-1} & \text{if } r = 0. \end{cases} \tag{2}$$

$$\frac{\partial^2 \mu(\mathbf{x}, \mathbf{w}, r)}{\partial x_k^2} = \begin{cases} (r-1) w_k \left[\frac{x_k}{\mu(\mathbf{x}, \mathbf{w}, r)} \right]^{r-2} \sum_{i \neq k} \mathbf{w_i} \mathbf{x_i^r} & \text{if } r \neq 0, \\ w_k \, (w_k - 1) g(\mathbf{x}, \mathbf{w}) x_k{}^{-}2 & \text{if } r = 0. \end{cases} \tag{3}$$

For fixed $\begin{pmatrix} r < 1 \\ r > 1 \end{pmatrix}$ and $\mathbf{w}$, (2) and (3) imply that the function (1) is increasing and $\begin{pmatrix} \text{concave} \\ \text{convex} \end{pmatrix}$ with



Aggregation Schemes. Fig. 1 Nature of the function in a particular case

respect to each x_k, implying that the aggregated value increases at $\begin{pmatrix}\text{decreasing}\\\text{increasing}\end{pmatrix}$ rate with respect to each component of **x**. These properties are desirable for aggregating the $\begin{pmatrix}\text{development}\\\text{deprivation}\end{pmatrix}$ indicators, since the aggregated value of $\begin{pmatrix}\text{development}\\\text{deprivation}\end{pmatrix}$ is expected to $\begin{pmatrix}\text{rise}\\\text{fall}\end{pmatrix}$ from the $\begin{pmatrix}\text{floor to the ceiling value}\\\text{ceiling to the floor value}\end{pmatrix}$ at decreasing rate with respect to each component of **x**. For given **x** and **w**, the function (1) with any value of r, $\begin{pmatrix}r<1\\r>1\end{pmatrix}$, could be used to aggregate the $\begin{pmatrix}\text{development}\\\text{deprivation}\end{pmatrix}$ indicators if the public aversion to social inequalities should be incorporated.

What value of r should one use in practice? There is no simple answer to this question, since the answer depends upon the society's degree of preference for social equality. If a society has no preference for social equality, then one can use $r = 1$ in aggregating development or deprivation indicators, which is still a common practice in developing countries, even though the public efforts for bringing marginalized sub-populations into the mainstream has become a major agenda of development.

If a society has preference for social equality, then subjective judgment in the choice of r seems to be unavoidable. For the purpose of monitoring and evaluation, such judgment does not seem to be a serious issue as long as a fixed value of r is decided upon. In this context, Anand and Sen suggested using $r = -1$ for aggregating the indicators of development when $n = 2$ (UNDP 1995), and $r = 3$ for aggregating the indicators of deprivation when $n = 3$ (UNDP 1997). A lot of research work still needs to be done in this area for producing social-equality sensitive indicators of development or deprivation.

Cross References

►Composite Indicators
►Lorenz Curve
►Role of Statistics: Developing Country Perspective

References and Further Reading

Hardy GH, Littlewood JE, Polya G (1952) Inequalities. Cambridge University Press, London
Morris MD (1979) Measuring the condition of the world's poor: the physical quality of life index. Frank Case, London
UNDP (1991) Human Development Report 1991, Financing Human Development Oxford University Press, New York
UNDP (1995) Human Development Report 1995, Gender and Human Development. Oxford University Press, New York
UNDP (1997) Human Development Report 1997, Human Development to Eradicate Poverty. Oxford University Press, New York

Agriculture, Statistics in

Gavin J. S. Ross
Rothamsted Research, Harpenden, UK

The need to collect information on agricultural production has been with us since the dawn of civilization. Agriculture was the main economic activity, supplying both food for growing populations and the basis for taxation. The Sumerians of Mesopotamia before 3000 BC developed writing systems in order to record crop yields and livestock numbers. The Ancient Egyptians recorded the extent and productivity of arable land on the banks of the Nile. Later conquerors surveyed their new possessions, as in the Norman conquest of England which resulted in the Domesday Book of 1086, recording the agricultural potential of each district in great detail.

The pioneers of scientific agriculture, such as J.B. Lawes and J.H.Gilbert at Rothamsted, England, from 1843 onwards, insisted on accurate measurement and recording as the first requirement for a better understanding of the processes of agricultural production. The Royal Statistical Society (RSS) was founded in 1834 with its symbol of a sheaf of corn, implying that the duty of statisticians was to gather numerical information, but for others to interpret the data. Lawes published numerous papers on the variability of crop yields from year to year, and later joined the Council of the RSS. By 1900 agricultural experiments were conducted in several countries, including Germany, the Netherlands and Ireland, where W.S. Gosset, publishing under the name of "Student," conducted trials of barley varieties for the brewing industry.

In 1919 R.A. Fisher was appointed to analyze the accumulated results of 70 years of field experimentation at Rothamsted, initiating a revolution in statistical theory and practice. Fisher had already published the theoretical explanation of Student's t-distribution and the sampling distribution of the correlation coefficient, and challenged Karl Pearson's position that statistical analysis was only possible with large samples. His first task was to study the relationship between rainfall and crop yields on the long-term experiments, for which he demanded a powerful mechanical calculator, the "Millionaire." Introducing orthogonal polynomials to fit the yearly weather patterns and to eliminate the long-term trend in crop yield, he performed multiple regressions on the rainfall components, and developed the variance ratio test (later the F-distribution) to justify which terms to include using what became the ►analysis of variance. If the results were of minor interest to farmers, the methods used were of enormous importance in establishing the new methodology of curve fitting, regression analysis and the analysis of variance.

Fisher's work with agricultural scientists brought him a whole range of statistical challenges. Working with small samples he saw the role of the statistician as one who extracts the information in a sample as efficiently as possible. Working with non-normally distributed data he proposed the concept of likelihood, and the method of maximum likelihood to estimate parameters in a model. The early field experiments at Rothamsted contained the accepted notion of comparison of treatments with controls at the same location, and some plots included factorial combinations of fertilizer sources. Fisher saw that in order to apply statistical methods to assess the significance of observed effects it was necessary to introduce ►randomization and replication. Local control on land of varying fertility could be improved by blocking, and for trends in two directions he introduced Latin Square designs. The analysis of factorial experiments could be expressed in terms of main effects and interaction effects, with the components of interaction between blocks and treatments regarded as the basic residual error variance.

Fisher's ideas rapidly gained attention and his ideas and methods were extended to many fields beyond agricultural science. George Snedecor in Iowa, Mahalanobis and C.R. Rao in India, were early disciples, and his assistants included L.H.C. Tippett, J. Wishart and H. Hotelling. He was visited in 1926 by J. Neyman, who was working with agricultural scientists in Poland. In 1930 he was joined by Frank Yates who had experience of ►least squares methods as a surveyor in West Africa. Fisher left Rothamsted in 1933 to pursue his interests in genetics, but continued to collaborate with Yates. They introduced Balanced Incomplete Blocks and Lattice designs, and Split Plot designs with more than one component of error variance. Their *Statistical Tables*, first published in 1938, were widely used for many decades later.

Yates expanded his department to provide statistical analysis and consulting to agricultural departments and institutes in Britain and the British Empire. Field experimentation spread to South America with W.L Stevens, and his assistants W.G. Cochran, D.J. Finney and O. Kempthorne became well-known statistical innovators in many applications. During World War II Yates persuaded the government of the value of sample surveys to provide information about farm productivity, pests and diseases and fertilizer use. He later advised Indian statisticians on

the design and analysis of experiments in which small farmers in a particular area might be responsible for one plot each.

In 1954 Yates saw the potential of the electronic computer in statistical research, and was able to acquire the first computer devoted to civilian research, the Elliott 401. On this computer the first statistical programs were written for the analysis of field experiments and surveys, for bioassay and ►probit analysis, for multiple regression and multivariate analysis, and for model fitting by maximum likelihood. All the programs were in response to the needs of agricultural scientists, at field or laboratory level, including those working in animal science. Animal experiments typically had unequal numbers of units with different treatments, and iterative methods were needed to fit parameters by least squares or maximum likelihood. Animal breeding data required lengthy computing to obtain components of variance from which to estimate heritabilities and selection indices. The needs of researcher workers in fruit tree research, forestry, glasshouse crops and agricultural engineering all posed different challenges to the statistical profession.

In 1968 J.A. Nelder came to Rothamsted as head of the Statistics Department, having been previously at the National Vegetable Research Station at Wellesbourne, where he had explored the used of systematic designs for vegetable trials, and had developed the well-used Simplex Algorithm with R. Mead to fit ►nonlinear models. With more powerful computers it was now possible to combine many analyses into one system, and he invited G.N. Wilkinson from Adelaide to include his general algorithm for the analysis of variance in a more comprehensive system that would allow the whole range of nested and crossed experimental designs to be handled, along with facilities for regression and multivariate analysis. The program GENSTAT is now used world-wide in agricultural and other research settings.

Nelder worked with R.M. Wedderburn to show how the methodology of Probit Analysis (fitting binomial data to a transformed regression line) could be generalized to a whole class of ►Generalized Linear Models. These methods were particularly useful for the analysis of multiway contingency tables, using logit transformations for binomial data and log transformations for positive data with long-tailed distributions. The applications may have been originally in agriculture but found many uses elsewhere, such as in medical and pharmaceutical research.

The needs of soil scientists brought new classes of statistical problems. The classification of soils was complicated by the fact that overlapping horizons with different properties did not occur at the same depth, although samples were essential similar but displaced. The method of Kriging, first used by South African mining engineers, was found to be useful in describing the spatial variability of agricultural land, with its allowance for differing trends and sharp boundaries.

The need to model responses to fertilizer applications, the growth of plants and animals, and the spread of weeds, pests and diseases led to developments in fitting non-linear models. While improvements in the efficiency of numerical optimization algorithms were important, attention to the parameters to be optimized helped to show the relationship between the model and the data, and which observations contributed most to the parameters of interest. The limitations of agricultural data, with many unknown or unmeasurable factors present, makes it necessary to limit the complexity of the models being fitted, or to fit common parameters to several related samples.

Interest in spatial statistics, and in the use of models with more than one source of error, has led to developments such as the powerful REML algorithm. The use of intercropping to make better use of productive land has led to appropriate developments in experimental design and analysis.

With the increase in power of computers it became possible to construct large, complex models, incorporating where possible known relationships between growing crops and all the natural and artificial influences affecting their growth over the whole cycle from planting to harvest. These models have been valuable in understanding the processes involved, but have not been very useful in predicting final yields. The statistical ideas developed by Fisher and his successors have concentrated on the choices which farmers can make in the light of information available at the time, rather than to provide the best outcomes for speculators in crop futures. Modeling on its own is no substitute for continued experimentation.

The challenge for the 21st century will be to ensure sustainable agriculture for the future, taking account of climate change, resistance to pesticides and herbicides, soil degradation and water and energy shortages. Statistical methods will always be needed to evaluate new techniques of plant and animal breeding, alternative food sources and environmental effects.

About the Author

Gavin J.S. Ross has worked in the Statistics Department at Rothamsted Experimental Station since 1961, now as a retired visiting worker. He served under Frank Yates,

John Nelder and John Gower, advising agricultural workers, and creating statistical software for nonlinear modelling and for cluster analysis and multivariate analysis, contributing to the GENSTAT program as well as producing the specialist programs MLP and CLASP for his major research interests. His textbook *Nonlinear Estimation* (Springer 1990) describes the use of stable parameter transformations to fit and interpret nonlinear models. He served as President of the British Classification Society.

Cross References

►Analysis of Multivariate Agricultural Data
►Farmer Participatory Research Designs
►Spatial Statistics
►Statistics and Climate Change

References and Further Reading

Cochran WG, Cox GM (1957) Experimental designs, 2nd edn. Wiley, New York
Finney DJ (1962) An introduction to statistical science in agriculture. Edinburgh, Oliver and Boyd
Fisher RA (1924) The influence of rainfall on the yield of wheat at Rothamsted. Phil Trans Roy Soc London B 213:89–142
Mead R, Curnow RM (1983) Statistical methods in agriculture and experimental biology, 2nd edn. Chapman and Hall, London
Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. Biometrika 58(3): 545–554
Webster R, Oliver MA (2007) Geostatistics for environmental scientists, 2nd edn. Wiley, New York
Yates F (1981) Sampling methods for censuses and surveys, 4th edn. Griffin, London

Akaike's Information Criterion

Hirotugu Akaike†
Former Director General of the Institute of Statistical Mathematics and a Kyoto Prize Winner
Tokyo, Japan

The Information Criterion $I(g:f)$ that measures the deviation of a model specified by the probability distribution f from the true distribution g is defined by the formula

$$I(g:f) = E\log g(X) - E\log f(X).$$

Here E denotes the expectation with respect to the true distribution g of X. The criterion is a measure of the deviation of the model f from the true model g, or the best possible model for the handling of the present problem.

The following relation illustrates the significant characteristic of the log likelihood:

$$I(g:f_1) - I(g:f_2) = -E(\log f_1(X) - \log f_2(X)).$$

This formula shows that for an observation x of X the log likelihood $\log f(x)$ provides a relative measure of the closeness of the model f to the truth, or the goodness of the model. This measure is useful even when the true structure g is unknown.

For a model $f(X/\mathbf{a})$ with unknown parameter $\mathbf{a}$ the maximum likelihood estimate $\mathbf{a}(x)$ is defined as the value of $\mathbf{a}$ that maximizes the likelihood $f(x/\mathbf{a})$ for a given observation x. Due to this process the value of $\log f(x/\mathbf{a}(x))$ shows an upward bias as an estimate of $\log f(X/\mathbf{a})$. Thus to use $\log f(x/\mathbf{a}(x))$ as the measure of the goodness of the model $f(X/\mathbf{a})$, it must be corrected for the expected bias.

In typical application of the method of maximum likelihood this expected bias is equal the dimension, or the number of components, of the unknown parameter $\mathbf{a}$. Thus the relative goodness of a model determined by the maximum likelihood estimate is given by

AIC = −2 (log maximum likelihood − (number of parameters)).

Here log denotes natural logarithm. The coefficient −2 is used to make the quantity similar to the familiar chi-square statistic in the test of dimensionality of the parameter.

AIC is the abbreviation of An Information Criterion.

About the Author

Professor Akaike died of pneumonia in Tokyo on 4th August 2009, aged 81. He was the Founding Head of the first Department of Statistical Science in Japan. "Now that he has left us forever, the world has lost one of its most innovative statisticians, the Japanese people have lost the finest statistician in their history and many of us a most noble friend" (Professor Howell Tong, from "The Obituary of Professor Hirotugu Akaike." *Journal of the Royal Statistical Society, Series A,* March, 2010). Professor Akaike had sent his Encyclopedia entry on May 14 2009, adding the following sentence in his email: "This is all that I could do under the present physical condition."

Cross References

►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Cp Statistic
►Kullback-Leibler Divergence
►Model Selection

Akaike's Information Criterion: Background, Derivation, Properties, and Refinements

Joseph E. Cavanaugh[1], Andrew A. Neath[2]
[1]Professor
The University of Iowa, Iowa City, IA, USA
[2]Professor
Southern Illinois University Edwardsville, Edwardsville, IL, USA

Introduction

The ▶Akaike Information Criterion, AIC, was introduced by Hirotogu Akaike in his seminal 1973 paper "Information Theory and an Extension of the Maximum Likelihood Principle." AIC was the first model selection criterion to gain widespread attention in the statistical community. Today, AIC continues to be the most widely known and used model selection tool among practitioners.

The traditional maximum likelihood paradigm, as applied to statistical modeling, provides a mechanism for estimating the unknown parameters of a model having a specified dimension and structure. Akaike extended this paradigm by considering a framework in which the model dimension is also unknown, and must therefore be determined from the data. Thus, Akaike proposed a framework wherein both model estimation and selection could be simultaneously accomplished.

For a parametric candidate model of interest, the likelihood function reflects the conformity of the model to the observed data. As the complexity of the model is increased, the model becomes more capable of adapting to the characteristics of the data. Thus, selecting the fitted model that maximizes the empirical likelihood will invariably lead one to choose the most complex model in the candidate collection. ▶Model selection based on the likelihood principle, therefore, requires an extension of the traditional likelihood paradigm.

Background

To formally introduce AIC, consider the following model selection framework. Suppose we endeavor to find a suitable model to describe a collection of response measurements y. We will assume that y has been generated according to an unknown density $g(y)$. We refer to $g(y)$ as the *true* or *generating model*.

A model formulated by the investigator to describe the data y is called a *candidate* or *approximating model*. We will assume that any candidate model structurally corresponds to a parametric class of distributions. Specifically, for a certain candidate model, we assume there exists a k-dimensional parametric class of density functions

$$\mathcal{F}(k) = \{f(y|\,\theta_k) \mid \theta_k \in \Theta(k)\},$$

a class in which the parameter space $\Theta(k)$ consists of k-dimensional vectors whose components are functionally independent.

Let $L(\theta_k\,|\,y)$ denote the likelihood corresponding to the density $f(y|\,\theta_k)$, i.e., $L(\theta_k\,|\,y) = f(y|\,\theta_k)$. Let $\hat{\theta}_k$ denote a vector of estimates obtained by maximizing $L(\theta_k\,|\,y)$ over $\Theta(k)$.

Suppose we formulate a collection of candidate models of various dimensions k. These models may be based on different subsets of explanatory variables, different mean and variance/covariance structures, and even different specifications for the type of distribution for the response variable. Our objective is to search among this collection for the fitted model that "best" approximates $g(y)$.

In the development of AIC, optimal approximation is defined in terms of a well-known measure that can be used to gauge the similarity between the true model $g(y)$ and a candidate model $f(y|\,\theta_k)$: the *Kullback–Leibler information* (Kullback and Leibler 1951; Kullback 1968). The Kullback–Leibler information between $g(y)$ and $f(y|\,\theta_k)$ with respect to $g(y)$ is defined as

$$I(\theta_k) = \mathrm{E}\left\{\log\frac{g(y)}{f(y|\,\theta_k)}\right\},$$

where $\mathrm{E}(\cdot)$ denotes the expectation under $g(y)$. It can be shown that $I(\theta_k) \geq 0$ with equality if and only if $f(y|\,\theta_k)$ is the same density as $g(y)$. $I(\theta_k)$ is not a formal metric, yet we view the measure in a similar manner to a distance: i.e., as the disparity between $f(y|\,\theta_k)$ and $g(y)$ grows, the magnitude of $I(\theta_k)$ will generally increase to reflect this separation.

Next, define

$$d(\theta_k) = \mathrm{E}\{-2\log f(y|\,\theta_k)\}.$$

We can then write

$$2I(\theta_k) = d(\theta_k) - \mathrm{E}\{-2\log g(y)\}.$$

Since $\mathrm{E}\{-2\log g(y)\}$ does not depend on θ_k, any ranking of a set of candidate models corresponding to values of $I(\theta_k)$ would be identical to a ranking corresponding to values of $d(\theta_k)$. Hence, for the purpose of discriminating among various candidate models, $d(\theta_k)$ serves as a valid substitute for $I(\theta_k)$. We will refer to $d(\theta_k)$ as the *Kullback discrepancy*.

To measure the separation between between a fitted candidate model $f(y|\,\hat{\theta}_k)$ and the generating model

$g(y)$, we consider the Kullback discrepancy evaluated at $\hat{\theta}_k$:

$$d(\hat{\theta}_k) = \mathrm{E}\{-2\log f(y|\,\theta_k)\}|_{\theta_k=\hat{\theta}_k}.$$

Obviously, $d(\hat{\theta}_k)$ would provide an attractive means for comparing various fitted models for the purpose of discerning which model is closest to the truth. Yet evaluating $d(\hat{\theta}_k)$ is not possible, since doing so requires knowledge of the true distribution $g(\cdot)$. The work of Akaike (1973, 1974), however, suggests that $-2\log f(y|\,\hat{\theta}_k)$ serves as a biased estimator of $d(\hat{\theta}_k)$, and that the bias adjustment

$$\mathrm{E}\{d(\hat{\theta}_k)\} - \mathrm{E}\{-2\log f(y|\,\hat{\theta}_k)\} \tag{1}$$

can often be asymptotically estimated by twice the dimension of θ_k.

Since k denotes the dimension of θ_k, under appropriate conditions, the expected value of

$$\mathrm{AIC} = -2\log f(y|\,\hat{\theta}_k) + 2k$$

will asymptotically approach the expected value of $d(\hat{\theta}_k)$, say

$$\Delta(k) = \mathrm{E}\{d(\hat{\theta}_k)\}.$$

Specifically, we will establish that

$$\mathrm{E}\{\mathrm{AIC}\} + o(1) = \Delta(k). \tag{2}$$

AIC therefore provides an asymptotically unbiased estimator of $\Delta(k)$. $\Delta(k)$ is often called the *expected Kullback discrepancy*.

In AIC, the empirical log-likelihood term $-2\log f(y|\,\hat{\theta}_k)$ is called the *goodness-of-fit term*. The bias correction $2k$ is called the *penalty term*. Intuitively, models which are too simplistic to adequately accommodate the data at hand will be characterized by large goodness-of-fit terms yet small penalty terms. On the other hand, models that conform well to the data, yet do so at the expense of containing unnecessary parameters, will be characterized by small goodness-of-fit terms yet large penalty terms. Models that provide a desirable balance between fidelity to the data and parsimony should correspond to small AIC values, with the sum of the two AIC components reflecting this balance.

Derivation

To justify AIC as an asymptotically unbiased estimator of $\Delta(k)$, we will focus on a particular candidate class $\mathcal{F}(k)$. For notational simplicity, we will suppress the dimension index k on the parameter vector θ_k and its estimator $\hat{\theta}_k$.

The justification of (2) requires the strong assumption that the true density $g(y)$ is a member of the candidate class $\mathcal{F}(k)$. Under this assumption, we may define a parameter vector θ_o having the same size as θ, and write $g(y)$ using the parametric form $f(y|\,\theta_o)$. The assumption that $f(y|\,\theta_o) \in \mathcal{F}(k)$ implies that the fitted model is either correctly specified or overfit.

To justify (2), consider writing $\Delta(k)$ as indicated:

$$\begin{aligned}
&\Delta(k)\\
&= \mathrm{E}\{d(\hat{\theta})\}\\
&= \mathrm{E}\{-2\log f(y|\,\hat{\theta})\}\\
&\quad + \left[\mathrm{E}\{-2\log f(y|\,\theta_o)\} - \mathrm{E}\{-2\log f(y|\,\hat{\theta})\}\right] \qquad (3)\\
&\quad + \left[\mathrm{E}\{d(\hat{\theta})\} - \mathrm{E}\{-2\log f(y|\,\theta_o)\}\right]. \qquad (4)
\end{aligned}$$

The following lemma asserts that (3) and (4) are both within $o(1)$ of k.

We assume the necessary regularity conditions required to ensure the consistency and ▸asymptotic normality of the maximum likelihood vector $\hat{\theta}$.

Lemma

$$\mathrm{E}\{-2\log f(y|\,\theta_o)\} - \mathrm{E}\{-2\log f(y|\,\hat{\theta})\} = k + o(1), \tag{5}$$

$$\mathrm{E}\{d(\hat{\theta})\} - \mathrm{E}\{-2\log f(y|\,\theta_o)\} = k + o(1). \tag{6}$$

Proof

Define

$$\mathrm{I}(\theta) = \mathrm{E}\left[-\frac{\partial^2 \log f(y|\,\theta)}{\partial\theta\partial\theta'}\right]$$

$$\text{and } \mathcal{I}(\theta, y) = \left[-\frac{\partial^2 \log f(y|\,\theta)}{\partial\theta\partial\theta'}\right].$$

$\mathrm{I}(\theta)$ denotes the *expected Fisher information matrix* and $\mathcal{I}(\theta, y)$ denotes the *observed Fisher information matrix*.

First, consider taking a second-order expansion of $-2\log f(y|\,\theta_o)$ about $\hat{\theta}$, and evaluating the expectation of the result. Since $-2\log f(y|\,\theta)$ is minimized at $\theta = \hat{\theta}$, the first-order term disappears, and we obtain

$$\begin{aligned}
\mathrm{E}\{-2\log f(y|\,\theta_o)\} &= \mathrm{E}\{-2\log f(y|\,\hat{\theta})\}\\
&\quad + \mathrm{E}\left\{(\hat{\theta} - \theta_o)'\{\mathcal{I}(\hat{\theta}, y)\}(\hat{\theta} - \theta_o)\right\}\\
&\quad + o(1).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathrm{E}\{-2\log f(y|\,\theta_o)\} - \mathrm{E}\{-2\log f(y|\,\hat{\theta})\}\\
&\quad = \mathrm{E}\left\{(\hat{\theta} - \theta_o)'\{\mathcal{I}(\hat{\theta}, y)\}(\hat{\theta} - \theta_o)\right\} + o(1). \qquad (7)
\end{aligned}$$

Next, consider taking a second-order expansion of $d(\hat{\theta})$ about θ_o, again evaluating the expectation of the

result. Since $d(\theta)$ is minimized at $\theta = \theta_o$, the first-order term disappears, and we obtain

$$\begin{aligned}\mathrm{E}\{d(\hat{\theta})\} &= \mathrm{E}\{-2\log f(y|\,\theta_o)\} \\ &\quad + \mathrm{E}\left\{(\hat{\theta}-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}-\theta_o)\right\} \\ &\quad + o(1).\end{aligned}$$

Thus,

$$\begin{aligned}&\mathrm{E}\{d(\hat{\theta})\} - \mathrm{E}\{-2\log f(y|\,\theta_o)\} \\ &\quad = \mathrm{E}\left\{(\hat{\theta}-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}-\theta_o)\right\} + o(1). \qquad (8)\end{aligned}$$

Recall that by assumption, $\theta_o \in \Theta(k)$. Therefore, the quadratic forms

$$(\hat{\theta}-\theta_o)'\{\mathcal{I}(\hat{\theta},y)\}(\hat{\theta}-\theta_o) \text{ and } (\hat{\theta}-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}-\theta_o)$$

both converge to centrally distributed chi-square random variables with k degrees of freedom. Thus, the expectations of both quadratic forms are within $o(1)$ of k. This fact along with (7) and (8) establishes (5) and (6). □

Properties

The previous lemma establishes that AIC provides an asymptotically unbiased estimator of $\Delta(k)$ for fitted candidate models that are correctly specified or overfit. From a practical perspective, AIC estimates $\Delta(k)$ with negligible bias in settings where n is large and k is comparatively small. In settings where n is small and k is comparatively large (e.g., $k \approx n/2$), $2k$ is often much smaller than the bias adjustment, making AIC substantially negatively biased as an estimator of $\Delta(k)$.

If AIC severely underestimates $\Delta(k)$ for higher dimensional fitted models in the candidate collection, the criterion may favor the higher dimensional models even when the expected discrepancy between these models and the generating model is rather large. Examples illustrating this phenomenon appear in Linhart and Zucchini (1986, 86–88), who comment (p. 78) that "in some cases the criterion simply continues to decrease as the number of parameters in the approximating model is increased."

AIC is *asymptotically efficient* in the sense of Shibata (1980, 1981), yet it is not *consistent*. Suppose that the generating model is of a finite dimension, and that this model is represented in the candidate collection under consideration. A consistent criterion will asymptotically select the fitted candidate model having the correct structure with probability one. On the other hand, suppose that the generating model is of an infinite dimension, and therefore lies outside of the candidate collection under consideration. An asymptotically efficient criterion will asymptotically select the fitted candidate model which minimizes the mean squared error of prediction.

From a theoretical standpoint, asymptotic efficiency is arguably the strongest optimality property of AIC. The property is somewhat surprising, however, since demonstrating the asymptotic unbiasedness of AIC as an estimator of the expected Kullback discrepancy requires the assumption that the candidate model of interest subsumes the true model.

Refinements

A number of AIC variants have been developed and proposed since the introduction of the criterion. In general, these variants have been designed to achieve either or both of two objectives: (1) to relax the assumptions or expand the setting under which the criterion can be applied, (2) to improve the small-sample performance of the criterion.

In the Gaussian linear regression framework, Sugiura (1978) established that the bias adjustment (1) can be exactly evaluated for correctly specified or overfit models. The resulting criterion, with a refined penalty term, is known as "corrected" AIC, or AICc. Hurvich and Tsai (1989) extended AICc to the frameworks of Gaussian nonlinear regression models and time series autoregressive models. Subsequent work has extended AICc to other modeling frameworks, such as autoregressive moving average models, vector autoregressive models, and certain ►generalized linear models and ►linear mixed models.

The Takeuchi (1976) information criterion, TIC, was derived by obtaining a general, large-sample approximation to each of (3) and (4) that does not rely on the assumption that the true density $g(y)$ is a member of the candidate class $\mathcal{F}(k)$. The resulting approximation is given by the trace of the product of two matrices: an information matrix based on the score vector, and the inverse of an information matrix based on the Hessian of the log likelihood. Under the assumption that $g(y) \in \mathcal{F}(k)$, the information matrices are equivalent. Thus, the trace reduces to k, and the penalty term of TIC reduces to that of AIC.

Bozdogon (1987) proposed a variant of AIC that corrects for its lack of consistency. The variant, called CAIC, has a penalty term that involves the log of the determinant of an information matrix. The contribution of this term leads to an overall complexity penalization that increases with the sample size at a rate sufficient to ensure consistency.

Pan (2001) introduced a variant of AIC for applications in the framework of generalized linear models fitted

using generalized estimating equations. The criterion is called QIC, since the goodness-of-fit term is based on the empirical quasi-likelihood.

Konishi and Kitagawa (1996) extended the setting in which AIC has been developed to a general framework where (1) the method used to fit the candidate model is not necessarily maximum likelihood, and (2) the true density $g(y)$ is not necessarily a member of the candidate class $\mathcal{F}(k)$. Their resulting criterion is called the generalized information criterion, GIC. The penalty term of GIC reduces to that of TIC when the fitting method is maximum likelihood.

AIC variants based on computationally intensive methods have also been proposed, including cross-validation (Stone 1977; Davies et al. 2005), bootstrapping (Ishiguro et al. 1997; Cavanaugh and Shumway 1997; Shibata 1997), and Monte Carlo simulation (Hurvich et al. 1990; Bengtsson and Cavanaugh 2006). These variants tend to perform well in settings where the sample size is small relative to the complexity of the models in the candidate collection.

About the Authors

Joseph E. Cavanaugh is Professor of Biostatistics and Professor of Statistics and Actuarial Science at The University of Iowa. He is an associate editor of the *Journal of the American Statistical Association* (2005–present) and the *Journal of Forecasting* (1999–present). He has published over 60 refereed articles.

Andrew Neath is a Professor of Mathematics and Statistics at Southern Illinois University Edwardsville. He has been recognized for his work in science education. He is an author on numerous papers, merging Bayesian views with model selection ideas. He wishes to thank Professor Miodrag Lovric for the honor of an invitation to contribute to a collection containing the works of so many notable statisticians.

Cross References

►Akaike's Information Criterion
►Cp Statistic
►Kullback-Leibler Divergence
►Model Selection

References and Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F (eds) Proceedings of the 2nd International symposium on information theory. Akadémia Kiadó, Budapest, pp 267–281

Akaike H (1974) A new look at the statistical model identification. IEEE T Automat Contra AC-19:716–723

Bengtsson T, Cavanaugh JE (2006) An improved Akaike information criterion for state-space model selection. Comput Stat Data An 50:2635–2654

Bozdogan H (1987) Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52:345–370

Cavanaugh JE, Shumway RH (1997) A bootstrap variant of AIC for state-space model selection. Stat Sinica 7:473–496

Davies SL, Neath AA, Cavanaugh JE (2005) Cross validation model selection criteria for linear regression based on the Kullback-Leibler discrepancy. Stat Methodol 2:249–266

Hurvich CM, Shumway RH, Tsai CL (1990) Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. Biometrika 77:709–719

Hurvich CM, Tsai CL (1989) Regression and time series model selection in small samples. Biometrika 76:297–307

Ishiguro M, Sakamoto Y, Kitagawa G (1997) Bootstrapping log likelihood and EIC, an extension of AIC. Ann I Stat Math 49:411–434

Konishi S, Kitagawa G (1996) Generalised information criteria in model selection. Biometrika 83:875–890

Kullback S (1968) Information Theory and Statistics. Dover, New York

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:76–86

Linhart H, Zucchini W (1986) Model selection. Wiley, New York

Pan W (2001) Akaike's information criterion in generalized estimating equations. Biometrics 57:120–125

Shibata R (1980) Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Ann Stat 80:147–164

Shibata R (1981) An optimal selection of regression variables. Biometrika 68:45–54

Shibata R (1997)Bootstrap estimate of Kullback-Leibler information for model selection. Stat Sinica 7:375–394

Stone M (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J R Stat Soc B 39:44–47

Sugiura N (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat A7:13–26

Takeuchi K (1976) Distribution of information statistics and criteria for adequacy of models. Mathematical Sciences 153:12–18(in Japanese)

Algebraic Statistics

Sonja Petrović[1], Aleksandra B. Slavković[2]
[1]Research Assistant Professor
University of Illinois at Chicago, Chicago, IL, USA
[2]Associate Professor
The Pennsylvania State University, University Park, PA, USA

Algebraic statistics applies concepts from algebraic geometry, commutative algebra, and geometric combinatorics to better understand the structure of statistical models, to

improve statistical inference, and to explore new classes of models. Modern algebraic geometry was introduced to the field of statistics in the mid 1990s. Pistone and Wynn (1996) used Gröbner bases to address the issue of confounding in design of experiments, and Diaconis and Sturmfels (1998) used them to perform exact conditional tests. The term *algebraic statistics* was coined in the book by Pistone et al. (2001), which primarily addresses experimental design. The current algebraic statistics literature includes work on contingency tables, sampling methods, graphical and latent class models, and applications in areas such as statistical disclosure limitation (e.g., Dobra et al. (2009)), and computational biology and phylogenetics (e.g., Pachter and Sturmfels (2005)).

Algebraic Geometry of Statistical Models

Algebraic geometry is a broad subject that has seen an immense growth over the past century. It is concerned with the study of algebraic varieties, defined to be (closures of) solution sets of systems of polynomial equations. For an introduction to computational algebraic geometry and commutative algebra, see Cox et al. (2007).

Algebraic statistics studies statistical models whose parameter spaces correspond to real positive parts of algebraic varieties. To demonstrate how this correspondence works, consider the following simple example of the independence model of two binary random variables, X and Y, such that joint probabilities are arranged in a 2×2 matrix $p := [p_{ij}]$. The model postulates that the joint probabilities factor as a product of marginal distributions: $p_{ij} = p_{i+}p_{+j}$, where $i, j \in \{1, 2\}$. This is referred to as an *explicit* algebraic statistical model. Equivalently, the matrix p is of rank 1, that is, its 2×2 determinant is zero: $p_{11}p_{22} - p_{12}p_{21} = 0$. This is referred to as an *implicit* description of the independence model. In algebraic geometry, the set of rank-1 matrices, where we allow p_{ij} to be arbitrary complex numbers, is a classical object called a *Segre variety*. Thus, the independence model is the real positive part of the Segre variety. Exponential family models, in general, correspond to *toric varieties*, whose implicit description is given by a set of binomials. For a broad, general definition of algebraic statistical models, see Drton and Sullivant (2007).

By saying that "we understand the algebraic geometry of a model," we mean that we understand some basic information about the corresponding variety, such as: degree, dimension and codimension (i.e., degrees of freedom); the defining equations (i.e., the implicit description of the model); the singularities (i.e., degeneracy in the model). The current algebraic statistics literature demonstrates that understanding the geometry of a model can be useful for statistical inference (e.g., exact conditional inference, goodness-of-fit testing, parameter identifiability, and maximum likelihood estimation). Furthermore, many relevant questions of interest in statistics relate to classical open problems in algebraic geometry.

Algebraic Statistics for Contingency Tables

A paper by Diaconis and Sturmfels (1998) on algebraic methods for discrete probability distributions stimulated much of the work in algebraic statistics on contingency tables, and has led to two general classes of problems: (1) algebraic representation of a statistical model, and (2) conditional inference. The algebraic representation of the independence model given above generalizes to any k-way table and its corresponding hierarchical log-linear models (e.g., see Dobra et al. (2009)). A standard reference on log-linear models is Bishop et al. (1975).

Most of the algebraic work for contingency tables has focused on geometric characterizations of log-linear models and estimation of cell probabilities under those models. Algebraic geometry naturally provides an explicit description of the closure of the parameter space. This feature has been utilized, for example, by Eriksson et al. (2006) to describe polyhedral conditions for the nonexistence of the MLE for log-linear models. More recently, Petrović et al. (2010) provide the first study of algebraic geometry of the p_1 random graph model of Holland and Leinhardt (1981).

Conditional inference relies on the fact that data-dependent objects are a convex bounded set, $P_t = \{\mathbf{x} : x_i \in \mathbb{R}_{\geq 0}, \mathbf{t} = \mathbf{A}\mathbf{x}\}$, where x is a table, A is a design matrix, and $\mathbf{t}$ a vector of constraints, typically margins, that is, sufficient statistics of a log-linear model. The set of all integer points inside P_t is referred to as a *fiber*, which is the support of the conditional distribution of tables given $\mathbf{t}$, or the so-called *exact distribution*. Characterization of the fiber is crucial for three statistical tasks: counting, sampling and optimization. Diaconis and Sturmfels (1998) provide one of the fundamental results in algebraic statistics regarding sampling from exact distributions. They define a Markov basis, a set of integer valued vectors in the kernel of A, which is a smallest set of moves needed to perform a ▶random walk over the space of tables and to guarantee connectivity of the chain. In Hara et al. (2010), for example, the authors use Markov bases for exact tests in a multiple logistic regression. The earliest application of Markov bases, counting and optimization was in the area of statistical disclosure limitation for exploring issues of confidentiality with the release of contingency table data; for an overview,

see Dobra et al. (2009), and for other related topics, see Chen et al. (2006), Onn (2006), and Slavković and Lee (2009).

Graphical and Mixture Models

Graphical models (e.g., Lauritzen (1996)) are an active research topic in algebraic statistics. Non-trivial problems, for example, include complete characterization of Markov bases for these models, and counting the number of solutions of their likelihood equations. Geiger et al. (2006) give a remarkable result in this direction: *decomposable* graphical models are precisely those whose Markov bases consist of squarefree quadrics, or, equivalently, those graphical models whose maximum likelihood degree is 1. More recently, Feliz et al. (2010) made a contribution to the mathematical finance literature by proposing a new model for analyzing default correlation.

►Mixture models, including latent class models, appear frequently in statistics, however, standard asymptotics theory often does not apply due to the presence of singularities (e.g., see Watanabe (2009)). Singularities are created by marginalizing (smooth) models; geometrically, this is a projection of the corresponding variety. Algebraically, mixture models correspond to *secant varieties*. The complexity of such models presents many interesting problems for algebraic statistics; e.g., see Fienberg et al. (2009) for the problems of maximum likelihood estimation and parameter identifiability in latent class models. A further proliferation of algebraic statistics has been supported by studying mixture models in phylogenetics (e.g., see Allman et al. (2010)), but many questions about the geometry of these models still remain open.

Further Reading

There are many facets of algebraic statistics, including generalizations of classes of models discussed above: experimental design, continuous multivariate problems, and new connections between algebraic statistics and information geometry. For more details see Putinar and Sullivant (2008), Drton et al. (2009), Gibilisco et al. (2009), and references given therein. Furthermore, there are many freely available algebraic software packages (e.g., 4ti2 (4ti2 team), CoCoA (CoCoATeam)) that can be used for relevant computations alone, or in combination with standard statistical packages.

Acknowledgments

Supported in part by National Science Foundation grant SES-0532407 to the Department of Statistics, Pennsylvania State University.

Cross References

►Categorical Data Analysis
►Confounding and Confounder Control
►Degrees of Freedom
►Design of Experiments: A Pattern of Progress
►Graphical Markov Models
►Logistic Regression
►Mixture Models
►Statistical Design of Experiments (DOE)
►Statistical Inference
►Statistical Inference: An Overview

References and Further Reading

4ti2 team. 4ti2 – a software package for algebraic, geometric and combinatorial problems on linear spaces. http://WWW.4ti2.de

Allman E, Petrović S, Rhodes J, Sullivant S (2010) Identifiability of two-tree mixtures under group-based models. IEEE/ACM Trans Comput Biol Bioinfor. In press

Bishop YM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Cambridge, MA (Reprinted by Springer, 2007)

Chen Y, Dinwoodie I, Sullivant S (2006) Sequential importance sampling for multiway tables. Ann Stat 34(1):523–545

CoCoATeam. CoCoA: a system for doing computations in commutative algebra. http://cocoa.dima.unige.it

Cox D, Little J, O'Shea D (2007) Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra, 3rd edn. Springer, New York

Diaconis P, Sturmfels B (1998) Algebraic algorithms for sampling from conditional distributions. Ann Stat 26:363–397

Dobra A, Fienberg SE, Rinaldo A, Slavković A, Zhou Y(2009) Algebraic statistics and contingency table problems: estimations and disclosure limitation. In: Emerging Applications of Algebraic Geometry: IMA volumes in mathematics and its applications, 148:63–88

Drton M, Sturmfels B, Sullivant S (2009) Lectures on algebraic statistics, vol39. Oberwolfach seminars, Birkhäuser

Eriksson N, Fienberg SE, Rinaldo A, Sullivant S (2006) Polyhedral conditions for the nonexistence of the mle for hierarchical log-liner models. J Symb Comput 41(2):222–233

Feliz I, Guo X, Morton J, Sturmfels B (2010) Graphical models for correlated default. Math Financ (in press)

Fienberg SE, Hersh P, Zhou Y (2009) Maximum likelihood estimation in latent class models for contingency table data. In: Gibilisco P, Riccomagno E, Rogantin M, Wynn H (eds) Algebraic and geometric methods in statistics. Cambridge University Press, London, pp 27–62

Geiger D, Meek C, Sturmfels B (2006) On the toric algebra of graphical models. Ann Stat 34(3):1463–1492

Gibilisco P, Riccomagno E, Rogantin M, Wynn H (2009) Algebraic and geometric methods in statistics, Cambridge University press

Hara H, Takemura A, Yoshida R (2010) On connectivity of fibers with positive marginals in multiple logistic regression. J Multivariate Anal 101(4):909–925

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs (with discussion). J Am Stat Assoc 76:33–65

Lauritzen SL (2006) Graphical models. Clarendon, Oxford
Onn S (2006) Entry uniqueness in margined tables. Lect Notes Comput Sci 4302:94–101
Pachter L, Sturmfels B (2005) Algebraic statistics for computational biology. Cambridge University Press, New York, NY
Petrović S, Rinaldo A, Fienberg SE (2010) Algebraic statistics for a directed random graph model with reciprocation. In: Viana MAG, Wynn H (eds) Algebraic methods in statistics and probability, II, Contemporary Mathematics. Am Math Soc 516
Pistone G, Wynn H (1996) Generalised confounding with Gröbner bases. Biometrika 83(3):653–666
Pistone G, Riccomagno E, Wynn H (2001) Algebraic statistics: computational commutative algebra in statistics. CRC, Boca Raton
Putinar M, Sullivant S (2008) Emerging applications of algebraic geometry. Springer, Berlin
Slavković AB, Lee J (2010) Synthetic two-way contingency tables that preserve conditional frequencies. Stat Methodal 7(3): 225–239
Watanabe S (2009) Algebraic geometry and statistical learning theory: Cambridge monographs on applied and computational mathematics, 25, NewYork, Cambridge University Press

Almost Sure Convergence of Random Variables

Herold Dehling
Professor
Ruhr-Universität Bochum, Bochum, Germany

Definition and Relationship to Other Modes of Convergence

Almost sure convergence is one of the most fundamental concepts of convergence in probability and statistics. A sequence of random variables $(X_n)_{n\geq 1}$, defined on a common probability space $(\Omega, \mathcal{F}, P)$, is said to converge almost surely to the random variable X, if

$$P(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1.$$

Commonly used notations are $X_n \xrightarrow{a.s.} X$ or $\lim_{n\to\infty} X_n = X\,(a.s.)$. Conceptually, almost sure convergence is a very natural and easily understood mode of convergence; we simply require that the sequence of numbers $(X_n(\omega))_{n\geq 1}$ converges to $X(\omega)$ for almost all $\omega \in \Omega$. At the same time, proofs of almost sure convergence are usually quite subtle.

There are rich connections of almost sure convergence with other classical modes of convergence, such as convergence in probability, defined by $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$ for all $\epsilon > 0$, convergence in distribution, defined by $\lim_{n\to\infty} Ef(X_n) = Ef(X)$ for all real-valued bounded, continuous functions f, and convergence in L_p, defined by $\lim_{n\to\infty} E|X_n - X|^p = 0$. Almost sure convergence implies convergence in probability, which again implies convergence in distribution, but not vice versa. Almost sure convergence neither implies nor is it implied by convergence in L_p. A standard counterexample, defined on the probability space $[0,1]$, equipped with the Borel σ-field and Lebesgue measure, is the sequence $X_n(\omega) = 1_{[\frac{j}{2^k}, \frac{j+1}{2^k}]}(\omega)$, if $n = 2^k + j$, $k \geq 0$, $0 \leq j < 2^k$. The sequence $(X_n)_{n\geq 1}$ converges to zero in probability and in L_p, but not almost surely. On the same probability space, the sequence defined by $X_n = n^{1/p} 1_{[0,\frac{1}{n}]}$ provides an example that converges to zero almost surely, but not in L_p.

Although convergence in probability does not imply almost sure convergence, there is a partial result in this direction. If $(X_n)_{n\geq 1}$ converges in probabilty to X, one can find a subsequence $(n_k)_{k\geq 1}$ such that $X_{n_k} \xrightarrow{a.s.} X$.

Skorohod's almost sure representation theorem is a partial converse to the fact that almost sure convergence implies convergence in distribution. If $(X_n)_{n\geq 1}$ converges in distribution to X, one can find a sequence of random variables $(Y_n)_{n\geq 1}$ and a random variable Y such that X_n and Y_n have the same distribution, for each n, X and Y have the same distribution, and $\lim_{n\to\infty} Y_n = Y$ almost surely. Originally proved by Skorohod (1956) for random variables with values in a separable metric space, this representation theorem has been extended by Dudley (1968) to noncomplete spaces and later by Wichura (1970) to nonseparable spaces.

By some standard arguments, one can show that almost sure convergence of $(X_n)_{n\geq 1}$ to X is equivalent to

$$\lim_{n\to\infty} P(\sup_{k\geq n} |X_k - X| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0.$$

Thus almost sure convergence holds, if the series $\sum_{k\geq 1} P(|X_k - X| \geq \epsilon)$ converges. In this case, the sequence $(X_n)_{n\geq 1}$ is said to *converge completely* to X.

Important Almost Sure Convergence Theorems

Historically the earliest and also the best known almost sure convergence theorem is the *Strong Law of Large Numbers*, established originally by Borel (1909). Given an i.i.d. sequence $(X_k)_{k\geq 1}$ of random variables that are uniformly distributed on $[0,1]$, Borel showed that

$$\frac{1}{n} S_n \xrightarrow{a.s.} E(X_1),$$

where $S_n := \sum_{k=1}^n X_k$ denotes the partial sum. Later, this was generalized to sequences with arbitrary distributions. Finally, Kolmogorov (1930) could show that the existence of first moments is a necessary and sufficient condition for the strong law of large numbers for i.i.d. random variables.

Hsu and Robbins (1947) showed complete convergence in the law of large numbers, provided the random variables have finite second moments; Baum and Katz (1965) showed that this condition is also necessary.

Birkhoff (1931) proved the *Ergodic Theorem*, i.e., the validity of the strong law of large numbers for stationary ergodic sequences $(X_k)_{k\geq 1}$ with finite first moments. Kingman (1968) generalized this to the *Subadditive Ergodic Theorem*, valid for doubly indexed subadditive process $(X_{s,t})$ satisfying a certain moment condition. Doob (1953) established the *Martingale Convergence Theorem*, which states that every L_1-bounded submartingale converges almost surely.

The *Marcinkiewicz-Zygmund Strong Law of Large Numbers* (1937) is a sharpening of the law of large numbers for partial sums of i.i.d. random variables, stating that for $1 \leq p < 2$ we have

$$\lim_{n\to\infty} \frac{1}{n^{1/p}} \sum_{k=1}^{n} (X_k - E(X_k)) = 0 \text{ a.s.},$$

if and only if the random variables have finite p-th moments. Note that for $p = 2$ this result is false as it would contradict the central limit theorem (see ▶Central Limit Theorems).

For i.i.d. random variables with finite variance $\sigma^2 \neq 0$, Hartman and Wintner (1941) proved the *Law of the Iterated Logarithm*, stating that

$$\limsup_{n\to\infty} \frac{1}{\sqrt{2\sigma^2 n \log\log n}} \sum_{k=1}^{n} (X_i - E(X_1)) = 1 \text{ a.s.},$$

and that the corresponding lim inf equals −1. In the special case of a symmetric ▶random walk, this theorem had been established earlier by Khintchin (1924). The law of the iterated logarithm gives a very precise information about the behavior of the centered partial sum.

Strassen (1964) proved the *Functional Law of the Iterated Logarithm*, which concerns the normalized partial sum process, defined by

$$f_n(\frac{k}{n}) := \frac{1}{\sqrt{2\sigma^2 n \log\log n}} \sum_{i=1}^{k} (X_i - E(X_i)), 0 \leq k \leq n,$$

and linearly interpolated in between. The random sequence of functions $(f_n)_{n\geq 1}$ is almost surely relatively compact and has the following set of limit points

$$K = \{x \in C[0,1] : x \text{ is absolutely continuous and } \int_0^1 (x'(t))^2 dt \leq 1\}.$$

The functional law of the iterated logarithm gives a remarkably sharp information about the behavior of the partial sum process.

The *Almost Sure Invariance Principle*, originally established by Strassen (1964) is an important technical tool in many limit thorems. Strassen's theorem states that for i.i.d. random variables with finite variance, one can define a standard Brownian motion (see ▶Brownian Motion and Diffusions) $W(k)$ satisfying

$$\sum_{k=1}^{n} (X_k - E(X_1)) - \sigma W(n) = o(\sqrt{n \log\log n}), \text{a.s.}.$$

Komlos et al. (1975) gave a remarkable sharperning of the error term in the almost sure invariance principle, showing that for $p > 2$ one can find a standard Brownian motion $(W_t)_{t\geq 0}$ satisfying

$$\sum_{k=1}^{n} (X_k - E(X_1)) - \sigma W(n) = o(n^{1/p}), \text{a.s.}.$$

if and only if the random variables have finite p-th moments. In this way, results that hold for Brownian motion can be carried over to the partial sum process. E.g., many limit theorems in the statistical analysis of change-points are proved by a suitable application of strong approximations.

In the 1980s, Brosamler, Fisher and Schatte independently discovered the *Almost Sure Central Limit Theorem*, stating that for partial sums $S_k := \sum_{i=1}^{k} X_i$ of an i.i.d. sequence $(X_i)_{i\geq 1}$ with mean zero and variance σ^2

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{k=1}^{n} \frac{1}{k} 1_{\{S_k/\sigma\sqrt{k} \leq x\}} = \Phi(x),$$

where $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ denotes the standard normal distribution function. The remarkable feature of this theorem is that one can observe the central limit theorem, which in principle is a distributional limit theorem, along a single realization of the process.

In 1933, Glivenko and Cantelli independently discovered a result that is now known as the *Glivenko–Cantelli Theorem* (see ▶Glivenko-Cantelli Theorems). Given a sequence $(X_k)_{k\geq 1}$ of i.i.d random variables with distribution function $F(x) := P(X \leq x)$, we define the empirical distribution function $F_n(x) = \frac{1}{n} \sum_{k=1}^{n} 1_{\{X_k \leq x\}}$. The Glivenko–Cantelli theorem states that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{a.s.} 0.$$

This theorem is sometimes called the fundamental theorem of statistics, as it shows that it is possible to recover the

distribution of a random variable from a sequence of observations.

Almost sure convergence has been established for U-statistics, a class of sample statistics of great importance in mathematical statistics. Given a symmetric kernel $h(x, y)$, we define the bivariate U-statistic

$$U_n := \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} h(X_i, X_j).$$

Hoeffding (1961) proved the *U-Statistic Strong Law of Large Numbers*, stating that for any integrable kernel and i.i.d. random variables $(X_i)_{i \ge 1}$,

$$U_n \xrightarrow{a.s.} Eh(X_1, X_2).$$

Aaronson et al. (1996) established the corresponding *U-Statistic Ergodic Theorem*, albeit under extra conditions. The *U-statistic Law of the Iterated Logarithm*, in the case of i.i.d. data (X_i) was established by Sen (1972). In the case of degenerate kernels, i.e., kernels satisfying $Eh(x, X_1) = 0$, for all x, this was sharpened by Dehling et al. (1985) and Dehling (1989). Their *Degenerate U-Statistic Law of the Iterated Logarithm* states that

$$\limsup_{n \to \infty} \frac{1}{n \log \log n} \sum_{1 \le i < j \le n} h(X_i, X_j) = c_h, \quad \text{a.s.},$$

where c_h is the largest eigenvalue (see ►Eigenvalue, Eigenvector and Eigenspace) of the integral operator with kernel $h(x, y)$. A functional version as well as an almost sure invariance principle were established by the same authors.

Proofs of Almost Sure Convergence

In most situations, especially in applications in Statistics, almost sure convergence is proved by identifying a given sequence as a a continuous function of a sequence of a type studied in one of the basic theorems on almost sure convergence.

The proofs of the basic almost sure convergence theorems are quite subtle and require a variety of technical tools, such as exponential inequalities, maximal inequalities, truncation techniques and the Borel-Cantelli lemma (see ►Borel–Cantelli Lemma and Its Generalizations).

About the Author

Herold Dehling (born 1954 in Westrhauderfehn, Germany) is Professor of Mathematics at the Ruhr-Universität Bochum, Germany. From 1988 to 2000, he was on the faculty of the University of Groningen, The Netherlands. Prior to that he held postdoc positions at Boston University and at the University of Göttingen. Herold Dehling studied Mathematics at the University of Göttingen and the University of Illinois at Urbana-Champaign. He obtained his Ph.D. in 1981 at Göttingen. Herold Dehling is an elected member of the International Statistical Institute. In 2005 he was awarded the Prix Gay-Lussac–Humboldt of the Republic of France. Herold Dehling conducts research in the area of asymptotic methods in probability and statistics, with special emphasis on dependent processes. He has published more than 75 research papers in probability and statistics. Herold Dehling is co-author of three books, *Kansrekening* (Epsilon Uitgaven, Utrecht 2005, with J. N. Kalma), *Einführung in die Wahrscheinlichkeitsrechnung und Statistik* (Springer, Heidelberg 2004, with B. Haupt) and *Stochastic modelling in process technology* (Elsevier Amsterdam 2007, with T. Gottschalk and A. C. Hoffmann). Moreover, he is coeditor of the books *Empirical Process Techniques for Dependent Data* (Birkhäuser, Boston 2002, with T. Mikosch and M. Sorensen) and *Weak Dependence in Probability, Analysis and Number Theory* (Kendrick Press, Utah 2010, with I. Berkes, R. Bradley, M. Peligrad and R. Tichy).

Cross References

►Brownian Motion and Diffusions
►Convergence of Random Variables
►Ergodic Theorem
►Random Variable
►Weak Convergence of Probability Measures

References and Further Reading

Aaronson J, Burton RM, Dehling H, Gilat D, Hill T, Weiss B (1996) Strong laws for L- and U-statistics. Trans Am Math Soc 348:2845–2866

Baum LE, Katz M (1965) Convergence rates in the law of large numbers. Trans Am Math Soc 120:108–123

Birkhoff GD (1931) Proof of the Ergodic theorem. Proc Nat Acad Sci USA 17:656–660

Borel E (1909) Les probabilités dénombrables et leurs application arithmétique. Rendiconti Circolo Mat Palermo 27: 247—271

Brosamler G (1988) An almost everywhere central limit theorem. Math Proc Cambridge Philos Soc 104:561–574

Cantelli FP (1933) Sulla determinazione empirica della leggi di probabilita. Gior Ist Ital Attuari 4:421–424

Csörgo M, Révész P (1981) Strong approximations in probability and statistics. Academic, New York

Dehling H, Denker M, Philipp W (1985) Invariance principles for von Mises and U-Statistics. Z Wahrsch verw Geb 67: 139–167

Dehling H (1989) The functional law of the iterated logarithm for von-Mises functionals and multiple Wiener integrals. J Multiv Anal 28:177–189

Dehling H (1989) Complete convergence of triangular arrays and the law of the iterated logarithm for U-statistics. Stat Prob Lett 7:319–321

Doob JL (1953) Stochastic processes. Wiley, New York

Dudley RM (1968) Distances of probability measures and random variables. Ann Math Stat 39:1563–1572
Fisher A (1989) Convex invariant means and a pathwise central limit theorem. Adv Math 63:213–246
Glivenko VI (1933) Sulla determinazione empirica della leggi di probabilita. Gior Ist Ital Attuari 4:92–99
Hartmann P, Wintner A (1941) On the law of the iterated logarithm. Am J Math 63:169–176
Hoeffding W (1961) The strong law of large numbers for U-statistics. University of North Carolina, Institute of Statistics Mimeograph Series 302
Hsu PL, Robbins H (1947) Complete convergence and the law of large numbers. Proc Nat Acad Sci USA 33:25–31
Khintchin A (1924) Über einen Satz der Wahrscheinlichkeitsrechnung. Fund Math 6:9–20
Kingman JFC (1968) The ergodic theory of subadditive stochastic processes. J R Stat Soc B 30:499–510
Kolmogorov AN (1930) Sur la loi forte des grandes nombres. Comptes Rendus Acad Sci Paris 191:910–912
Komlos J, Major P, Tusnady G (1975) An approximation of partial sums of independent RVs and the sample DF I. Z Wahrsch verw Geb 32:111–131
Marcinkiewicz J, Zygmund A (1937) Sur les fonctions indépendantes. Fund Math 29:60–90
Schatte P (1988) On strong versions of the central limit theorem. Math Nachr 137:249–256
Sen PK (1972) Limiting behavior of regular functionals of empirical distributions for stationary mixing processes. Z Wahrsch verw Geb 25:71–82
Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York
Skorohod AV (1956) Limit theorems for stochastic processes. Theory Prob Appl 1:261–290
Stout WF (1974) Almost sure convergence. Academic, New York
Strassen V (1964) An invariance principle for the law of the iterated logarithm. Z Wahrsch verw Geb 3:211–226
Van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge
Wichura MJ (1970) On the construction of almost uniformly convergent random variables with given weakly convergent image laws. Ann Math Stat 41:284–291

Analysis of Areal and Spatial Interaction Data

Gunter Spöck[1], Jürgen Pilz[2]
[1]Associate Professor
University of Klagenfurt, Klagenfurt, Austria
[2]Professor
University of Klagenfurt, Klagenfurt, Austria

Areal Data

Areal data y_i are data that are assigned to spatial regions A_i, $i = 1, 2, \ldots, n$. Such data and spatial areas naturally arise at different levels of spatial aggregation, like data assigned to countries, counties, townships, political districts, constituencies or other spatial regions that are featured by more or less natural boundaries. Examples for data y_i might be the number of persons having a certain chronic illness, number of enterprises startups, average income, population density, number of working persons, area of cultivated land, air pollution, etc. Like all spatial data, also areal data are marked by the fact that they exert spatial correlation to the data from neighboring areas. Tobler (1970) expresses this in his first law of geography: "everything is related to everything else, but near things are more related than distant things." It is this spatial correlation which is investigated, modeled and taken into account in the analysis of areal data.

Spatial proximity matrix. A mathematical tool that is common to almost all areal analysis methods is the so-called $(n \times n)$ spatial proximity matrix $\mathbf{W}$, each of whose elements, w_{ij}, represents a measure of spatial proximity of area A_i and area A_j. According to Bailey and Gatrell (1995) some possible criteria might be:

- $w_{ij} = 1$ if A_j shares a common boundary with A_i and $w_{ij} = 0$ else.
- $w_{ij} = 1$ if the centroid of A_j is one of the k nearest centroids to that of A_i and $w_{ij} = 0$ else.
- $w_{ij} = d_{ij}^{\gamma}$ if the inter-centroid distance $d_{ij} < \delta$ $(\delta > 0, \gamma < 0)$; and $w_{ij} = 0$ else.
- $w_{ij} = \frac{l_{ij}}{l_i}$, where l_{ij} is the length of common boundary between A_i and A_j and l_i is the perimeter of A_i.

All diagonal elements w_{ii} are set to 0. The spatial proximity matrix $\mathbf{W}$ must not be symmetric. For instance, case 2 and case 4 above lead to asymmetric proximity matrices. For more proximity measures we refer to Bailey and Gatrell (1995) and any other textbook on areal spatial analysis like Anselin (1988).

Spatial Correlation Measures

Global measures of spatial correlation. The global Moran index I, first derived by Moran (1950), is a measure for spatial correlation of areal data having proximity matrix $\mathbf{W}$. Defining $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ and $\bar{y}$, the mean of the data y_i, $i = 1, 2, \ldots, n$, the global Moran index may be written

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n} (y_i - \bar{y})^2}. \qquad (1)$$

Thus the global Moran index may be be interpreted as measuring correlation between $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ and the spatial lag-variable $\mathbf{Wy}$. But the Moran index does not necessarily take values between -1 and 1. Its expectation for independent data y_i is $\mathrm{E}[I] = -\frac{1}{n-1}$. Values of the Moran index larger than this value thus are an indication of

positive global spatial correlation; values smaller than this value indicate negative spatial correlation.

A global correlation measure similar to the variogram known from classical geostatistics is the Geary-index (Geary's c, Geary 1954):

$$c = \frac{n-1}{2S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2}$$

Under the independence assumption for the $\mathbf{y_i}$ its expectation is $\mathrm{E}[c] = 1$. Values of c larger than 1 indicate negative correlation and values smaller than 1 positive correlation.

The significance for Moran's I and Geary's c may be tested by means of building all $n!$ permutations of the y_i, $i = 1, 2, \ldots, n$, assigning them to the the different areas A_j, $j = 1, 2, \ldots, n$, calculating for each permutation Moran's I or Geary's c and then considering the distributions of these permuted spatial correlation statistics. True correlation statistics at the lower or upper end of these distributions are an indication of significance of the global correlation measures.

A map often useful for detecting spatial clusters of high or low values is the so-called LISA map. It may be shown that Moran's I is exactly the upward slope of the regression line between the regressors $(\mathbf{y} - \mathbf{1_n}\bar{y})$ and the spatial lag-variables $\mathbf{W}(\mathbf{y} - \mathbf{1_n}\bar{y})$ as responses, where the matrix $\mathbf{W}$ is here standardized to have rows which sum up to one. The corresponding scatterplot has four quadrants PP, NN, PN and NP, with P and N indicating positive and negative values for the regressors and responses. If one codes these four classes into which the pairs $[y_i - \bar{y}, \sum_{j=1}^{n} w_{ij}(y_j - \bar{y})]$ may fall with colors and visualizes these colors in a map of the areas one can easily detect clusters of areas that are surrounded by low or high neighboring values.

Both statistics, the Moran I and Geary's c make a global assumption of second order stationarity, meaning that the y_i, $i = 1, 2, \ldots, n$ all have the same constant mean and variance. If one doubts that this condition is fully met one has to rely on local measures of spatial correlation, for local versions of Moran's I and Geary's c see Anselin (1995).

Spatial Linear Regression

A problem frequently occuring in areal data analysis is the regression problem. Response variables y_i and corresponding explanatory vectors $\mathbf{x_i}$ are observed in spatial areas A_i, $i = 1, 2, \ldots, n$ and one is interested in the linear regression relationship $y_i \approx \mathbf{x_i}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an unknown regression parameter vector to be estimated. Subsuming all row vectors $\mathbf{x_i}^T$ in the $(n \times p)$ design matrix $\mathbf{X}$ and writing $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ the ordinary ▶least squares solution to this regression problem, which does not take account of spatial correlation, is known to be $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. If the data in $\mathbf{y}$ are known to be correlated the above ordinary least squares estimator is known to be inefficient and statistical significance tests in this regression model are known to be misleading. Problems may be resolved by considering the generalized least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{y}$, where the covariance matrix Σ is measuring the correlation between the data in $\mathbf{y}$. All regression procedures used in areal data analysis deal more or less with the modeling and estimation of this covariance structure Σ and the estimation of β. In all subsequent sections we will assume that the spatial proximity matrix $\mathbf{W}$ is standardized such that its rows sum up to one.

Simultaneous autoregressive model (SAR). The SAR model is given as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},\ \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\epsilon}. \tag{3}$$

Here λ is an unknown parameter, $-1 < \lambda < 1$, measuring spatial correlation; the parameters λ and $\boldsymbol{\beta}$ are to be estimated. The error vector $\boldsymbol{\epsilon}$ has uncorrelated components with constant unknown variances σ^2, like $\mathbf{u}$ it has expectation zero. The two equations may be combined to get

$$\mathbf{y} = \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} - \lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Obviously $\mathbf{y}$ is modeled as being influenced also by the spatial lag-variables $\mathbf{W}\mathbf{y}$ and the spatial lag-regression $\mathbf{W}\mathbf{X}\boldsymbol{\beta}$. The coefficient λ is measuring the strength of this influence. The covariance matrix of $\mathbf{u}$ may be shown to be $\mathrm{cov}[\mathbf{u}] = \sigma^2((\mathbf{I_n} - \lambda\mathbf{W})^T(\mathbf{I_n} - \lambda\mathbf{W}))^{-1}$. An estimation procedure for the SAR model is implemented in the R-package spdep, Bivand (2006). It is based on the Gaussian assumption for $\mathbf{y}$ and iteratively calculates maximum (profile) likelihood estimates for σ^2 and λ and generalized least squares estimates for $\boldsymbol{\beta}$ based on the covariance matrix $\mathrm{cov}[\mathbf{u}]$ and the estimates for σ^2 and λ calculated a step before.

Spatial lag model. The so-called spatial lag model may be written

$$\mathbf{y} = \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{4}$$

It is simpler in structure than the SAR model because the lag-regression term $-\lambda\mathbf{W}\mathbf{X}\boldsymbol{\beta}$ is missing. For its estimation, again, an iterative profile likelihood procedure similar to the SAR procedure may be used.

Spatial Durbin model. The spatial Durbin model is a generalization of the SAR model and given as

$$\mathbf{y} = \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{5}$$

with $\mathbf{W}\mathbf{X}\boldsymbol{\gamma}$ having its own regression parameter vector $\boldsymbol{\gamma}$. By means of the restriction $\boldsymbol{\gamma} = -\lambda\boldsymbol{\beta}$ the Durbin model

becomes equivalent to a SAR model. The so-called common factor test (Florax and de Graaf 2004), a likelihood ratio test, can be used to decide between the two hypotheses, - SAR-model and spatial Durbin model. As an alternative to the above models one may also use a SAR model with a lag-error component

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \lambda \mathbf{W}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}. \tag{6}$$

Deciding between models. For the investigation whether a SAR model, a spatial lag model or ordinary least squares give the best fit to the data one may adopt Lagrange multiplier tests as described in Florax and de Graaf (2004). Interestingly, these tests are based on ordinary least squares residuals and for this reason are easily calculable. Breitenecker (2009) gives a nice overview on all the possibilities related to testing models.

Geographically weighted regression. Fotheringham et al. (2002) propose, as an alternative to the above mentioned regression models, geographically weighted regression. The proposed methodology is particularly useful when the assumption of stationarity for the response and explanatory variables is not met and the regression relationship changes spatially. Denoting by (u_i, v_i) the centroids of the spatial areas A_i, $i = 1, 2, \ldots, n$, where the responses y_i and explanatory vectors $\mathbf{x_i}$ are observed, the model for geographically weighted regression may be written

$$y_i = \mathbf{x_i}^T \boldsymbol{\beta}(u_i, v_i) + \epsilon_i, i = 1, 2, \ldots, n. \tag{7}$$

The regression vector $\boldsymbol{\beta}(u_i, v_i)$ is thus dependent on the spatial location (u_i, v_i) and is estimated by means of a weighted least squares estimator that is locally dependent on a diagonal weight matrix $\mathbf{C_i}$:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = (\mathbf{X}^T \mathbf{C_i} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{C_i} \mathbf{y}$$

The diagonal elements $c_{jj}^{(i)}$ of $\mathbf{C_i}$ are defined by means of a kernel function, e.g. $c_{jj}^{(i)} = \exp(-d_{ij}/h)$. Here d_{ij} is a value representing the distance beetween A_i and A_j; d_{ij} may either be Euclidean distance or any other metric measuring distance between areas. Further, h is the bandwidth measuring how related areas are and can be determined by means of crossvalidating the residuals from the regression or based on the ▶Akaike's information criterion (Brunsdon et al. 1998). Selecting the bandwidth h too large results in oversmoothing of the data. On the other hand a bandwidth too small allows for too less data during estimation.

All areal analysis methods discussed so far are implemented in the R-packages spdep and spgwr, (Bivand 2006, 2009). Methods for counting data, as they frequently appear in epidemiology, and Bayesian methods are not dealt with here; for those methods the interested reader is referred to Lawson (2009).

Spatial Interaction Data

This is a further category of spatial data which is related to modeling the "flow" of people and/or objects between a set of origins and a set of destinations. In contrast with areal (and geostatistical) data, which are located at points or in areas, spatial interaction data are related to pairs of points, or pairs of areas. Typical examples arise in health services (e.g., flow to hospitals), transport of freight goods, population migration and journeys-to-work. Good introductory material on spatial interaction models can be found in Haynes and Fotheringham (1984).

The primary objective is to model *aggregate* spatial interaction, i.e. the volume of flows, not the flows at an individual level. Having m origins and n destinations with associated flow data considered as random variables Y_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, n)$, the general spatial interaction model is of the form

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij};\ i = 1, \ldots, m;\ j = 1, \ldots, n \tag{8}$$

where $E(Y_{ij}) = \mu_{ij}$ and ε_{ij} are error terms with $E(\varepsilon_{ij}) = 0$. The goal is then to find suitable models for μ_{ij} involving flow propensity parameters of the origins i, attractiveness parameters of the destinations j, and the effects of the "distances" d_{ij} between them. Here, the quantities d_{ij} may be real (Euclidean) distances, travel times, costs of travel or any other measure of the separation between origins and destinations. One of the most widely used classes of models for μ_{ij} is the so-called *gravity model*

$$\mu_{ij} = \alpha_i \beta_j \exp(\gamma\, d_{ij}) \tag{9}$$

involving origin parameters α_i, destination parameters β_j and a scaling parameter γ. Under the assumption that the Y_{ij} are independent Poisson random variables with mean μ_{ij}, this model can be treated simply as a particular case of a generalised linear model with a logarithmic link. Model fitting can then proceed by deriving maximum likelihood estimates of the parameters using iteratively weighted least squares (IRLS) techniques. The above gravity models can be further enhanced when replacing the parameters β_j by some function of observed covariates $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jk})^T$ characterising the attractiveness of each of the destinations $j = 1, \ldots, n$. Again, this is usually done in a log-linear way, and the model becomes

$$\mu_{ij} = \alpha_i \exp(g(\boldsymbol{x}_j, \boldsymbol{\theta}) + \gamma\, d_{ij}) \tag{10}$$

where g is some function (usually linear) of the vector of destination covariates and a vector of associated parameters $\boldsymbol{\theta}$. Contrary to (9), which reproduces both the total flows from any origin and the total observed flows to each destination, the new model (10) is only *origin-constrained*. The obvious counterpart to (10) is one which is *destination-constrained*:

$$\mu_{ij} = \beta_j \exp(h(\boldsymbol{z}_i, \boldsymbol{\omega}) + \gamma\, d_{ij})$$

where h is some function of origin characteristics $\boldsymbol{z}_i$ and a vector of associated parameters $\boldsymbol{\omega}$. Finally, when modeling both α_i and β_j as functions of observed characteristics at origins and destinations, we arrive at the *unconstrained model*

$$\log \mu_{ij} = h(\boldsymbol{z}_i, \boldsymbol{\omega}) + g(\boldsymbol{x}_j, \boldsymbol{\theta}) + \gamma\, d_{ij} \qquad (11)$$

In population migration one often uses a particular form of (11), where $\boldsymbol{z}_i$ and $\boldsymbol{x}_j$ are taken to be univariate variables meaning the logarithms of the population P_i and P_j at origin i and destination j, respectively. Adding an overall scaling parameter τ to reflect the general tendency for migration, the following simple model results:

$$Y_{ij} = \tau P_i^{\omega} P_j^{\theta} \exp(\gamma\, d_{ij}) + \varepsilon_{ij} \qquad (12)$$

Likewise, in all the above models one can introduce more complex distance functions than $\exp(\gamma\, d_{ij})$. Also, as mentioned before, d_{ij} could be replaced by a general separation term s_{ij} embracing travel time, actual distance and costs of overcoming distance.

The interaction models considered so far are only models for μ_{ij}, the mean flow from i to j. Thus, they are only first order models, no second order effects are included and the maximum likelihood methods for estimating the parameters of the gravity models rest on the explicit assumption that fluctuations about the mean are independent. Up to now, there has been only little work done on taking account of spatially correlated errors in interaction modeling. To address such problems, pseudo-likelihood-methods are in order. Good references for further reading on spatial interaction models are Upton and Fingleton (1989), Bailey and Gatrell (1995) and Anselin and Rey (2010).

Spatial interaction models have found broad attention among (economic) geographers and within the GIS community, but have received only little attention in the spatial statistics community. The book by Anselin and Rey (2010) forms a bridge between the two different worlds. It contains a reprint of the original paper by Getis (1991), who first suggested that the family of spatial interaction models is a special case of a general model of spatial autocorrelation. Fischer et al. (2010) present a generalization of the Getis-Ord statistic which enables to detect local non-stationarity and extend the log-additive model of spatial interaction to a general class of spatial econometric origin-destination flow models, with an error structure that reflects origin and/or destination autoregressive spatial dependence. They finally arrive at the general spatial econometric model (3), where the design matrix $\boldsymbol{X}$ includes the observed explanatory variables as well as the origin, destination and separation variables, and $\boldsymbol{W}$ is a row-standardized spatial weights matrix.

About the Author

For biography of the author Jürgen Pilz *see* the entry ►Statistical Design of Experiments (DOE).

Cross References

►Data Depth
►Gaussian Processes
►Geostatistics and Kriging Predictors
►Model-Based Geostatistics
►Spatial Point Pattern
►Spatial Statistics
►Statistical Ecology

References and Further Reading

Anselin L (1988) Spatial econometrics: methods and models. Kluwer Academic, Dordrecht

Anselin L (1955) Local indicators of spatial association – LISA. Geogr Anal 27:93–115

Anselin L, Rey SJ (eds) (2010) Perspectives on spatial data analysis. Springer, Berlin

Bailey T, Gatrell A (1995) Interactive spatial data analysis. Longman Scientific and Technical, New York

Breitenecker R (2009) Raeumliche lineare modelle und autokorrelationsstrukturen in der gruendungsstatistik. Ibidem, Stuttgart

Bivand R (2006) SPDEP: spatial dependence: weighting schemes, statistics and models. R-package Version 0.3-12

Bivand R (2009) SPGWR: geographically weighted regression. R-package Version 0.6-2

Brunsdon C, Fotheringham S, Charlton M (1998) Geographically weighted regression – modelling spatial non-stationary. The Statistician 47:431–443

Fischer MM, Reismann M, Scherngell Th (2010) Spatial interaction and spatial autocorrelation. In: Rey SJ, Anselin A (eds) perspective on spatial data analysis. Springer, Berlin, pp 61–79

Florax R, de Graaf T (2004) The performance of diagnostic tests for spatial dependence in linear regression models: a meta-analysis of simulation studies. In: Anselin L et al (eds) Advances in spatial econometrics: methodology, *tools and applications*. Springer, Berlin, pp 29–77

Fotheringham S, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester

Geary R (1954) The contiguity ratio and statistical mapping. Inc Stat 5:115–145
Getis A (1991) Spatial interaction and spatial autocorrelation: a cross-product approach. Environ plann A23:1269–1277
Haynes KF, Fotheringham AS (1984) Gravity and spatial models. Sage, London
Lawson A (2009) Bayesian disease mapping: hierarchical modeling in spatial epidemiology. CRC, Chapman and Hall, New York
Moran P (1950) Notes on continuous stochastic phenomena. Biometrica 37:17–23
Tobler W (1970) A computer model simulating urban growth in the Detroit region. Econ Geogr 46:234–240
Upton GJG, Fingleton B (1989) Spatial data analysis by example, vol. 2. Wiley, Chichester

Analysis of Covariance

James J. Cochran
Associate Professor
Louisiana Tech University, Ruston, LA, USA

Introduction

The Analysis of Covariance (generally known as ANCOVA) is a statistical methodology for incorporating quantitatively measured independent observed (not controlled) variables in a designed experiment. Such a quantitatively measured independent observed variable is generally referred to as a covariate (hence the name of the methodology – analysis of covariance). Covariates are also referred to as concomitant variables or control variables.

If we denote the general linear model (GLM) associated with a completely randomized design as

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, m$$

where
Y_{ij} = the ith observed value of the response variable at the jth treatment level
μ = a constant common to all observations
τ_j = the effect of the jth treatment level
ε_{ij} = the random variation attributable to all uncontrolled influences on the ith observed value of the response variable at the jth treatment level

For this model the within group variance is considered to be the experimental error, and this implies that the treatments have similar effects on all experimental units. However, in some experiments the effect of the treatments on the experimental units varies systematically with some characteristic that varies across the experimental units. For example, one may test for a difference in the efficacy of a new medical treatment and an existing treatment protocol by randomly assigning the treatments to patients (experimental units) and testing for a difference in the outcomes. However, if the ▶randomization results in the placement of a disproportionate number of young patients in the group that receives the new treatment and/or placement of a disproportionate number of elderly patients in the group that receives the existing treatment, the results will be biased if the treatment is more (or less) effective on young patients than it is on elderly patients. Under such conditions one could collect additional information on the patients' ages and include this variable in the model. The resulting general linear model

$$Y_{ij} = \mu + \tau_j + \beta X_{ij} + \varepsilon_{ij},\ i = 1, \ldots, n_j,\ j = 1, \ldots, m.$$

where
X_{ij} = the ith observed value of the covariate at the jth treatment level,
β = the estimated change in the response that corresponds to a one unit increase in the value of the covariate at a fixed level of the treatment

is said to be a completely randomized design ANCOVA model and describes an experimental design GLM one factor experiment with a single covariate.

Note that the addition of covariate(s) can accompany many treatment and design structures. This article focuses on the simple one way treatment structure in a completely randomized design for the sake of simplicity and brevity.

Purpose of ANCOVA

There are three primary purposes for including a covariate in the ▶analysis of variance of an experiment:

1. To increase the precision of estimates of treatment means and inferences on differences in the response between treatment levels by accounting for concomitant variation on quantitative but uncontrollable variables. In this respect covariates are the quantitative analogies to blocks (which are qualitative/categorical) in that they are (1) not controlled and (2) used to remove a systematic source of variation from the experimental error. Note that while the inclusion of a covariate will result in a decrease in the experimental error, it will also reduce the degrees of freedom associated with the experimental error, and so inclusion of a covariate in an experimental model will not always result in greater precision and power.

2. To allow for the assessment of the nature of the relationship between the covariate(s) and the response variable after taking into consideration the treatment effects. In this respect covariates are analogous to independent variables in linear regression, and their associated slopes can provide important insight into the nature of the relationship between the response and the covariate.
3. To statistically adjust comparisons of the response between groups for imbalances in quantitative but uncontrollable variables. In this respect covariates are analogous to stratification and are of particular importance in situations where stratification on the covariate is impractical or infeasible.

Applications of ANCOVA

Typical applications of analysis of covariance include:

- Clinical trials in which quantitative but uncontrollable variables such as the weight, height, and age of the patients may influence the effectiveness of a treatment protocol.
- Marketing research in which quantitative but uncontrollable variables such as the pretest rating of a product given by a respondent may influence the respondent's posttest rating (i.e., after exposure to the test condition) of the product.
- Education experiments in which quantitative but uncontrollable variables such as the age, intelligence (if this can be measured), and prior scholastic performance of the students may influence the effectiveness of a pedagogical approach.
- Agricultural experiments in which quantitative but uncontrollable variables such as rainfall and historical yield of fruit bearing trees may influence the yield during an experiment.

Comparing Treatments in ANCOVA

Least squares means (or *LS* means) are generally used to compare treatment effects in experiments that include one or more covariates. *LS* means (which are sometimes referred to as marginal means, estimated marginal means, or adjusted treatment means) are the group means when the covariate is set equal to its grand mean $\overline{X}_m$ (mean of the covariate over all observations across all treatments). These are easily calculated by substituting the grand mean of the covariate into the estimated general linear model, i.e.,

$$\hat{Y}_j = \mu + \tau_j + \beta\overline{X}_m,\ j = 1, \ldots, m$$

Standard errors for *LS* means are typically calculated and used (in conjunction with the ▶asymptotic normality of *LS* means) to conduct inference on individual *LS* means and contrasts based on the *LS* means.

Assumptions of ANCOVA

In addition to the standard ANOVA assumptions:

- Independence of error terms
- Homogeneity of variance of the error terms across treatments
- Normality of the error terms across treatments

One must also consider the *regression assumptions* when performing statistical inference with ANCOVA. The regression assumptions include:

- A linear relationship exists between the covariate and the response variable.

  If no relationship exists between the covariate and response, there is no reason to include the covariate in the experiment or resulting model. If the relationship between the covariate and the response variable is nonlinear, the inclusion of a covariate in the model will not remove all variation in the observed values of the response that can potentially be accounted for by the covariate. The nature of the relationship between the covariate and the response can be assessed with scatter plots of these two variables by treatment. If a nonlinear relationship exists between the covariate and the response, one can utilize a polynomial ANCOVA model.
- Homogeneity of the regression slopes associated with the covariate (i.e., parallel regression lines across treatments).

  The calculations of the *LS* means are predicated on the lack of existence of a response by covariate interaction. If this assumption is violated, the adjustment to the response variable for a common value of the covariate is misleading. This assumption can be assessed through either scatter plots of the covariate and the response by treatment or through the inclusion of a treatment-covariate interaction in the model.

If the sample results suggest that any of these assumptions are not satisfied, inference based on the model may not be valid.

Alternatives to ANCOVA

Bonate (2000) provides a good discussion of alternatives to ANCOVA in pretest-posttest designs; he considers the relative merits of difference scores, relative change functions, various blocking methods, and repeated-measures analysis. Several authors have suggested more general nonparametric alternatives to ANCOVA based on an analysis

of covariance of the ranks of the response and covariance. Some notable examples of these approaches have been suggested by Quade (1967, 1982), Puri and Sen (1969), McSweeney and Porter (1971), Burnett and Barr (1977), Shirley (1981), Conover and Iman (1982), Chang (1993), Lesaffre and Senn (2003), and Tsangari and Akritas (2004).

About the Author

For biography *see* the entry ►Role of Statistics in Advancing Quantitative Education.

Cross References

►Analysis of Variance
►General Linear Models
►Multivariate Analysis of Variance (MANOVA)
►Nonparametric Models for ANOVA and ANCOVA Designs
►Rank Transformations
►Statistical Fallacies: Misconceptions, and Myths
►Statistics: An Overview

References and Further Reading

Bonate PL (2000) Analysis of pretest-posttest designs. Chapman and Hall/CRC, Boca Raton

Burnett TD, Barr DR (1977) A nonparametric analogy of analysis of covariance. Educ Psychol Meas 37(2):341–348

Chang GH (1993) Nonparametric analysis of covariance in block designs. Dissertation (Texas Tech University)

Conover WJ, Iman RL (1982) Analysis of covariance using the rank transformation. Biometrics 38:715–724

Doncaster CP, Davey AJH (2007) Analysis of variance and covariance: how to choose and construct models for the life sciences. Cambridge University Press, Cambridge

Huitema BE (1980) The analysis of covariance and alternatives. Wiley, New York

Lesaffre E, Senn S (2003) A note on non-Parametric ANCOVA for covariate adjustment in randomized clinical trials. Stat Med 22(23):3583–3596

McSweeney M, Porter AC (1971) Small sample properties of nonparametric index of response and rank analysis of covariance. Presented at the Annual Meeting of the American Educational Research Association, New York

Milliken GA, Johnson DE (2002) Analysis of messy data vol.3: analysis of covariance. Chapman and Hall, New York

Puri ML, Sen PK (1969) Analysis of covariance based on general rank scores. Ann Math Stat 40:610–618

Quade D (1967) Rank analysis of covariance. J Am Stat Assoc 62:1187–1200

Quade D (1982) Nonparametric analysis of covariance by matching. Biometrics 38:597–611

Rutherford A (2001) Introducing ANOVA and ANCOVA: a GLM approach. Sage, Los Angeles

Shirley EA (1981) a Distribution-Free Method for Analysis of Covariance Based on Ranked Data. J Appl Stats 30:158–162

Tsangari H, Akritas MG (2004) Nonparametric ANCOVA with two and three covariates. J Multivariate Anal 88(2):298–319

Analysis of Multivariate Agricultural Data

Asghar Ali
Professor and Chairman
Bahauddin Zakariya University, Multan, Pakistan

Agricultural research is most oftenly based on observational studies and experimentation resulting in multi-response variables. The selection of appropriate variety to grow; amount and types of fertilizers, insecticides and pesticides to apply; the irrigation system to use; the plant sowing technology to apply and to assess the soil fertility through chemical analysis of macro and micro nutrients available in the soil are the major areas of interest for the researcher to work on for the improvement of the agricultural productivity in terms of quality and quantity. The role of Statistics in planning agricultural research, designing experiments, data collection, analysis, modeling and interpretation of agricultural results is very well established. The basic principles and theoretical development of experimental designs pioneered by R. A. Fisher are the result of collaborative work of agricultural scientists and statisticians. In the process of experimentation and observational studies, the researcher is keen to have as many data information as possible so that nothing is left unattended related to the phenomenon under study as there will be no chance to repeat the experiment till the next season of the crop and it will not be less than a miracle if data from one year of the crop is consistent with the results of second year, no matter how much care is taken to keep the experimental conditions identical.

Agricultural data obtained through experimentation is initially analyzed using ►analysis of variance technique and then depending on the nature of treatments/factors applied, either the approach of multiple comparisons or ►response surface methodology is used to explore further the hidden features of the data. For example, the experimenter might be interested to compare different varieties of a particular crop such that there are two local varieties (V_1, V_2) in practice; three varieties are imported (V_3, V_4, V_5) and two new varieties (V_6, V_7) are developed by a local agricultural institute. If results obtained from analysis of variance conclude that performance of the varieties is significantly different from each other then obvious questions arise are to test the difference between the following variety comparisons: $[V_1$ and $V_2]$; $[V_6$ and $V_7]$; $[(V_1, V_2)$ and $(V_6, V_7)]$; $[(V_1, V_2, V_6, V_7)$ and $(V_3, V_4, V_5)]$; and if V_4 is a hybrid variety, then one has two more comparisons to test i.e., $[V_4$ with $(V_3, V_5)]$

and $[V_3$ with $V_5]$. These contrasts are orthogonal to each other but it will not always be the case, other techniques of multiple comparisons will have to be used then, which are available in almost all the books on experimental designs. On the other hand, if multifactor experiments are conducted to determine the appropriate levels of the applied factors on which optimum response is achieved. For this purpose data sets are modeled in adequate functional forms and the researcher is intended to fit simple functional form. Ordinary polynomials are the most popular functional forms which are used to model experimental data from many fields of scientific research. If first order polynomial is fitted, the researcher very simply states that the concerned factor has linear effect and the interpretation is made accordingly that with increase of levels of factor will result in increase (or decrease) in the response. The second order polynomials are used with the expectation that it will be possible to identify the levels of the applied factors to get the optimum response. A number of response functions that have been widely used by the agricultural and biological researchers have been discussed by Mead and Pike (1975). It should not be taken as granted that one response function considered applicable to one sort of situation will also be applicable to other similar situations; it is advisable that graphical approach be used to guess the appropriate functional form of the response under consideration. An extremely useful concept that is revealed by Nelder (1966) is known as Inverse Polynomial Response Functions (IPRF). It emphasizes that in agricultural research the effect of increasing a factor indefinitely is either to produce a saturation effect, in which case the response does not exceed a finite amount, or to produce a toxic effect, in which case the response eventually fall to zero and the response curve has no build-in-symmetry.

Nelder (1966) and Pike (1977) advocated these surfaces as giving responses that are nonnegative and bounded if regression coefficients are constrained to be positive and it is further assumed that $Var(Y) \propto [E(Y)]^2$. Extension in the ideas has been developed by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989) for the response variables that may not be normal and that the expected response may be a function of the linear predictors rather than just the linear predictors itself. Ali (1983) and Ali et al. (1986) have objected on placing constrains on the parameters as it will violate all the properties of good estimators and will no longer follow the distributional structure required for valid inferences. Their experience of examining many sets of data leads them not to expect all regression coefficients to be positive. Taking into account the error structure and functional form used for IPRF, Ali (1983) proposed the form of a response function called as Log Linear Response Functions (LLRF) based on the logarithmic transformation of the response variable and assuming that $\log Y \sim N(E(\log Y), \sigma^2)$. The estimation of regression coefficients achieved by carrying out a multiple regression of $\log Y$ on the terms required fitting the data adequately; the resulting estimators are therefore Minimum Variance Linear Unbiased Estimators. It is simple to estimate the variance-covariance matrix of these estimators and to test hypotheses concerning parameters by the usual linear regression methods. On the theoretical grounds the LLRF model therefore has much to commend it. The assumption that $\log Y_i$ follows the normal distribution may not always be true; in such cases it is recommended that Box-Cox family of transformation may be used under the same structure of the response function as has been used for LLRF and IPRF.

In order to produce an adequate prediction the researcher is usually uncertain as to which of the large number of terms should be included in the final model. The main point to bear in mind is that it should have as many terms as necessary so that maximum variation of the data is explained and as few terms as possible so that it can easily be interpreted. Ali (1983) argued that for summarizing the data from agricultural experiments the terms in the final model are required to be selected in a conforming order by preferring main effect terms over the interactions and lower order terms over the higher. It is further to remember that the inverse terms describe the rising ridge of the surfaces, the linear terms describe the optimum region and the higher degree terms contribute in explaining the falling portion of the surfaces. It is therefore recommended that for building the appropriate model one should concentrate on selection the inverse and linear terms along with their associated interaction terms. One who is not convinced with such types of model building method has the option to use the approach established by Nelder (1977).

The methods used for selection of final model are mainly based on the Minimum Mean Square Error criterion. It is possible to find more than one models which fulfills this criterion. In such cases one should select the one which has reasonable shape of the response surface, capable to determine the values of quantitative factors at which the response is an optimum, statistically significant regression coefficients and simple functional form.

There is no ambiguity to recognize the agricultural research as multifactor and multi-response and that these responses are measured at different stages of the maturity of the crop and that these are interrelated with one another. The univariate analyses of these variables therefore have partial impact on the true findings of research. Multivariate analyses are therefore natural and essential to consider the

data by giving due weight to the interrelationships among the variables under study. One possible approach which is widely used by the researchers is to study the correlation matrix of the variables. This approach only facilitates to assess the relationship among the pairs of variables and it can be extended to triplets of variables by considering the partial correlations and the multiple correlations among those. As a result there would be $\frac{1}{6}p(p^2 + 1)$ pairs and triplets to consider and it will certainly be confusing if the number p of variables under consideration is large.

To overcome this difficulty, Principal Component Analysis (PCA) can be used. It is a multivariate technique that has its aim the explanation of relationships among several difficult-to-interpret, correlated variables in terms of a few conceptually meaningful components which are uncorrelated with each other and are capable of accounting for nearly all the variation present in the observed data. PCA therefore finds a linear transformation of original variables into a new set of variables called as principal components which are uncorrelated with each other; are capable of accounting for the variation of the obtained data and are derived in such a way that the first few of them can often provide enough information about the data and so the dimensionality of the problem can considerably be reduced. The variables with higher component loadings in a particular principal component are considered to be the important ones and it is assumed that the principal component is the representative of these variables; hence it is interpreted only in terms of these variables. This approach of interpretation of principal components is acceptable if principal components are extracted using the correlation matrix R. The variance–covariance matrix Σ is as well used to derive principal components; since the principal component technique is scale dependent, the principal component loadings with this approach will therefore be much influenced by the unit of the measurements of the variables under consideration, hence, the interpretation of principal components just based on the magnitude of the loading may become questionable and misleading. Ali et al. (1985) suggested using the correlation between the original variables and the principal component for selection of representative variables in a particular principal component instead of using principal component loadings.

PCA is extremely useful technique when interest lies in investigating the interrelationship within a set of variables; when the relationship of two sets of variables, within and among the sets is of interest, the PCA is not a valid technique. The agricultural researchers always encounter with such types of problems where the assessment of relationship among and within the twosets is essential e.g. the interdependence of nutritional status and vegetative related characteristics with the crop yield related characteristics is pivotal. For such cases, ►Canonical Correlation Analysis (CCA) technique developed by Hotelling (1936) is of great benefit. It has certain maximal properties similar to those of PCA and in a way is an extension of the multiple regression analysis. The object of this approach is to find the linear functions of the variables for each of the sets such that the correlation between these linear functions is as high as possible. After locating such a pair of linear functions which are maximally correlated with each other, we look for other pairs of linear functions which are maximally correlated subject to the restriction that the new pair of linear functions must be uncorrelated with all other previously located functions. For the purpose of interpretation of the results, it is proposed to use correlation between the canonical variates and the original variables instead of canonical weights as has been already proposed for interpretation of PCA results. Details of PCA and CCA may be found in Mardia et al. (1979) and Jolliffe (2002).

About the Author

Asghar Ali holding M.Sc and D.Phil in Statistics from Sussex University and Post Doctorate from University of Kent at Canterbury, UK is Professor of Statistics at Bahauddin Zakariya University (BZU) Multan, Pakistan. Since 1975, he is serving BZU in various capacities: Presently, he is Chairman, Department of Statistics and also for a period of three years, he held the position of chairmanship, Department of Computer Science. Especially commendable have been his period as Principal, College of Agriculture, BZU. He has been publishing regularly related to data analysis in the field of agricultural sciences which is now being cited in related text books.

Cross References

►Agriculture, Statistics in
►Canonical Correlation Analysis
►Farmer Participatory Research Designs
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Principal Component Analysis

References and Further Reading

Ali A (1983) Interpretation of multivariate data: Comparison of several methods of interpreting multivariate data from a series of nutritional experiments, University of Sussex, Unpublished PhD thesis

Ali A, Clarke GM, Trustrum K (1985) Principal component analysis applied to some data from fruit nutrition experiments. The Statistician 34:365–370

Ali A, Clarke GM, Trustrum K (1986) Log-linear response functions and their use to model data from plant nutrition experiments. J Sci Food & Agric 37:1165–1177
Hotelling H (1936) Relation between two sets of variates. Biometrika 28:321–377
Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, USA
Mardia KV, Kent JT, Bibi JM (1979) Multivariate analysis. Academic, London
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London
Mead R, Pike DJ (1975) A review of response surface methodology from a biometric viewpoint. Biometrics 31(4):803–851
Nelder JA (1966) Inverse polynomials, a useful group of multifactor response functions. Biometrics 22:128–141
Nelder JA (1977) A reformation of linear models (with discussion), J R Stat Soc A140:48–76
Nelder JA, Wedderburn WM (1972) Generalized linear models. J R Stat Soc (General) A135(3):370–384
Pike DJ (1977) Inverse polynomials: A study of parameter estimation procedures and comparison of the performance of several experimental design criteria. University of Reading, Unpublished PhD thesis

Analysis of Variance

Gudmund R. Iversen
Professor Emeritus
Swarthmore College, Swarthmore, PA, USA

Analysis of variance is the name given to a collection of statistical methods originally used to analyze data obtained from experiments. The experiments make us of a quantitative dependent variable, also known as a metric variable or an interval or ratio variable, and one or more qualitative independent variables, also known as categorical or nominal variables. These analysis methods grew out of agricultural experiments in the beginning of the twentieth century, and the great English statistician Sir Ronald Fisher developed many of these methods. As an example, the dependent variable could be the yield in kilos of wheat from different plots of land and the independent variable could by types of fertilizers used on the plots of land.

Experimental Design

The way an experiment is run affects the particular analysis of variance method used for the analysis of the data. Experiments are designed according to different plans, and the choice of the design of the experiment affects which analysis of variance method being used. Without going into details about designs of experiments, an experiment could follow a factorial design, a randomized block design, a Latin square design, etc. There exist too many designs of experiments and accompanying analysis of variance methods for the analysis of the resulting data to cover all of them in this short presentation. But it is possible to present the underlying features of all analysis of variance methods.

Analysis of Variance and Multiple Regression

But first it is worth noting that analysis of variance is closely related to regression analysis. Indeed, it is possible to see both analyses as special cases of the so-called general linear model. In particular, using ▶dummy variables for the independent variables in analysis of variance, the analysis quickly turns into a regression analysis. The main difference is that when data are collected through a properly designed experiment, it is possible to conclude that there is a causal effect of the independent variable(s) on the dependent variable. When data are collected through observational, studies there may be a causal effect of the independent variable(s) or not.

Statistical Software

Much of the early work on analysis of variance consisted of finding efficient ways of making the necessary computations with the use of simple calculators. With the introduction of modern statistical software for electronic computers, this line of work is now less important. Instead, statisticians have worked on showing the similarities of the computations needed for both analysis of variance and multiple regression, and the old distinction between the two approaches to data analysis is no longer of any importance. However, statistical software packages still make a distinction between the two, and the output from the two methods often look very different.

One-Way Analysis of Variance

This name is given to the design where there is one independent nominal variable with several categories and a quantitative dependent variable with a unit of measurement and often a meaningful zero. An example of an experiment could be where students are randomly assigned to two different groups and the students in one group were taught using a new method of teaching while the students in the second group, as a control group, were taught using the old method. The random assignment to the different groups means that the effects of all other variables, for

example gender, is canceled out, and any observed difference between the two groups is causally due to the teaching method being used. In this simple case, the statistical method is the same as the *t*-test for the difference between two groups. From a regression point of view we could use a dummy variable and assign the value 0 to all the students in the control group and the value 1 to all the students in the experiment group. In this case, the intercept of the regression line would equal the mean of the dependent variable for the control group and the slope of the line would equal the difference between the means of the two groups. Thus, the *t*-test for the null hypothesis that the population regression line has zero slop becomes the same as the *t*-test for the difference between the two means.

The fundamental question in an analysis of variance is whether the population means for different groups are equal or not. But the methods for analysis of variance use variances to answer the question about means, thus the name analysis of variance. The analysis is based on identifying two factors that determine the values of the dependent variable. One such factor is the net effect of all factors except the independent variable, known as the residual variable, and the other factor is the independent variable.

The Residual Sum of Squares

If the residual variable had no effect, then all the values of the dependent variable for the control group would be equal to each other, and all the values of the dependent variable for the experimental group would be equal to each other. The best estimates of these two values would be the mean of the dependent variable for the group. To the extent that the values within each group are not equal, is due to the residual variable. Thus, the effect of the residual variable for a single observation can be seen as the difference between the observed value and the group mean. For each observation we now have such a difference. One way to summarize the values of these differences for a group is to square each difference and add all these squares. We then have a sum of squares for each of the two groups, and by adding these two sums we have a measure of the overall effect of the residual variable. If the dependent variable is known as Y and y_{ij} is the ith observation in the jth group and $\bar{y}_j$ is the mean in the jth group, then the residual sum of squares RSS can be written

$$RSS = \sum \sum (y_{ij} - \bar{y}_j)^2.$$

Note that there are many other ways we could combine these differences. For example, we could have taken the absolute value of each difference and added those differences instead of using squares. Thus, the final conclusion from the analysis should include a statement that the conclusion is based on squares and not some other mathematical operation. Even though nobody does include such a warning, it should be made clear that the analysis is based on squares.

The Treatment Sum of Squares

We also need a measure of how different the two groups are from each other. One way to do that is find how different the group means are from the overall mean. If the treatment variable has no effect, then the two group means would be equal and equal to the overall mean. One way to measure how different the group means are from the overall mean is to take each group mean and subtract the overall mean. By squaring each difference and weighing each square by the number of observations in the group n_j, then the treatment sum of squares between the groups GSS can be written

$$GSS = \sum \sum n_j (\bar{y}_j - \bar{y})^2.$$

The *F*-Test

The residual sum of squares is also known as the within group sum of squares and the group sum of squares is sometimes known as the between group squares. The final step consists of making a comparison between the two sums of squares. If the residual sum of squares is large in comparison with the group sum of squares, then it seems that the difference between the group means is not statistically significant. For this comparison we take into how many groups we have, here 2 and in general k groups, and how many observations n there are all together. A mathematical development shows that we should compute the ratio

$$F = \frac{GSS/(k-1)}{RSS/(n-k)}$$

This is known as the F-ratio and is named in honor of Ronald Fisher. It gives rise to the F-distribution, and the distribution has been extensively tabulated. The two numbers $(k-1)$ and $(n-k)$ are the so-called degrees of freedom, and they are used to take into account how many groups there are in the experiment and how many observations there are in the experiment. For example, for a 5% significance level with $k = 2$ groups and $n = 30$ observations, the critical value of F on 1 and 28 degrees of freedom equals 4.20. Thus, for any observed value of F larger than 4.20, we conclude that there is a statistically significant difference between the two groups. In this case, had we done a *t*-test

for the difference between the special case of two group means, the critical value of t becomes $\sqrt{4.20} = 2.05$.

Other Analyses

It is possible to generalize to an experiment with more than just two groups. The null hypothesis of equal group means is tested the same way as with two groups, and the computations follow the same plan as above. With two or more independent variables the analysis becomes more extensive. We can still represent the independent variables by dummy variables and do a regression analysis. But that way it is easy to overlook the possible interaction effect of the two independent variables. This means we could have an effect of the independent variables together over and beyond their separate effects. Finally, in analysis of variance we distinguish between using all values of the independent variables (Model I) and only using a sample of possible values (Model II).

About the Author

Dr. Gudmund Iversen is Professor Emeritus of Statistics at Swarthmore College, Swarthmore PA 19081. He chaired the Department of Mathematics and Statistics at three different intervals and directed the College's Center for Social and Policy Studies for several years. Prior to Swarthmore he taught statistics at the University of Michigan and also directed the Summer Training Program, Inter-university Consortium for Political and Social Research. He was a visiting lecturer in the American Statistical Association visiting lecture program. He has been a visiting professor at Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim, West Germany 1986, at Department of Political Science, University of Oslo, Norway (1978–1979), at The Graduate School of Social Work and Social Research, Bryn Mawr College, spring 1990, fall 2000, and at School of Social Policy and Practice at the University of Pennsylvania 1999. He was a member of the joint committee of Mathematical Association of America and American Statistical Association on the statistics curriculum (1991–1997). He was Associate book review editor, *Journal of American Statistical Association* (1986–1989), and Associate editor, *Journal of Statistics Education* (1993–2000). He has published 22 articles and 10 books on statistics and statistical education.

Cross References

►Agriculture, Statistics in
►Analysis of Covariance
►Analysis of Multivariate Agricultural Data
►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Data Analysis
►Experimental Design: An Introduction
►F Distribution
►Farmer Participatory Research Designs
►General Linear Models
►Graphical Analysis of Variance
►Multiple Comparison
►Multiple Comparisons Testing from a Bayesian Perspective
►Multivariate Analysis of Variance (MANOVA)
►Multivariate Data Analysis: An Overview
►Nonparametric Models for ANOVA and ANCOVA Designs
►Parametric Versus Nonparametric Tests
►Rank Transformations
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences
►Statistical Software: An Overview
►Statistics: An Overview
►Tests for Homogeneity of Variance

References and Further Reading

Hinkelmann K, Kempthorne O (2008) Design and analysis of experiments, I and II, 2nd edn. Wiley, New York
Iversen GR, Norpoth H (1987) Analysis of variance, 2nd edn. Sage Baverly Hills, CA

Analysis of Variance Model, Effects of Departures from Assumptions Underlying

Hardeo Sahai[1], Mohammed I. Ageel[2], Anwer Khurshid[3]
[1]Professor
University of Puerto Rico, San Juan, Puerto Rico
[2]Founder and President of the Saudi Association of Statistical Sciences, Professor
Jazan University, Jazan, Saudi Arabia
[3]Professor
University of Karachi, Karachi, Pakistan

Introduction

Every statistical model has its own underlying "assumptions" that must be verified to validate the results. In some situations, violations of these assumptions will not change substantive research conclusions, while in others,

violation of assumptions can be critical to meaningful research. For a meaningful and conclusive data analysis by ►Analysis of Variance (ANOVA), the following assumptions are needed:

(a) Errors be normally distributed
(b) Errors have same variances (homogeneity of variances)
(c) Errors be independently distributed

However, the question arising is What would be the effects of any departure from the assumptions of the model on the inferences made? The answer is simple: It may either influence the probability of making Type I error (i.e., incorrectly rejecting null hypothesis) or a Type II error (i.e., failing to reject a null hypothesis when it is false). For a thorough discussion of the topic, the reader is referred to Scheffé (1959), Miller (1986), Snedecor and Cochran (1989), Sahai and Ojeda (2005), and Sahai and Ageel (2000). Some of the main findings are discussed in the following section.

Effects of Departures from Assumptions

Departures from Normality

For fixed effects model, due to the central limit theorem (see ►Central Limit Theorems) the lack of normality causes no problems in large samples, as long as the assumptions hold. In general, when true ►randomization occurs the violations of normality is acceptable. Also, heterogeneity of variances can result in nonnormality, so ensuring homogeneity of variances may also result in normality. Only highly skewed distributions would have a marked effect either on the level of significance or the power of the F test. However, it is worth mentioning that kurtosis of the error distribution (either more or less peaked than a normal distribution) is more important than skewness of the distribution in terms of the effects on inferences. Both analytical results (see, e.g., Scheffé 1959:345–351) and the empirical studies by Pearson (1931), Geary (1947), Gayen (1950), Box and Anderson (1955), Boneau (1960, 1962), Srivastava (1959), Bradley (1964), Tiku (1964, 1971), and Donaldson (1968) attest to the conclusion that lack of normality would have little effect of F test either in terms of level of significance or power. Hence, the F test is generally robust against departures from normality (in skewness and/or kurtosis) if sample sizes are large or even if moderately large. For instance, the specified level of significance might be 0.05, whereas the actual level for a nonnormal error distribution might vary from 0.044 to 0.052 depending on the sample size and the magnitude of the kurtosis. Generally, the actual level of significance in the presence of positive kurtosis (platykurtic) is slightly higher than the specified one and the real power of the test for positive kurtosis is slightly higher than the normal one. If the underlying population has negative kurtosis (leptokurtic), the actual power of the test will be slightly lower than the normal one (Glass et al. 1972). Single interval estimates of the factor level means and contrasts and some of the multiple comparison methods are also not much affected by the lack of normality provided the sample sizes are not too small. The robustness of multiple comparison tests in general has not been as thoroughly studied. Among few studies in this area is that of Brown (1974). Some other studies have investigated the robustness of several multiple comparison procedures, including Tukey and Scheffé, for exponential and chi-square distributions and found little effect on both α and power (see, e.g., Petrinovich and Hardyck 1969; Keselman and Rogan 1978). Dunnett (1982) reported that Tukey is conservative both with respect to α and power for long-tailed distributions and to ►outliers. Similarly, Ringland (1983) found that Scheffé was conservative for distributions with influence to outliers.

Lange and Ryan (1989) gave several examples that show that nonnormality of random effects is, indeed, encountered in practice. For random effects model, the lack of normality has more serious implications than fixed effects model. The estimates of the variance components are still unbiased, but the actual confidence coefficients for interval estimates of σ_e^2, σ_α^2, $\sigma_\alpha^2/\sigma_e^2$ may be substantially different from the specified one (Singhal and Sahai 1992). Moreover, when testing the null hypothesis, if the variance of a random effect is some specified value different from zero, the test is not robust to the assumption of normality. For some numerical results of this, the reader is referred to Arvesen and Schmitz (1970) and Arvesen and Layard (1975). However, if one is concerned only with a test of hypothesis $\sigma_\alpha^2 = 0$, then slight departures from normality have only minor consequences for the conclusions reached when the sample size is reasonably large (see, e.g., Tan and Wong 1980; Singhal et al. 1988).

Departures from Equal Variances

Both the analytical derivations by Box (1954) and the empirical studies indicate that if the variances are unequal, the F test for the equality of means under fixed effects model is only slightly affected provided there is no remarkable difference in sample sizes and the parent populations are approximately normally distributed. When the variances are unequal, an approximate test similar to the approximate t test when two group variances are unequal may be used (Welch 1956). Generally, unequal error variances increase the actual level of significance slightly

higher than the specified level and result in a slight elevation of the power function to a degree related to the magnitude of differences among variances. If larger variances are associated with larger sample sizes, the level of significance will be slightly less than the nominal value, and if they are associated with smaller sample sizes, the level of significance will be slightly greater than the nominal value (Horsnell 1953; Kohr and Games 1974). Similarly, the Scheffé's multiple comparison procedure based on the ►*F* distribution is not affected to any appreciable degree by unequal variances if the sample sizes are approximately equal. Thus, the *F* test and the related analyses are robust against unequal variances if the sample sizes are nearly equal.

On the other hand, when different number of cases appear in various samples, violation of the assumption of homogeneous variances can have serious effects in the validity of the final inference (see, e.g., Scheffé 1959; Welch 1956; James 1951; Box 1954; Brown and Forsythe 1974; Bishop and Dudewicz 1978; Tan and Tabatabai 1986). Krutchkoff (1988) made an extensive simulation study to determine the size and power of several analysis of variance procedures, including the *F* test, Kruskal–Wallis test, and a new procedure called the *K* test. It was found that both the *F* test and the Kruskal–Wallis test are highly sensitive whereas the *K* test is relatively insensitive to the heterogeneity of variances. Kruskal–Wallis test, however, is not as sensitive to the unequal error variances as the *F* test and was found to be more robust to nonnormality (when the error variances are equal) than either the *F* test or the *K* test. Thus, whenever possible, the experimenter should try to achieve the same number of cases in each factor level unless the assumption of equal population variances can reasonably be assured in the experimental context. The use of equal sample sizes for all factor levels not only tends to minimize the effects of unequal variances using the *F* test, but also simplifies the computational procedure.

For random effects model, however, the lack of homoscedasticity or unequal error variances can have serious effects on inferences about the variance components, even when all factor levels contain equal sample sizes. Boneau (1960) has shown that when variances are different in the various groups and sample sizes are small and different, ANOVA can produce highly misleading results.

Departures from Independence of Error Terms

Lack of independence can result from biased measurements or possibly from a poor allocation of treatments to experimental units. Nonindependence of the error terms can have important effects on inferences for both fixed and random effects models. If this assumption is not met, the *F* ratio may be strongly affected severely in serious errors in inferences (Scheffé 1959). The direction of the effect depends on the nature of the dependence of the error terms. In most cases encountered in practice, the dependence tends to make the value of the ratio too large and consequently the significance level will be smaller than it should be (although the opposite can also be true). Since the remedy of violation of this assumption is often difficult, every possible effort should be made to obtain independent random samples. The use of randomization in various stages of the study can be most important protection against independence of error terms. In general, great care should be taken to ensure that the data are based on independent observations, both between and within groups, i.e., each observation is in no way related to any of the other observations. Although, dependency among the error terms creates a special problem in any analysis of variance, it is not necessary that the observations themselves must be completely independent for applying the random effects model.

In summary, ANOVA is very robust to violations of the assumptions, as long as only one assumption is violated. If two or more assumptions are severely violated the results are not to be trusted. Further if the data are:

(a) Not normally distributed, but satisfies the homogeneity of variance and independent assumptions, the findings may still be valid.
(b) Normally distributed and are independent samples, but does not satisfy the homogeneity of variance assumption, the findings may still be valid.

The above review and discussion are restricted to the one-way analysis of variance. A similar finding for two-way classification without and with interaction can be found in Sahai and Ageel (2000).

Tests for Departures from Assumptions

As we have seen in the preceding section, the analysis of variance procedure is robust and can tolerate certain departures from the specified assumptions. It is, nevertheless, recommended that whenever a departure is suspected it should be checked out. In this section, we shall briefly state the tests for normality and homoscedasticity.

Tests for Normality

A relatively simple technique to determine the appropriateness of the assumption of normality is to graph the data points on a normal probability paper. If a straight line can be drawn through the plotted points, the assumption of normality is considered to be reasonable. Some formal tests for normality are the chi-square goodness of fit test, and the

tests for skewness and kurtosis that are often used as supplements to the chi-square test (see ▶Chi-Square Tests). For a detailed discussion of these tests refer to Sahai and Ageel (2000).

The tests mentioned above are some of the classical tests of normality. Over the years, a large number of other techniques have been developed for testing for departures from normality. Some powerful omnibus tests proposed for the problem are Shapiro–Wilk's test (Shapiro and Wilk 1965), Shapiro–Francia's test (Shapiro and Francia 1972), and D'Agostino's test (D'Agostino 1971).

For a discussion of tests especially designed for detecting outliers see Barnett and Lewis (1994). Robust estimation procedures have also been employed in detecting extreme observations. The procedures give less weight to data values that are extreme in comparison to the rest of the data. Robust estimation techniques have been reviewed by Hampel et al. (1986).

Tests for Homoscedasticity

If there are just two populations, the equality of two population variances can be tested by using the usual F test. However, more than two population, rather than making all pairwise F tests, we want a single test that can be used to verify the assumption of equality of population variances. There are several tests available for this purpose. The three most commonly used tests are the Bartlett's, Hartley's, and Cochran's tests. The ▶Bartlett's test (Bartlett 1937a, b) compares the weighted arithmetic and geometric means of the sample variances. The Hartley's test (Hartley 1950) compares the ratio of the largest to the smallest variance. The Cochran's test (Cochran 1941) compares the largest sample variance with the average of all the sample variances. For a full description of these procedures and illustration of their applications with examples see Sahai and Ageel (2000). These tests, however, have lower power than is desired for most applications and are adversely affected by nonnormality. Detailed practical comments on Bartlett's, Hartley's, and Cochran's tests are also given by Sahai and Ageel. In recent years, there have appeared a number of tests in the literature that are less sensitive to normality in the data and are found to have a good power for a variety of population distributions see Levene (1960). Following Levene (1960), a number of other robust procedures have been proposed, which are essentially based on techniques of applying ANOVA to transformed scores. For example, Brown and Forsythe (1974a) proposed applying an ANOVA to the absolute deviations from the mean. A somewhat different approach known as ▶jackknife was proposed by Miller (1968) where the original scores in each group are replaced by the contribution of that observation to the group variance. O'Brien (1979, 1981) proposed a procedure, which is a blend of Levene's squared deviation scores and the jackknife. In recent years, there have been a number of studies investigating the robustness of these procedures. For a further discussion and details, the reader is referred to Conover et al. (1981), Olejnik and Algina (1987), and Ramsey (1994).

Corrections for Departures from Assumptions of the Model

Departure from independence could arise in an experiment in which experimental units or plots are laid out in a field so that adjacent plots give similar yields. Lack of independence can also result from correlation in time rather than in space. If the data set in a given problem violates the assumptions of the analysis of variance model, a choice of possible corrective measures is available. One approach is to modify the model. However, this approach has the disadvantage that more often than not the modified model involves fairly complex analysis. Another approach may be to consider using some nonparametric tests. A third approach to be discussed in this section is to use transformations on the data. Sometimes it is possible to make an algebraic transformation of the data to make them appear more nearly normally distributed, or to make the variances of the error terms constant. Conclusions derived from the statistical analyses performed on the transformed data are also applicable to the original data. In this section, we briefly discuss some commonly used transformations to correct for the lack of normality and homoscedasticity. An extremely thorough and detailed monograph on transformation methodology has been prepared by Thöni (1967). An excellent and thorough introduction and a bibliography of the topic can be found in a review paper by Hoyle (1973). For a more recent bibliography of articles on transformations see Draper and Smith (1981:683–684).

Transformations to Correct Lack of Normality

Some transformations to correct for the departures from normality are logarithmic transformation, square-root transformation, and arcsine transformation.

Transformations to Correct Lack of Homoscedasticity

There are several types of data in which the variances of the error terms are not constant. If there is evidence of some systematic relationship between treatment mean and variance, homogeneity of the error variance may be achieved through an appropriate transformation of the data. Bartlett (1936) has given a formula for deriving such transformations provided the relationship between μ_i and σ_e^2 is known. In many cases where the nature of the relationship is not clear, the experimenter can, through trial

and error, find a transformation that will stabilize the variance. We give some commonly employed transformations to stabilize the variance. These are logarithmic transformation, square-root transformation, reciprocal transformation, arcsine transformation, and power transformation. For a detailed discussion of these transformations and their applicability refer to Sahai and Ageel (2000).

These are some of the more commonly used transformations. Still other transformations can be found applicable for various other relationships between the means and the variances. Further, the transformations to stabilize the variance also often make the population distribution nearly normal. For equal sample sizes, however, these transformations may not usually be necessary. Moreover, the use of such transformations may often result in different group means. It is possible that the means of the original scores are equal but the means of the transformed scores are not, and vice versa. Further, the means of transformed scores are often changed in ways that are not intuitively meaningful or are difficult to interpret.

Acknowledgment

We are grateful to Professor Miodrag Lovric, editor-in chief, for his nice editorial efforts and for his valuable suggestions and comments that led to the improvement in the contribution.

About the Authors

Dr. Hardeo Sahai held Professorial and visiting Professorial positions at the University of Puerto Rico, Mayaguez and San Juan, University of Ceara, Brazil, University of Granada, Spain, University of Veracruzana, Mexico, University of Nacional de Colombia, University of Nacional de Trujillo, Peru. He has received the University of Kentuky Outstanding Alumnus award, Medal of honor University of Granada (Spain), Plaque of honor University of Nacional de Trujillo (Peru). He has published over 150 papers in the statistical (bio)medical and epidemiological literature and is coauthor of the several books which include: *Statistics in Epidemiology: Methods, Techniques and Applications* (with Anwer Khurshid, CRC Press, Boca Raton, Florida, 1996), and *Analysis of Variance for Random Models, Vol. 1: Balanced Data and Vol. 2: Unbalanced Data* (with Mario M. Ojeda, Birkhäuser 2004). Professor Sahai is a Fellow of the American Statistical Association, Royal Statistical Society and Elected Member of the International Statistical Institute.

Dr. Mohammed Ibrahim Ali Ageel was a Full Professor and Chairman of Mathematics, Department of Mathematics, King Saud University (KSU), Saudi Arabia, and later the Chairman of Mathematics Department, King Khalid University (KKU), Saudi Arabia. He was the Dean of Graduate School, King Khalid University, Saudi Arabia. Professor Ageel was also a Full Professor of Mathematics and Dean of Engineering, Najran University (NU), Saudi Arabia. Ageel is currently a Full Professor of Mathematics, Jazan University, Jazan, Saudi Arabia. He is a Founder and President of Saudi Association of Statistical Sciences (SASS). He is an Elected member of the International Statistical Institute, and was elected as a Fellow of the Royal Statistical Society. He has published more than 50 research papers and articles in both theoretical and applied areas. Professor Ageel is a coauthor of the book *The Analysis of Variance: Fixed, Random and Mixed Models* (with Hardeo Sahai, Birkhäuser, Boston 2000).

Anwer Khurshid is a Professor at the Department of Statistics, University of Karachi, Pakistan. During 2004–2010 he had a faculty position at the Sultan Qaboos University, Oman and University of Nizwa, Oman. He is author or coauthor of more than 70 papers and two books (both with Professor Hardeo Sahai). In recognition of his teaching and research contributions Professor Khurshid was awarded a certificate of appreciation in 1997 by the Chancellor, University of Karachi, Pakistan.

Cross References

►Analysis of Variance
►Bartlett's Test
►Heteroscedasticity
►Normality Tests
►Robust Inference
►Robust Statistics
►Tests for Homogeneity of Variance

References and Further Reading

Arvesen JN, Layard MWJ (1975) Asymptotically robust tests in unbalanced variance component models. Ann Stat 3:1122–1134

Arvesen JN, Schmitz TH (1970) Robust procedures for variance component problems using the jackknife. Biometrics 26:677–686

Barnett VD, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, New York

Bartlett MS (1936) The square root transformation in the analysis of variance. J R Stat Soc 3:68–78

Bartlett MS (1937a) Properties of sufficiency and statistical tests. Proc R Soc Lond Ser A 160:268–282

Bartlett MS (1937b) Some examples of statistical methods of research in agriculture and applied biology. J R Stat Soc Suppl 4: 137–183

Boneau CA (1960) The effects of violation of assumptions underlying the t-test. Psychol Bull 57:49–64

Boneau CA (1962) A comparison of the power of the U and t tests. Psychol Rev 59:246–256

Box GEP (1954) Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. Ann Math Stat 25:290–302

Box GEP, Anderson SL (1955) Permutation theory in the derivation of robust criteria and the study of departures from assumption. J R Stat Soc Ser B 17:1–26

Bradley JV (1964) Studies in research methodology, VI. The central limit effect for a variety of populations and the robustness of Z, t, and F. Technical report no. 7 AMRL-54-123. Aerospace Medical Research Laboratories, Wright-Patterson Air Force Base, Dayton, Ohio

Brown RA (1974) Robustness of the studentized range statistic. Biometrika 61:171–175

Brown MB, Forsythe AB (1974a) Robust tests for the equality of variances. J Am Stat Assoc 69:364–367

Brown MB, Forsythe AB (1974b) The small size sample behavior of some statistics which test the equality of several means. Technometrics 16:129–132

Brown MB, Forsythe AB (1974c) The ANOVA and multiple comparisons for data with heterogeneous variances. Biometrics 30:719–724

Cochran WG (1941) The distribution of the largest of a set of estimated variances as a fraction of their total. Ann Eugen 11:47–52

Conover WJ, Johnson ME, Johnson MM (1981) A comparative study of tests for homogeneity of variances with applications to outer continental shelf bidding data. Technometrics 23:351–361 (Corrigendum Technometrics 26:302)

D'Agostino RB (1971) An omnibus test of normality for moderate and large size samples. Biometrika 58:341–348

Donaldson TS (1968) Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. J Am Stat Assoc 63:660–676

Draper NR, Smith H (1981) Applied regression analysis, 2nd edn. Wiley, New York

Dunnett CW (1982) Robust multiple comparisons. Commun Stat Part A: Theory Methods 11:2611–2629

Gayen AK (1950) The distribution of the variance ratio in random samples of any size drawn from non-normal universes. Biometrika 37:236–255

Geary RC (1947) Testing for normality. Biometrika 34:209–242

Glass GV, Peckham PD, Sanders JR (1972) Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Rev Educ Res 42:239–288

Hampel FR, Rochetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York

Horsnell G (1953) The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika 40: 128–136

Hoyle MH (1973) Transformations – an introduction and a bibliography. Int Stat Rev 41:203–223

James GS (1951) The comparison of several groups of observations when the ratio of population variances are unknown. Biometrika 38:324–329

Keselman HJ, Rogan JC (1978) A comparison of modified-Tukey and Scheffé methods of multiple comparisons for pairwise contrasts. J Am Stat Assoc 73:47–51

Kohr RL, Games PA (1974) Robustness of the analysis of variance, the Welch procedure, and a Box procedure to heterogeneous variances. J Exp Educ 43:61–69

Krutchkoff RG (1988) One way fixed effects analysis of variance when the error variances may be unequal. J Stat Comput Simul 30:259–271

Lange N, Ryan L (1989) Assessing normality in random effects models. Ann Stat 17:624–642

Levene H (1960) Robust tests for equality of variances. In: Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB (eds) Contributions to probability and statistics. Stanford University Press, Stanford, pp 278–292

Miller RG Jr (1968) Jackknifing variances. Ann Math Stat 39: 567–582

O'Brien RG (1979) An improved ANOVA method for robust tests of additive models for variances. J Am Stat Assoc 74:877–880

O'Brien RG (1981) A simple test for variance effects in experimental designs. Psychol Bull 89:570–574

Olejnik SF, Algina J (1987) Type I error rates and power estimates of selected parametric and nonparametric tests of scales. J Educ Stat 12:45–61

Pearson ES (1931) The analysis of variance in cases of non-normal variation. Biometrika 23:114–133

Petrinovich LF, Hardyck CD (1969) Error rates for multiple comparison methods. Psychol Bull 71:43–54

Ramsey PH (1994) Testing variances in psychological and educational research. J Educ Stat 19:23–42

Ringland JT (1983) Robust multiple comparisons. J Am Stat Assoc 78:145–151

Sahai H, Ageel MI (2000) The analysis of variance: fixed, random and mixed models. Birkhäuser/Springer, Boston

Sahai H, Ojeda M (2005) Analysis of variance for random models: unbalanced data. Birkhauser, USA

Scheffé H (1959) The analysis of variance. Wiley, New York

Shapiro SS, Francia RS (1972) An approximate analysis of variance test for normality. J Am Stat Assoc 67:215–216

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52:591–611

Singhal RA, Sahai H (1992) Sampling distribution of the ANOVA estimator of between variance component in samples from a non-normal universe. J Stat Comput Simul 43:19–30

Singhal RA, Tiwari CB, Sahai H (1988) A selected and annotated bibliography on the robustness studies to non-normality in variance component models. J Jpn Stat Soc 18: 195–206

Snedecor GW, Cochran WG (1989) Statistical methods, 8th edn. Iowa State University Press, Ames

Srivastava ABL (1959) Effects of non-normality on the power of the analysis of variance test. Biometrika 46:114–122

Tan WY, Tabatabai MA (1986) Some Monte Carlo studies on the comparison of several means under heteroscedasticity and robustness with respect to departure from normality. Biom J 28:801–814

Tan WY, Wong SP (1980) On approximating the null and non-null distributions of the F ratio in unbalanced random effects models from non-normal universes. J Am Stat Assoc 75: 655–662

Thöni H (1967) Transformation of variables used in the analysis of experimental and observational data: a review. Technical report no. 7. Statistical Laboratory, Iowa State University, Ames

Tiku ML (1964) Approximating the general non-normal variance-ratio sampling distributions. Biometrika 51:83–95

Tiku ML (1971) Power function of F-test under non-normal situations. J Am Stat Assoc 66:913–916

Welch BL (1956) On linear combinations of several variances. J Am Stat Assoc 51:132–148

Anderson–Darling Tests of Goodness-of-Fit

Theodore W. Anderson
Professor of Statistics and Economics, Emeritus
Stanford University, Stanford, CA, USA

Introduction

A "goodness-of-fit" test is a procedure for determining whether a sample of n observations, $x_1, \ldots, x_n$, can be considered as a sample from a given specified distribution. For example, the distribution might be a normal distribution with mean 0 and variance 1. More generally, the specified distribution is defined as

$$F(x) = \int_{-\infty}^{x} f(y)dy, \quad -\infty < x < \infty, \tag{1}$$

where $f(y)$ is a specified density. This density might be suggested by a theory, or it might be determined by a previous study of similar data.

When X is a random variable with distribution function $F(x) = \Pr\{X \le x\}$, then $U = F(X)$ is a random variable with distribution function

$$\Pr\{U \le u\} = \Pr\{F(X) \le u\} = u, \quad 0 \le u \le 1. \tag{2}$$

The model specifies $u_1 = F(x_1), \ldots, u_n = F(x_n)$ as a sample from the distribution (2), that is, the standard uniform distribution (see ▶Uniform Distribution in Statistics) on the unit interval $[0,1]$ written $U(0,1)$.

A test of the hypothesis that $x_1, \ldots, x_n$ is a sample from a specified distribution, say $F^0(x)$, is equivalent to a test that $u_1 = F^0(x_1), \ldots, u_n = F^0(x_n)$ is a sample from $U(0,1)$. Define the *empirical distribution function* as

$$F_n(x) = \frac{k}{n}, \quad -\infty < x < \infty, \tag{3}$$

if k of $(x_1, \ldots, x_n)$ are $\le x$. A goodness-of-fit test is a comparison of $F_n(x)$ with $F^0(x)$. The hypothesis $H_0 : F(x) = F^0(x)$, $-\infty < x < \infty$, is rejected if $F_n(x)$ is very different from $F^0(x)$. "Very different" is defined here as

$$W_n^2 = n\int_{-\infty}^{\infty} \left[F_n(x) - F^0(x)\right]^2 \psi\left[F^0(x)\right] dF^0(x)$$
$$= n\int_{-\infty}^{\infty} \left[F_n(x) - F^0(x)\right]^2 \psi\left[F^0(x)\right] f^0(x)dx \tag{4}$$

being large; here (1) holds and $\psi(z)$ is a weight function such that $\psi(z) \ge 0$, and $f^0(x)$ is the density of $F^0(x)$.

If $\psi(z) = 1$, the statistic W_n^2 is the Cramér-von Mises statistic, denoted by $n\omega^2$. Anderson and Darling (1952) gave a table of the limiting distribution of $n\omega^2$ as $n \to \infty$. For example, the 5% significance point is .46136 and the 1% significance point is .74346.

The Anderson–Darling Statistic

For a given x and hypothetical distribution $F^0(\cdot)$, the random variable $n F_n(x)$ has a ▶binomial distribution with probability $F^0(x)$. The expected value of $n F_n(x)$ is $nF^0(x)$ and the variance is $n F^0(x)\left[1 - F^0(x)\right]$. The definition of the goodness-of-fit statistic (4) permits the choice of weight function $\psi(\cdot)$. In particular the investigator may want to emphasize the tails of the presumed distribution $F^0(x)$. In that case the choice is

$$\psi(u) = \frac{1}{u(1-u)}. \tag{5}$$

Then for a specified x

$$\sqrt{n}\frac{F_n(x) - F^0(x)}{\sqrt{F^0(x)\left[1 - F^0(x)\right]}} \tag{6}$$

has mean 0 and variance 1 when the null hypothesis is true. The Anderson–Darling statistic is

$$A_n^2 = n\int_{-\infty}^{\infty} \frac{\left[F_n(x) - F^0(x)\right]^2}{F^0(x)\left[1 - F^0(x)\right]} dF^0(x). \tag{7}$$

It was shown in Anderson and Darling (1954) that (7) can be written as

$$A_n^2 = -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)\left[\log u_{(j)} + \log\left(1 - u_{(n-j+1)}\right)\right] \tag{8}$$

where $u_{(j)} = F^0\left(x_{(j)}\right)$ and $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ is the ordered sample.

Anderson and Darling found the limiting distribution of A_n^2 [for weight function (5)]. In the next section the development of this distribution is outlined. The 5% significance point of the limiting distribution is 2.492 and the 1% point is 3.880. The mean of this limiting distribution is 1 and the variance is $2(\pi^2 - 9)/3 \sim .57974$.

Outline of Derivation

Let $u = F^0(x)$, $u_i = F^{(0)}(x_i)$, $i = 1, \ldots, n$, and $u_{(i)} = F^{(0)}(x_{(i)})$, $i = 1, \ldots, n$. Let $G_n(u)$ be the empirical distribution function of $u_1, \ldots, u_n$; that is

$$G_n(u) = \frac{k}{n}, \quad 0 \le u \le 1, \tag{9}$$

if k of $u_1, \ldots, u_n$ are $\le u$. Thus

$$G_n\left[F^0(x)\right] = F_n^0(x), \tag{10}$$

and

$$W_n^2 = n \int_0^1 [G_n(u) - u]^2 \psi(u) du, \tag{11}$$

when the null hypothesis $F(x) = F^{(0)}(x)$ is true. For every u $(0 \le u \le 1)$

$$Y_n(u) = \sqrt{n}\,[G_n(u) - u] \tag{12}$$

is a random variable, and the set of these may be considered as a stochastic process with parameter u. Thus

$$\Pr\left\{W_n^2 \le z\right\} = \Pr\left\{\int_0^1 Y_n^2(u)\psi(u)du \le z\right\} = A_n(z), \tag{13}$$

say. For a fixed set $u_1, \ldots, u_k$ the k-variate distribution of $Y_n(u_1), \ldots, Y_n(u_k)$ approaches a multivariate normal distribution (see ▶Multivariate Normal Distributions) as $n \to \infty$ with mean and covariance function

$$\mathcal{E}[Y_n(u)] = 0, \quad \mathcal{E}Y_n(u)Y_n(v) = \min(u,v) - uv. \tag{14}$$

The limiting process of $\{Y_n(u)\}$ is a Gaussian process $y(u)$, $0 \le u \le 1$, and $\mathcal{E}y(u) = 0$ and $\mathcal{E}y(u)y(v) = \min(u,v) - uv$. Let

$$a(z) = \Pr\left\{\int_0^1 y^2(u)\psi(u)du \le z\right\}. \tag{15}$$

Then $A_n(z) \to a(z)$, $0 \le z < \infty$. The mathematical problem for the Anderson–Darling statistic is to find the distribution function $a(z)$ when $\psi(u) = 1/u(1-u)$.

We briefly sketch the procedure to find the distribution of $\int_0^1 z^2(u)du$, where $z(u)$ is a Gaussian stochastic process with $\mathcal{E}z(u) = 0$ and $\mathcal{E}z(u)z(v) = k(u,v)$. When the kernel is continuous and square integrable (as is the case here), it can be written as

$$k(u,v) = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} f_j(u) f_j(v), \tag{16}$$

where λ_j is an eigenvalue and $f_j(u)$ is the corresponding normalized eigenfunction of the integral equation

$$\lambda \int_0^1 k(u,v) f(u) du = f(v), \tag{17}$$

$$\int_0^1 f_j^2(u)du = 1, \quad \int_0^1 f_i(u)f_j(u)du = 0, \quad i \ne j. \tag{18}$$

Then the process can be written

$$z(u) = \sum_{j=1}^{\infty} \frac{1}{\sqrt{\lambda_j}} X_j f_j(u), \tag{19}$$

where $X_1, X_2, \ldots,$ are independent $N(0,1)$ variables. Then

$$\int_0^1 z^2(u)du = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} X_j^2, \tag{20}$$

with characteristic function

$$\mathcal{E}\exp\left[it\int_0^1 z^2(u)du\right] = \prod_{j=1}^{\infty} \mathcal{E}\left(\exp itX_j^2/\lambda_j\right) = \prod_{j=1}^{\infty}\left(1 - 2it/\lambda_j\right)^{-\frac{1}{2}}. \tag{21}$$

The process $Y_n^*(u) = \sqrt{\psi(u)}\,Y_n(u)$ has covariance function

$$k(u,v) = \sqrt{\psi(u)}\sqrt{\psi(v)}\,[\min(u,v) - uv]; \tag{22}$$

as $n \to \infty$, the process $Y_n^*(u)$ approaches $y^*(u) = \sqrt{\psi(u)}\,y(u)$ with covariance (22). The characteristic function of the limiting distribution of $n\omega^2$ is

$$\sqrt{\frac{\sqrt{2it}}{\sin\sqrt{2it}}} \tag{23}$$

for $\psi(u) = 1$, and that of the limiting distribution of A_n^2 is

$$\sqrt{\frac{-2\pi it}{\cos\left(\frac{\pi}{2}\sqrt{1+8it}\right)}}. \tag{24}$$

for $\psi(u) = 1/u(1-u)$.

The integral equation (17) can be transformed to a differential equation

$$h''(t) + \lambda\psi(t)\,h((t) = 0. \tag{25}$$

Anderson–Darling Tests with Unknown Parameters

When parameters in the tested distribution are not known, but are estimated efficiently, the covariance (14) is modified, and the subsequent limiting distribution theory for both $n\omega^2$ and A_n^2 follows the same lines as above, with this new covariance. If the parameters are location and/or scale, the limiting distributions do not depend on the true parameter values, but depend on the class of tested distributions. If the parameters are shape parameters, the limiting distribution depends on shape. Limiting distributions have been evaluated and percentage points given for a number of different tested distributions; see Stephens (1976, 1986). Tests for three parameter Weibull, and von Mises have been given by Lockhart and Stephens (1985, 1994).

The percentage points for these tests are much smaller than those given above for the case when parameters are known.

Acknowledgments

The assistance of Michael A. Stephens is gratefully acknowledged.

About the Author

Professor Anderson was born June 5, 1918 in Minneapolis, Minnesota. He is Past President, Institute of Mathematical Statistics (1963), Vice President, American Statistical Association (1971–1973), Fellow of the American Academy of Arts and Sciences (elected 1974), Member of the National Academy of Sciences (elected 1976). Professor Anderson has been awarded the R. A. Fisher Award of Committee of Presidents of Statistical Societies (1985) and Samuel S. Wilks Memorial Medal, American Statistical Association (1988). He holds four honorary doctorates. Professor Anderson has published over 170 articles in statistical, econometric, and mathematical journals, and seven books, including the internationally recognized text *An Introduction to Multivariate Statistical Analysis* (1958, Wiley; 3rd edition 2003). The 17th International Workshop in Matrices and Statistics was held in Tomar (Portugal July 2008), in honour of Professor Theodore Wilbur Anderson 90th birthday. *The Collected Papers of T. W. Anderson: 1943–1985* (edited by George P. H. Styan) were published by Wiley in 1990, comprising 109 papers and 16 commentaries, in a 2-volume set covering 1,681 pages.

Cross References

►Cramér-Von Mises Statistics for Discrete Distributions
►Jarque-Bera Test
►Kolmogorov-Smirnov Test
►Normality Tests
►Normality Tests: Power Comparison
►Omnibus Test for Departures from Normality
►Tests of Fit Based on The Empirical Distribution Function

References and Further Reading

Anderson TW, Darling DA (1952) Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes. Ann Math Stat 23:193–212
Anderson TW, Darling DA (1954) A test of goodness-of-fit. J Am Stat Assoc 49:765–769
Lockhart RA, Stephens MA (1985) Tests of fit for the von-Mises distribution. Biometrika 72:647–652
Lockhart RA, Stephens MA (1994) Estimation and tests of fit for the three-parameter Weibull distribution. J R Stat Soc B 56:491–500
Stephens MA (1976) Asymptotic results for goodness-of-fit statistics with unknown parameters. Ann Stat 4:357–369
Stephens MA (1986) In: D'Agostino R, Stephens MA (eds) Goodness-of-fit techniques, chap. 4. Marcel Dekker, New York

Approximations for Densities of Sufficient Estimators

Juan Carlos Abril
President of the Argentinean Statistical Society, Professor Universidad Nacional de Tucumán and Consejo Nacional de Investigaciones Científicas y Técnicas, San Miguel de Tucumán, Argentina

Introduction

Durbin (1980a) proposed a simple method for obtaining asymptotic expansions for the densities of sufficient estimators. The expansion is a series which is effectively in powers of n^{-1}, where n is the sample size, as compare with the ►Edgeworth expansion which is in powers of $n^{-1/2}$. The basic approximation is just the first term of this series. This has an error of order n^{-1} compare to the error of $n^{-1/2}$ in the usual asymptotic normal approximation (see ►Asymptotic Normality). The order of magnitude of the error can generally be reduced to order $n^{-3/2}$ by renormalization.

Suppose that the real m-dimensional random vector $\mathbf{S}_n = (S_{1n}, S_{2n}, \ldots, S_{mn})'$ has a density with respect to Lebesgue measure which depends on integer $n > N$ for some positive N and on $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is a subset of $\mathbb{R}^q$ for q an arbitrary positive integer.

Let

$$\mathbf{D}_n(\boldsymbol{\theta}) = n^{-1} E\left\{\mathbf{S}_n - E(\mathbf{S}_n)\right\}\left\{\mathbf{S}_n - E(\mathbf{S}_n)\right\}' \qquad (1)$$

which we assume is finite and positive-definite for all n and $\boldsymbol{\theta}$, and which we assume converges to a finite positive-definite matrix $\mathbf{D}(\boldsymbol{\theta}_0)$ as $n \to \infty$ and $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is a particular value of $\boldsymbol{\theta}$, usually the true value.

Let $\phi_n(\mathbf{z}, \boldsymbol{\theta}) = E(e^{i\mathbf{z}'\mathbf{S}_n})$ be the characteristic function of $\mathbf{S}_n$ where $\mathbf{z} = (z_1, z_2, \ldots, z_m)'$. Whenever the appropriate derivatives exists, let

$$\frac{\partial^j \log \phi_n(\widetilde{\mathbf{z}}, \boldsymbol{\theta})}{\partial \mathbf{z}^j}$$

denote the set of jth order derivatives $\partial^j \log \phi_n(\mathbf{z},\boldsymbol{\theta})/\partial z_1^{j_1}\cdots\partial z_m^{j_m}$ for all integers $j_1, j_2, \ldots, j_m \geq 0$ satisfying $\sum_k j_k = j$, evaluated at $\mathbf{z} = \tilde{\mathbf{z}}$. The jth cumulant $\boldsymbol{\kappa}_{nj}(\boldsymbol{\theta})$ of $\mathbf{S}_n$, where it exists, satisfies the relation

$$i^j \boldsymbol{\kappa}_{nj}(\boldsymbol{\theta}) = \frac{\partial^j \log \phi_n(\mathbf{0},\boldsymbol{\theta})}{\partial \mathbf{z}^j}. \tag{2}$$

In what follows, let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ be points in an open subset $\boldsymbol{\Theta}_0$ of $\boldsymbol{\Theta}$, and let r be a specified integer. We use the word limit in the sense of joint limit, and introduce three assumptions.

Assumption 1. If n is large enough $|\phi_n(\mathbf{z},\boldsymbol{\theta})|$ is integrable over $\mathbb{R}^m$, and if δ_1 is an arbitrary positive constant the limit of

$$n^{\frac{r}{2}-1} \int_{B_{\delta_1\sqrt{n}}} \left|\phi_n\left(\mathbf{z}/\sqrt{n},\boldsymbol{\theta}\right)\right| d\mathbf{z},$$

as $n \to \infty$ and $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$ is zero, where $B_{\delta_1\sqrt{n}}$ is the region $\|\mathbf{z}\| \geq \delta_1\sqrt{n}$ and $\|\cdot\|$denotes the Euclidean norm.

Assumption 2. The rth derivative $\partial^r \log \phi_n(\mathbf{z},\boldsymbol{\theta})/\partial \mathbf{z}^r$ exists for $\mathbf{z}$ in a neighborhood of the origin and the limit of

$$n^{-1}\frac{\partial^r \log \phi_n(\mathbf{z},\boldsymbol{\theta})}{\partial \mathbf{z}^r}$$

as $n \to \infty$, $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$ and $\|\mathbf{z}\| \to 0$ exists.

Assumption 3. The cumulant $\boldsymbol{\kappa}_{nj}(\boldsymbol{\theta}) = O(n)$ uniformly for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$ for $j = 3, \ldots, r-1$.

Now we present the Edgeworth expansion and the corresponding approximation to the density $h_n(\mathbf{x},\boldsymbol{\theta})$ of $\mathbf{X}_n = n^{-1/2}E\{\mathbf{S}_n - E(\mathbf{S}_n)\}$. Suppose that there is an integer $r \geq 3$ such that Assumptions 1–3 hold. Then there is a neighborhood $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta_2$ of $\boldsymbol{\theta}_0$ such that

$$h_n(\mathbf{x},\boldsymbol{\theta}) - \widehat{h}_n(\mathbf{x},\boldsymbol{\theta}) = o\left\{n^{-(r/2)+1}\right\} \tag{3}$$

uniformly in $\mathbf{x}$ and in $\boldsymbol{\theta}$ for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta_2$, where

$$\widehat{h}_n(\mathbf{x},\boldsymbol{\theta}) = \frac{|\mathbf{D}_n(\boldsymbol{\theta})|^{-1/2}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}\mathbf{x}'\mathbf{D}_n^{-1}(\boldsymbol{\theta})\mathbf{x}\right\} \left\{1 + \sum_{j=3}^{r} n^{-(j/2)+1} P_{nj}(\mathbf{x},\boldsymbol{\theta})\right\}, \tag{4}$$

and where $P_{nj}(\mathbf{x},\boldsymbol{\theta})$ is a generalized Edgeworth polynomial of order j the definition of which is given in Durbin (1980a). The practical construction of $P_{nj}(\mathbf{x},\boldsymbol{\theta})$ is described by Chambers (1967, pp. 368–369).

Approximations to the Densities of Sufficient Estimators

Suppose that $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$ is a matrix of observations of n continuos or discrete random $\ell \times 1$ vectors, not necessarily independent or identically distributed, with density

$$f(\mathbf{y},\boldsymbol{\theta}) = G(\mathbf{t},\boldsymbol{\theta})H(\mathbf{y}), \qquad \mathbf{y} \in \mathcal{Y}, \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{5}$$

where $\mathbf{t} = (t_1, \ldots, t_m)'$ is the value computed from $\mathbf{y}$ of an estimator $\mathbf{T}_n$ of the m-dimensional parameter $\boldsymbol{\theta}$, where $\mathcal{Y}$ and $\boldsymbol{\Theta}$ are observation and parameter spaces and where $\mathcal{Y}$ and H do not depend upon $\boldsymbol{\theta}$. We assume that $f(\mathbf{y},\boldsymbol{\theta}) > 0$ for all $\mathbf{y} \in \mathcal{Y}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. By the factorization theorem $\mathbf{T}_n$ is sufficient for $\boldsymbol{\theta}$.

Suppose that a transformation $\mathbf{y}_1, \ldots, \mathbf{y}_n \to t_1, \ldots, t_m, u_{m+1}, \ldots, u_{n\ell}$ exists such that on substituting for $\mathbf{y}$ on the right-hand side of (5) and integrating or summing out $u_{m+1}, \ldots, u_{n\ell}$ we obtain the marginal density $g(\mathbf{t},\boldsymbol{\theta})$ of $\mathbf{T}_n$ in the form $g(\mathbf{t},\boldsymbol{\theta}) = G(\mathbf{t},\boldsymbol{\theta})H_1(\mathbf{t})$ where H_1 does not depend upon $\boldsymbol{\theta}$. We therefore have

$$f(\mathbf{y},\boldsymbol{\theta}) = g(\mathbf{t},\boldsymbol{\theta})h(\mathbf{y}), \tag{6}$$

where $h(\mathbf{y}) = H(\mathbf{y})/H_1(\mathbf{t})$. The derivation of (6) from (5) has been given in this form to avoid measure-theoretic complications.

Suppose further that although functions $G(\mathbf{t},\boldsymbol{\theta})$ satisfying (5) can be deduced immediately from inspection of $f(\mathbf{y},\boldsymbol{\theta})$, the density $g(\mathbf{t},\boldsymbol{\theta})$ is unknown and we want to obtain an approximation to it for a particular value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$. Since (6) holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ we have

$$f(\mathbf{y},\boldsymbol{\theta}_0) = g(\mathbf{t},\boldsymbol{\theta}_0)h(\mathbf{y}). \tag{7}$$

On dividing (7) by (6) the unknown factor $h(\mathbf{y})$ is eliminated and we obtain immediately

$$g(\mathbf{t},\boldsymbol{\theta}_0) = \frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\boldsymbol{\theta})}g(\mathbf{t},\boldsymbol{\theta}). \tag{8}$$

If we substitute $\mathbf{t}$ for $\boldsymbol{\theta}$ in (8), as is legitimate since we have assumed that $\mathbf{t} \in \boldsymbol{\Theta}$, we obtain

$$g(\mathbf{t},\boldsymbol{\theta}_0) = \frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\mathbf{t})}g(\mathbf{t},\mathbf{t}). \tag{9}$$

The basic idea is to obtain an approximation $\widehat{g}(\mathbf{t},\boldsymbol{\theta}_0)$ for $g(\mathbf{t},\boldsymbol{\theta}_0)$ by substituting a series approximation $\widehat{g}(\mathbf{t},\mathbf{t})$ for $g(\mathbf{t},\mathbf{t})$ in (9), giving

$$\widehat{g}(\mathbf{t},\boldsymbol{\theta}_0) = \frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\mathbf{t})}\widehat{g}(\mathbf{t},\mathbf{t}). \tag{10}$$

In effect, the method rescales the approximation $\widehat{g}(\mathbf{t},\mathbf{t})$ at $\boldsymbol{\theta} = \mathbf{t}$ by the likelihood ratio $f(\mathbf{y},\boldsymbol{\theta}_0)/f(\mathbf{y},\mathbf{t})$.

A second idea is to substitute an Edgeworth series approximation $\widehat{g}(\mathbf{t},\widetilde{\boldsymbol{\theta}})$ for $g(\mathbf{t},\boldsymbol{\theta})$ in (8), where $\widetilde{\boldsymbol{\theta}}$ is chosen as the value of $\boldsymbol{\theta}$ for which the mean of the distribution of $\mathbf{T}_n$ coincides with $\mathbf{t}$. The reason for using this indirect approach instead of approximating $g(\mathbf{t},\boldsymbol{\theta})$ directly is that a straightforward Edgeworth approximation of $g(\mathbf{t},\boldsymbol{\theta})$, would normally be in powers of $n^{-1/2}$ whereas an Edgeworth approximation of $g(\mathbf{t},\mathbf{t})$ or $\widehat{g}(\mathbf{t},\widetilde{\boldsymbol{\theta}})$ is normally a series in powers of n^{-1}.

Suppose that $E(\mathbf{T}_n) = \boldsymbol{\theta} - \boldsymbol{\beta}_n(\boldsymbol{\theta})$, where $\boldsymbol{\beta}_n(\boldsymbol{\theta}) = O(n^{-1})$ uniformly for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, and that $n\mathbf{T}_n = \mathbf{S}_n$, where $\mathbf{S}_n$ satisfies the Assumptions 1–3 given above with $r = 4$. Maximum likelihood estimators often satisfy these assumptions. We make the following further assumption:

Assumption 4. Uniformly for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$,

$$|\mathbf{D}_n(\boldsymbol{\theta})| = |\mathbf{D}(\boldsymbol{\theta})|\left\{1 + O(n^{-1})\right\}.$$

The assumption is, of course, satisfied when $\mathbf{S}_n$ is a sum of independent and identically distributed vectors but it is also satisfied in other cases of interest, notably in some applications in time series analysis. We suppose that we require a single-term approximation which has an error of order n^{-1} at most.

Since $\mathbf{X}_n = n^{-1/2}E\left\{\mathbf{S}_n - E(\mathbf{S}_n)\right\} = \sqrt{n}\left\{\mathbf{T}_n - \boldsymbol{\theta} + \boldsymbol{\beta}_n(\boldsymbol{\theta})\right\}$, the value of $\mathbf{X}_n$ when $\mathbf{T}_n = \mathbf{t}$ and $\boldsymbol{\theta} = \mathbf{t}$ is $\mathbf{x} = \boldsymbol{\beta}_n(\boldsymbol{\theta})\sqrt{n}$. With $r = 4$, () gives

$$\widehat{h}_n(\mathbf{x},\mathbf{t}) = \frac{|\mathbf{D}_n(\mathbf{t})|^{-1/2}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}n\boldsymbol{\beta}_n(\mathbf{t})'\mathbf{D}_n^{-1}(\mathbf{t})\boldsymbol{\beta}_n(\mathbf{t})\right\} \times\left[1 + \sum_{j=3}^{4} n^{-(j/2)+1}P_{nj}\left\{\boldsymbol{\beta}_n(\boldsymbol{\theta})\sqrt{n},\mathbf{t}\right\}\right]. \tag{11}$$

Now $n\boldsymbol{\beta}_n(\mathbf{t})'\mathbf{D}_n^{-1}(\mathbf{t})\boldsymbol{\beta}_n(\mathbf{t}) = O(n^{-1})$ and the constant term of P_{n4} is $O(1)$. Moreover P_{n3} contains no constant term and hence is $O(n^{-1/2})$ when $\mathbf{x} = \boldsymbol{\beta}_n(\boldsymbol{\theta})\sqrt{n}$. We note that these orders of magnitude are uniform for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\theta}_0$ under the Assumptions 1–3. Because of Assumption 4, we have

$$\widehat{h}_n(\mathbf{x},\mathbf{t}) = \frac{|\mathbf{D}(\mathbf{t})|^{-1/2}}{(2\pi)^{m/2}}\left\{1 + O(n^{-1})\right\}$$

uniformly for $\mathbf{t}$ in a neighborhood of $\boldsymbol{\theta}_0$.

Let $h_n(\mathbf{x},\mathbf{t})$ be the true density of $\mathbf{X}_n$, then by (3)

$$\begin{aligned} h_n(\mathbf{x},\mathbf{t}) &= \widehat{h}_n(\mathbf{x},\mathbf{t}) + o(n^{-1}) \\ &= \frac{|\mathbf{D}(\mathbf{t})|^{-1/2}}{(2\pi)^{m/2}}\left\{1 + O(n^{-1})\right\} + o(n^{-1}). \end{aligned} \tag{12}$$

Since the term $o(n^{-1})$ is uniform for $\mathbf{t}$ in a neighborhood of $\|\mathbf{t} - \boldsymbol{\theta}_0\| < \delta_2$, where $\delta_2$is a suitably chosen positive constant independent of n, and since $|\mathbf{D}(\mathbf{t})|$ is continuous at $\boldsymbol{\theta}_0$ and hence is bounded away from zero for $\mathbf{t}$ in the neighborhood, the term $o(n^{-1})$ of (12) can be absorbed inside the curly bracket. We thus have uniformly

$$h_n(\mathbf{x},\mathbf{t}) = \frac{|\mathbf{D}(\mathbf{t})|^{-1/2}}{(2\pi)^{m/2}}\left\{1 + O(n^{-1})\right\}.$$

Transforming from $\mathbf{x}$ to $\mathbf{t}$ we obtain for the density of $\mathbf{T}_n$ at $\mathbf{T}_n = \boldsymbol{\theta} = \mathbf{t}$,

$$g(\mathbf{t},\mathbf{t}) = \left(\frac{n}{2\pi}\right)^{m/2}|\mathbf{D}(\mathbf{t})|^{-1/2}\left\{1 + O(n^{-1})\right\}. \tag{13}$$

Substituting in (9) we obtain

$$g(\mathbf{t},\boldsymbol{\theta}_0) = \left(\frac{n}{2\pi}\right)^{m/2}|\mathbf{D}(\mathbf{t})|^{-1/2}\frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\mathbf{t})}\left\{1 + O(n^{-1})\right\}, \tag{14}$$

uniformly in $\mathbf{t}$ for $\|\mathbf{t} - \boldsymbol{\theta}_0\| < \delta_2$.

Expression (14) is the basic approximation for the density of the sufficient estimator $\mathbf{T}_n$. The fact that the error is a proportional error which is uniform over the region $\|\mathbf{t} - \boldsymbol{\theta}_0\| < \delta_2$ is important since the limiting probability that $\mathbf{T}_n$ falls outside this region is zero.

Assuming appropriate regularity conditions to be satisfied, $\mathbf{D}^{-1}(\boldsymbol{\theta})$ is the limiting mean information matrix $\mathcal{I}(\boldsymbol{\theta})$, where

$$\mathcal{I}(\boldsymbol{\theta}) = \lim_{n\to\infty} E\left[-n^{-1}\frac{\partial^2 \log f(\mathbf{y},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right].$$

We then have for the basic approximation

$$g(\mathbf{t},\boldsymbol{\theta}_0) = \left(\frac{n}{2\pi}\right)^{m/2}|\mathcal{I}(\mathbf{t})|^{1/2}\frac{f(\mathbf{y},\boldsymbol{\theta}_0)}{f(\mathbf{y},\mathbf{t})}\left\{1 + O(n^{-1})\right\}, \tag{15}$$

uniformly in $\mathbf{t}$ for $\|\mathbf{t} - \boldsymbol{\theta}_0\| < \delta_2$.

The simplicity of the structure of this approximation should be noted. It consists of the normal approximation to the density when $\boldsymbol{\theta} = \mathbf{t}$, namely $\{n/(2\pi)\}^{m/2}|\mathcal{I}(\mathbf{t})|^{1/2}$, multiplied by the likelihood ratio $f(\mathbf{y},\boldsymbol{\theta}_0)/f(\mathbf{y},\mathbf{t})$.

Durbin (1980a) proved that when either (14) or (15) is integrated over any subset of $\mathbb{R}^m$, the error term remains $O(n^{-1})$. This result is, in fact, of great importance in practical situations since it demonstrates that the basic approximation can be integrated for inference purposes with an error which is of order n^{-1} at most. He proved as well that when the constant term of the approximation (14), and consequently also of (15), is adjusted to make the integral over the whole space equal to unity, the order of magnitude of the error is often reduced from $O(n^{-1})$ to $O_x(n^{-3/2})$, where $O_x(n^{-q})$ denotes a quantity which is $O(n^{-q})$ for each fixed $\mathbf{x} = \sqrt{n}\{\mathbf{t} - E(\mathbf{T}_n)\}$ but which is not $O(n^{-q})$ uniformly for all $\mathbf{x}$. This process of adjusting the constant term is generally called renormalization.

About the Author

Professor Abril is co-editor of the *Revista de la Sociedad Argentina de Estadstica* (Journal of the Argentinean Statistical Society).

Cross References

►Approximations to Distributions
►Edgeworth Expansion
►Properties of Estimators
►Sufficient Statistics

References and Further Reading

Abril JC (1985) Asymptotic expansions for time series problems with applications to moving average models. Ph.D. Thesis, The London School of Economics and Political Science, University of London, England
Barndorff-Nielsen O, Cox DR (1979) Edgeworth and saddle-point approximations with statistical applications. J R Stat Soc B 41:279–312
Bhattacharya RN, Ghosh JK (1978) On the validity of the formal Edgeworth expansion. Ann Statist 6:434–451
Bhattacharya RN, Rao RR (1976) Normal approximation and asymptotic expansions. Wiley, New York
Chambers JM (1967) On Methods of asymptotic approximation for multivariate distributions. Biometrika 54:367–383
Daniels HE (1954) Saddlepoint approximations in statistics. Ann Math Statist 25:631–650
Daniels HE (1956) The approximate distribution of serial correlation coefficients. Biometrika 43:169–185
Durbin J (1980a) Approximations for the densities of sufficient estimators. Biometrika 67:311–333
Durbin J (1980b) The approximate distribution of partial serial correlation coefficient calculated from residual from regression on Fourier series. Biometrika 67:335–349
Feller W (1971) An introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York
Hampel FR (1973) Some small sample asymptotics. Proc Prague Symp Asymptotic Stat 2:109–126
Loève M (1977) Probability theory, vol I, 4th edn. Springer, Berlin
Phillips PCB (1978) Edgeworth and saddlepoint approximations in a first order autoregression. Biometrika 65:91–98
Wallace DL (1958) Asymptotic approximations to distributions. Ann Math Stat 29:635–654

Approximations to Distributions

Juan Carlos Abril
President of the Argentinean Statistical Society, Professor Universidad Nacional de Tucumán and Consejo Nacional de Investigaciones Científicas y Técnicas, San Miguel de Tucumán, Argentina

Introduction

The exact probability distribution of estimators for finite samples is only available in convenient form for simple functions of the data and when the likelihood function is completely specified. Frequently, these conditions are not satisfied and the inference is based on approximations to the sample distribution. Typically, large sample methods based on the central limit theorem (see ►Central Limit Theorems) are generally used. For example, if T_n is an estimator of the parameter θ based on a sample of size n, it is sometimes possible to obtain functions $\sigma(\theta)$ such that the distribution of the random variable $\sqrt{n}(T_n - \theta)/\sigma(\theta)$ converges to the standard normal distribution as n tends to infinity. In such a case, it is very common to approximate the distribution of T_n by a normal distribution with mean θ and variance $\sigma^2(\theta)/n$.

These asymptotic approximations can be good even for very small samples. The mean of independent draws from a rectangular distribution has a bell-shaped density for n as small as three. But it is easy to construct examples where the asymptotic approximation is bad even when the sample has hundreds of observations. It is therefore desirable to know the conditions under which the asymptotic approximations are reasonable and to have alternative methods available when these approximations do not work properly. Most of the material discussed here is closely related with the topic *Asymptotic, higher order* which is presented as well in this Encyclopedia.

There is a good literature treating the theory and practice of approximations to distributions, but introductory

texts are relatively few. A very brief summary can be seen in Bickel and Doksum (1977), while some discussion is given in Johnson and Kotz (1970). The extension to asymptotic expansions can be seen in the excellent paper by Wallace (1958), although it is outdated. For a good treatment of the subject, an incursion upon the advanced probability and numerical analysis textbooks is needed. For those with enough time and patience, Chaps. 15 and 16 of Feller (1971) are well worth reading.

The Central Limit Theorem

The center of a large part of the asymptotic theory is the central limit theorem, initially formulated for sums of independent random variables. Let $\{Y_n\}$ be a sequence of independent random variables. Denote by H_n the distribution function of the standardized sum

$$X_n = \frac{\sum_{j=1}^{n} \{Y_j - E(Y_j)\}}{\sqrt{\left\{\sum_{j=1}^{n} V(Y_j)\right\}}},$$

where $V(Y_j)$ is the variance of Y_j, and by $\mathcal{N}(.)$ the standard normal distribution function. The central limit theorem then states that $\lim H_n(x) = \mathcal{N}(x)$, as $n \to \infty$, for every fixed x, provided only that the means and variances are finite. If the $\{Y_j\}$ are not identically distributed, an additional condition guaranteeing that the distributions are not too unbalanced is necessary.

For time series problems, for example, where in general the variables are not independent, there have been particularized versions of this theorem guaranteeing the asymptotic behavior of statistics used in this area. Good references are the textbook by Anderson (1971), Brockwell and Davis (1991), Hannan (1970), and Priestley (1982) where one can find an excellent treatment of the asymptotic theory applied to time series problems.

Some authors have shown that the order of magnitude of the errors in the central limit theorem is $O(n^{-1/2})$.

While the central limit theorem is very useful theoretically and often in practice, it is not always satisfactory since for small or moderate n the errors of the normal approximation may be too large.

Curve Fitting

The most simplest form for obtaining an approximation to a distribution is to look for a family of curves with the correct shape and select the member that fits best. If the moments, specially those of low order, of the true distribution are known, they can be used in the fitting process. Otherwise one can use Monte Carlo simulations or any other information about the true distribution.

Durbin and Watson (1971) describe a number of different approximations to the null distribution of the statistic d used for testing serial correlation in regression analysis. One of the most accurate is the beta approximation proposed by Henshaw (1966). Since d is between zero and four and it seems to have a unimodal density, it is reasonable to think that a linear transformation from a beta distributed variable can be a good approximation to the true distribution. Suppose that Y is a random variable with beta distribution function

$$\Pr(Y \le y) = \frac{1}{B(p,q)} \int_0^y t^{p-1}(1-t)^{q-1} dt = G(y; p.q),$$

where

$$B(p,q) = \int_0^\infty t^{p-1}(1-t)^{q-1} dt.$$

Then, for a and b constant, the random variable $a + bY$ has moments depending on p, q, a and b. These moments are easy to express analytically. Moreover, the moments of the Durbin–Watson's statistic d are simple functions of the matrix of explanatory variables. Equating the first four moments of d with the corresponding moments of $a + bY$, one obtains four equations with four unknowns. For a given matrix of explanatory variables these equations give a unique solution, p^*, q^*, a^* and b^* say. So $\Pr(d \le y)$ can be approximated by $G\{(y - a^*)/b^*; p^*, q^*\}$. This approximation gives good results in many cases. Theil and Nagar (1961) proposed a similar approximation but using the approximated moments of d instead of the true moments. Since these approximated moments are independent of the matrix of explanatory variables, Theil–Nagar's approximation does not depend on the data and can be tabulated without any problem. Unfortunately the approximated moments are not always accurate and the resulting approximation to the distribution is less satisfactory than Henshaw's approximation.

If one has enough information over the true density, the curve fitting methods give simple and correct approximations. However these methods are not so attractive when the purpose is not quantitative but qualitative. The comparison of alternative procedures is difficult because the curve fitting methods does not produce, in general, parametric families of curves easily comparable. If two statistics are approximately normal, they can be compared by their means and variances. If one statistic is approximately beta and another is approximately normal, the comparison between them is not easy since the usual parameters that describe one of the distributions are

not of much interest for obtaining information about the other. The flexibility that makes the curve fitting method so accurate is, as well, an inconvenience for using it in comparisons.

Transformations

Suppose that Y is a random variable and b a monotonically increasing function such that $b(Y)$ has a distribution function H which can be approximated by $\widehat{H}$. Since $\Pr(Y \le y)$ is equal to $\Pr\{b(Y) \le b(y)\}$, the distribution function of Y can be approximated by $\widehat{H}\{b(y)\}$. A well known example of this technique is Fisher's z transformation. The sample correlation coefficient $\widehat{\rho}$ based on a random sample from a bivariate normal population is very far from symmetry when the true coefficient ρ is large in absolute value. But, $z = b(\widehat{\rho}) = 2^{-1}\log\{(1+\widehat{\rho})/(1-\widehat{\rho})\}$ is almost symmetric and can be approximated by a normally distributed random variable with mean $2^{-1}\log\{(1+\rho)/(1-\rho)\}$ and variance n^{-1}. Therefore $\Pr(\widehat{\rho} \le y)$ can be approximated by $\mathcal{N}\{\sqrt{n}b(y) - \sqrt{n}b(\rho)\}$ for moderate sample size n.

The use of transformations for approximating distributions is an art. Sometimes, as in the case of the correlation coefficient, the geometry of the problem can suggest the appropriate transformation b. Since $\widehat{\rho}$ can be interpreted as the cosine of the angle between two normally distributed random vectors, an inverse trigonometric transformation can be useful. In other cases, arguments based on approximations to the moments are helpful. Suppose that $b(Y)$ can be expanded as a power series about $\mu = E(Y)$

$$b(Y) = b(\mu) + b'(\mu)(Y-\mu) + \frac{1}{2}b''(\mu)(Y-\mu)^2 + \cdots,$$

where $Y - \mu$ is in some sense small. so we can do

$$\begin{aligned} E(b) &\approx b(\mu) + \frac{1}{2}b''(\mu)E(Y-\mu)^2, \\ V(b) &\approx \{b'(\mu)\}^2 V(Y), \\ E\{b - E(b)\}^3 &\approx \{b'(\mu)\}^3 E(Y-\mu)^3 \\ &\quad + \frac{3}{2}\{b'(\mu)\}^2 b''(\mu)E(Y-\mu)^4, \end{aligned}$$

and choose b in such a way that these approximates moments are equal to the moments of the approximated distribution. If the approximated distribution is normal, we can require that the variance $V(b)$ be a constant independent of μ; or we can require that the third order moment be zero. If the moments of Y are (almost) known and the above approximation is used, the criterion leads to differential equations in $b(\mu)$. Note that Fisher's transformation of $\widehat{\rho}$ stabilizes the approximated variance of b making it independent of ρ.

Jenkins (1954) and Quenouille (1948) apply inverse trigonometric transformations to the case of the autocorrelation coefficient in time series. The use of transformations in econometrics seems, however, to be minimum due mainly to the fact that the method is closely related with univariate distributions.

Asymptotic Expansions

Frequently it is possible to decompose the problem of finding the distribution in a sequence of similar problems. If the sequence has a limit which can easily be found, one can obtain an approximation to the solution of the original problem by a solution of the limit problem. The sequence of the problem is indexed by a parameter, which usually is the sample size n. Suppose for instance that we want an approximation to the distribution of an estimator, computed from a sample, of a parameter θ. We define an infinite sequence $\widehat{\theta}_n$ of estimators, one for each sample size $n = 1, 2, \ldots$, and we consider the problem of obtaining the distribution of each $\widehat{\theta}_n$. Of course, it is necessary to have some description of the joint distribution of the observations for each n. Given such a sequence of problems, the asymptotic approach implies three steps:

(a) To look for a simple monotonic transformation $X_n = b(\widehat{\theta}_n; \theta, n)$ such that the estimator X_n is not very sensitive to n. Since the majority of estimators are centerd upon the true value of the parameter and they have a dispersion which decreases at the same rate as $n^{-1/2}$, the transformation $X_n = \sqrt{n}(\widehat{\theta}_n - \theta)$ is frequently used.

(b) To look for an approximation $\widehat{H}_n(x)$ to the distribution function $H_n(x) = \Pr(X_n \le x)$ such that, when n tends to infinity, the error

$$\left|\widehat{H}_n(x) - H_n(x)\right|$$

tends to zero.

(c) The distribution function of $\widehat{\theta}_n$ is approximated by $\widehat{H}_n$, i.e., $\Pr(\widehat{\theta}_n \le a) = \Pr\{X_n \le b_n(a;\theta,n)\}$ is approximated by $\widehat{H}_n\{b_n(a;\theta,n)\}$.

Let $\widehat{H}_n(x)$ be an approximation to the distribution function $H_n(x)$. If, for every x,

$$\lim_{n\to\infty} n^{(r/2)-1}\left|\widehat{H}_n(x) - H_n(x)\right| = 0, \quad r = 2, 3, \ldots,$$

we write

$$H_n(x) = \widehat{H}_n(x) + o\left\{n^{(r/2)-1}\right\}, \quad r = 2, 3, \ldots,$$

and we say that $\widehat{H}_n(x)$ is an approximation $o\left\{n^{(r/2)-1}\right\}$ or an approximation of order $r-1$. These names are used as well when approximating density functions. The asymptotic distribution is an approximation $o(n^0) = o(1)$ or a first order approximation. These concepts are related with the topic *Asymptotic, higher order* which is presented as well in this Encyclopedia.

The number n measures the velocity at which the error of approximation tends to zero as n tends to infinity. If we choose the transformation b such that H_n and $\widehat{H}_n$ vary gently with n, the value of r can give an indication of the error of approximation for moderate values of n.

There are two well known methods for obtaining high order approximations to distributions, both based on the Fourier inversion of the characteristic function. Let $\phi_n(z,\theta) = E\{\exp(izX_n)\}$ be the characteristic function of X_n and let $\psi_n(z,\theta) = \log\phi_n(z,\theta)$ be the cumulant generating function. If ϕ_n is integrable, the density function h_n of X_n can be written as

$$h_n(x;\theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixz}\phi_n(z,\theta)dz$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\{-ixz + \psi_n(z,\theta)\}\,dz. \qquad (1)$$

Frequently it is possible to expand $\psi_n(z,\theta)$ in power series where the successive terms are increasing powers of $n^{-1/2}$. In this case the integrand can be approximated by the first few terms of this series expansion. Integrating term by term, one obtains a series approximation to h_n; afterward integration will give an approximation to the distribution function. The approximation known as *Edgeworth approximation* or ►*Edgeworth expansion* consists in expanding $\psi_n(z,\theta)$ at $z = 0$. This method is the most frequently used in practice because of its relative simplicity. It does not require a complete knowledge of $\psi_n(z,\theta)$. It is enough if one knows the first low order cumulants of X_n. More details about this method is given in this Encyclopedia under the name *Edgeworth expansion*. The approximation known as *saddlepoint approximation* is obtained by expanding $\psi_n(z,\theta)$ at the "saddlepoint" value z^* where the integrand of (1) is maximized. This method, introduced by Daniels (1954), is more complex and requires a deeper knowledge of the function $\psi_n(z,\theta)$.When this knowledge is available, the method gives accurate approximations specially in the "tail" region of the distribution. Daniels (1956) and Phillips (1978) applied this method to some autocorrelation statistics in time series analysis. More details about this method is given in this Encyclopedia under the name *Saddlepoint approximations*.

Wallace (1958) gives an excellent introduction to the approximations based on expansions of the characteristic function. An exposition with emphasis on multivariate expansions can be found in Barndorff-Nielsen and Cox (1979). Durbin (1980) proposed a simple method for obtaining a second order approximation to the density of a large class of statistics. This method is discuss in this Encyclopedia under the name *Approximations for densities of sufficient estimators*.

Attitudes and Perspectives

The theory of approximate distributions, like the theory of exact distributions, depends on the assumptions made about the stochastic process which generates the data. The quality of the approximations will not be better than the quality of the specifications sustaining them. one certainly will not rely upon a theory of distribution unless the conclusions are so robust that they do not vary significantly in front of moderate changes of basic assumptions. Since the majority of the methods of approximation use information about the first four moments at least, while the usual asymptotic theory only need information about the first two moments, some loss of robustness has to be expected. However, if some idea about the degree of skewness and kurtosis is available, this information can be helpful to obtain better approximations to the distribution of statistics.

Recently there has been an increasing interest in asymptotic theory. Great efforts have been made in order to demonstrate that some statistics are asymptotically normal and efficient. Of course, the asymptotic theory is important to have an idea of the sample properties of a given statistical procedure. Unfortunately there has been some confusion with the use of the terms "asymptotic" and "approximated."The fact that a standardized estimator has an asymptotic normal distribution is purely a mathematical proposition about the limit of the probabilities measures under a set of previously specified assumptions. The fact that a given estimator is approximately normal suggests that, for this particular problem, one believes in the possibility of treating the estimator as if it was normal.

Sometimes, under certain circumstances, asymptotic arguments lead to good approximations, bet frequently they do not. A careful analyst, with some knowledge of statistical theory, a modest computer and a great amount of common sense can find reasonable approximations for a given inferential problem.

About the Author

Professor Abril is co-editor of the *Revista de la Sociedad Argentina de Estadstica* (Journal of the Argentinean Statistical Society).

Cross References

►Approximations for Densities of Sufficient Estimators
►Asymptotic Normality
►Asymptotic, Higher Order
►Central Limit Theorems
►Cornish-Fisher Expansions
►Edgeworth Expansion
►Limit Theorems of Probability Theory
►Saddlepoint Approximations
►Strong Approximations in Probability and Statistics

References and Further Reading

Abril JC (1985) Asymptotic expansions for time series problems with applications to moving average models. PhD thesis, The London School of Economics and Political Science, University of London, England
Anderson TW (1971) The statistical analysis of time series. Wiley, New York
Barndorff-Nielsen O, Cox DR (1979) Edgeworth and saddle-point approximations with statistical applications. J R Stat Soc B 41:279–312
Bickel PJ, Doksum KA (1977) Mathematical statistics. Holden-Day, San Francisco
Brockwell PJ, Davis RA (1991) Time series: theory and methods, 2nd edn. Springer, New York
Daniels HE (1954) Saddlepoint approximations in statistics. Ann Math Stat 25:631–650
Daniels HE (1956) The approximate distribution of serial correlation coefficients, Biometrika 43:169–185
Durbin J, (1980) Approximations for the densities of sufficient estimates. Biometrika 67:311–333
Durbin J, Watson GS (1971) Testing for serial correlation in least squares regression, III. Biometrika 58:1–19
Feller W (1971) An Introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York
Hannan EJ (1970) Multiple time series. Wiley, New York
Henshaw RC (1966) Testing single-equation least-squares regression models for autocorrelated disturbances. Econometrica 34:646–660
Jenkins GM (1954) An angular transformation for the serial correlation coefficient. Biometrika 41:261–265
Johnson NI, Kotz S (1970) Continuous univariate distributions, vol 1. Wiley, New York
Phillips PCB (1978) Edgeworth and saddlepoint approximations in a first order autoregression. Biometrika 65:91–98
Priestley MB (1982) Spectral analysis and time series. Academic, London
Quenouille MH (1948) Some results in the testing of serial correlation coefficients. Biometrika 35:261–284
Theil H, Nagar AL (1961) Testing the independence of regression disturbances. J Am Stat Assoc 56:793–806
Wallace DL (1958) Asymptotic approximations to distributions. Ann Math Stat 29:635–654

Association Measures for Nominal Categorical Variables

Tarald O. Kvålseth
Professor Emeritus
University of Minnesota, Minneapolis, MN, USA

As a means of summarizing the potential relationship between two (or more) random categorical variables X and Y, a number of measures of association have been proposed over the years. A historical review of such measures and new proposals have been presented in a series of papers by Goodman and Kruskal (1979) [see also Kendall and Stuart (1979), Ch. 33 and Liebetrau (1983)]. Such summary measures depend on whether X and Y are nominal or ordinal as well as on whether X and Y are to be treated symmetrically or asymmetrically. In the symmetric case, X and Y are treated equivalently and no causal relationship is assumed to exist between them. In the asymmetric case, a causal relationship between X and Y is considered to exist so that one variable is treated as the explanatory variable (X) and the other variable treated as the response variable (Y).

The focus here will be on the case when both X and Y are nominal categorical variables, i.e., no natural ordering exists for the variables. Association measures for both the symmetric and asymmetric case will be considered.

Symmetric Measures

For the variable X with I categories and the variable Y with J categories, their joint and marginal probabilities are defined as $\Pr(X = i, Y = j) = p_{ij}$, $\Pr(X = i) = p_{i+}$, and $\Pr(Y = j) = p_{+j}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$ where $\sum_1^2 i = 1I \sum_1^2 j = 1J p_{ij} = \sum_1^2 i = 1I p_{i+} = \sum_1^2 j = 1J p_{+j} = 1$. In terms of a two-way contingency table with I rows and J columns, the cell entries come from the joint distribution $\{p_{ij}\}$, with p_{ij} being the entry in cell (i, j), and $\{p_{i+}\}$ and $\{p_{+j}\}$ are the marginal distributions (totals) for the rows and columns, respectively. The conditional distribution of Y given X is defined in terms of $p_{j|i} = p_{ij}/p_{i+}$ for all i and j.

Several early suggested association measures were based on the (Pearson) coefficient of mean square contingency defined by

$$\Phi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{p_{ij}^2}{p_{i+}p_{+j}} - 1. \quad (1)$$

If the p_{ij} represent sample estimates (of population probabilities π_{ij}) $p_{ij} = n_{ij}/N$ based on the multinominal frequencies n_{ij} for all i, j and sample size $N = \sum_{1}^{2} i = 1I \sum_{1}^{2} j = 1J n_{ij}$, then it is recognized that $\Phi^2 = X^2/N$ where X^2 is the familiar Pearson chi-square goodness-of-fit statistic for testing the null hypothesis or independence between X and Y, i.e.,

$$X^2 = N\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{n_{i+}n_{+j}} - 1\right). \quad (2)$$

The most popular such association measure based on X^2 appears to be Cramér's (1946) V defined as

$$V = \sqrt{\frac{X^2}{N(M-1)}}, \quad M = \min\{I, J\}. \quad (3)$$

This V ranges in value between 0 and 1, inclusive, for any given I and J, with $V = 0$ if, and only if, X and Y are independent and $V = 1$ when there is no more than one non-zero entry in either each row or in each column. The V is invariant with any permutations of the rows or the columns. The estimated standard error of V is given in Bishop et al. (1975, p. 386), but its expression is rather messy.

Kendall and Stuart (1979, p. 606), have shown that V^2 is the mean squared canonical correlation. However, it has been argued that values of V are difficult to interpret since V has no obvious probabilistic meaning or interpretation. Nevertheless, V does reflect the divergence (or "distance") of the distribution $\{p_{ij}\}$ from the independence distribution $\{p_{i+}p_{+j}\}$ relative to the maximum divergence.

There is some uncertainty in the literature as to whether V or V^2 is the proper measure to use. This issue will be addressed in a section below. It may also be pointed out that a similar association measure can be formulated in terms of the likelihood-ratio statistic G^2, which has the same asymptotic chi-square distribution as χ^2 under the null hypothesis and is often used instead of χ^2. For the G^2 under independence, i.e., for

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{Nn_{ij}}{n_{i+}n_{+j}}\right) \quad (4)$$

and since $n_{ij} \le n_{i+}$ and $n_{ij} \le n_{+j}$ for all i and j, it follows that

$$\begin{aligned} G^2 \le G_X^2 &= 2\sum_{i=1}^{I} n_{i+}\log\left(\frac{N}{n_{i+}}\right) \text{ and } G^2 \le G_Y^2 \\ &= 2\sum_{j=1}^{J} n_{+j}\log\left(\frac{N}{n_{+j}}\right). \end{aligned} \quad (5)$$

Thus, analogously to V in (3), one could define the association measure

$$W = \sqrt{\frac{G^2}{\min\{G_X^2, G_Y^2\}}} \quad (6)$$

with G_X^2 and G_Y^2 as given in (5).

This new measure W can also be interpreted as the divergence ("distance") of the distribution $\{p_{ij}\}$ from the independence distribution $\{p_{i+}p_{+j}\}$ relative to its maximum [see also Kvålseth (1987)]. The W has the same type of properties as Cramér's V in (3) and can be expected to take on values quite similar to those of V. For instance, for the data

$$n_{11} = 20 \quad n_{12} = 15 \quad n_{13} = 25$$

$$n_{21} = 5 \quad n_{22} = 25 \quad n_{23} = 10$$

it is found from (3) and (6) that $V = .38$ and $W = .33$.

Asymmetric Measures

Goodman and Kruskal 1979 have discussed two different asymmetric association measures $(\lambda_{Y|X})$ and $(\tau_{Y|X})$ for the case when X can be considered to be the explanatory variable and Y the response variable. Such measures are frequently referred to as proportional reduction in error (PRE) measures since they can be interpreted in terms of the relative difference between two error probabilities P_Y and $P_{Y|X}$, i.e.,

$$PRE_{Y|X} = \frac{P_Y - P_{Y|X}}{P_Y} \quad (7)$$

where P_Y is the probability of error when predicting the Y – category of a randomly selected observation or item without knowing its X – category and $P_{Y|X}$ is the corresponding expected (weighted mean) error probability given its X – category.

The optimal prediction strategy would clearly be to predict that a randomly selected observation (item) would belong to a maximum-probability (modal) category, so that with

$$p_{+m} = \max\{p_{+1}, \ldots, p_{+J}\}; \text{ and } p_{im} = \max\{p_{i1}, \ldots, p_{iJ}\}, \quad i = 1, \ldots, I$$

the error probabilities P_Y and $P_{Y|X}$ become

$$P_Y = 1-p_{+m},\ P_{Y|X} = \sum_{i=1}^{I} p_{i+}\,(1-p_{im}/p_{i+}) = 1-\sum_{i=1}^{I} p_{im}. \quad (8)$$

From (7)–(8), the so-called Goodman–Kruskal *lambda* becomes

$$\lambda_{Y|X} = \frac{\sum_{1}^{2} i = 1 I p_{im} - p_{+m}}{1-p_{+m}} \quad (9)$$

which is the relative decrease in the error probability when predicting the Y-category as between not knowing and knowing the X-category.

Another asymmetric measure is based on a different prediction rule: Predictions are made according to the given probabilities. Thus, a randomly chosen observation (item) is predicted to fall in the jth category of Y with probability p_{+j} $(j = 1,\ldots,J)$ if its X-category is unknown. If, however, the observation is known to belong to the ith category of X, it is predicted to belong to the jth category of Y with the (conditional) probability p_{ij}/p_{i+} $(j = 1,\ldots,J)$. The error probabilities are then given by

$$P_Y = 1-\sum_{j=1}^{J} p_{+j}^2, \quad P_{Y|X} = \sum_{i=1}^{I} p_{i+}\left[1-\sum_{j=1}^{J}(p_{ij}/p_{i+})^2\right] \quad (10)$$

so that, from (7) and (10), the following so-called Goodman–Kruskal *tau* results:

$$\tau_{Y|X} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}^2/p_{i+} - \sum_{j=1}^{J} p_{+j}^2}{1-\sum_{j=1}^{J} p_{+j}^2} \quad (11)$$

which gives the relative reduction in the error probability when predicting an observation's Y-category as between its X-category not given and given.

Both measures in (9) and (11), and whose estimated standard errors are given elsewhere [e.g., Bishop et al. (1975, pp. 388–391), Goodman and Kruskal (1979), and Liebetrau (1983)], can assume values between 0 and 1, inclusive. Both equal 1 if, and only if, each row of the contingency table contains no more than one non-zero cell entry. Both are invariant under permutations of rows or of columns. However, their zero-value conditions differ. The $\tau_{Y|X} = 0$ if, and only if, X and Y are independent, whereas $\lambda_{Y|X} = 0$ if (1) X and Y are independent or (2) the modal probabilities p_{im} in all rows fall in the same column. This second condition is most likely to occur when the marginal distribution $\{p_{i+}\}$ is highly uneven (non-uniform). Thus, in cases of highly uneven $\{p_{i+}\}$, $\lambda_{Y|X}$ may be 0 or very small, while other measures such as $\tau_{Y|X}$ may be substantially larger. The high sensitivity of $\lambda_{Y|X}$ to $\{p_{i+}\}$ is one limitation of this measure that may lead to misleadingly low association values.

Symmetric version of *lambda* and *tau* can also be formulated in terms of weighted averages (Goodman and Kruskal 1979). Thus, in terms of the general expression in (7), a symmetric *PRE* could be formulated as the following weighted mean of $PRE_{Y|X}$ and $PRE_{X|Y}$:

$$PRE = \frac{P_Y - P_{Y|X} + P_X - P_{X|Y}}{P_Y + P_X}.$$

However, there would seem to be no strong reason for preferring such symmetricized measures over the V or W in (3) and (6).

It should be pointed out that asymmetric association measures can also be formulated in terms of relative reduction in variation, somewhat analogously to the coefficient of determination (R^2) used in regression analysis. This can be done by basically replacing the prediction error probabilities in (7) with appropriate measures of categorical variation (Agresti 2002, pp. 56–69).

Concluding Comment

For Cramér's V in (3), there is inconsistency in the literature concerning the use of V versus V^2 (and Cramér himself proposed V^2 (Cramér 1946, p. 443)). Also, concern has been expressed that different measures such as those in (9) and (11) can produce quite different results. Such issues are indeed important and are often overlooked.

As with any summary measure, and so it is with association measures, it is essential that a measure takes on values throughout its range that are reasonable in that they provide true or valid representations of the attribute being measured. In order to make such an assessment for the above association measures, consider the simple case of a 2×2 table with all the marginal probabilities equal to .5 and with the following cell entries:

$$p_{11} = (1-w)/4 p_{12} = (1+w)/4$$

$$p_{21} = (1+w)/4 p_{22} = (1-w)/4$$

with $0 \le w \le 1$. Each of these probabilities are seen to be the weighted mean of the corresponding probabilities for the case of perfect association and zero association (independence) for the given marginal probabilities, i.e.,

$$p_{11} = p_{22} = w(0)+(1-w)(.25), \quad p_{12} = p_{21} = w(.5)+(1-w)(.25)$$

In order for some association measure $A \in [0,1]$ to take on reasonable values in this ease, the only logical requirement is clearly that

$$A = w(A=1) + (1-w)(A=0) = w,\ w \in [0,1]$$

It is readily seen that the measures in (3) and (9) meet this requirement for all w, i.e., $V = w$ (and not V^2) and $\lambda_{Y|X} = w$ for the above $\{p_{ij}\}$ – distribution. However, it is seen that, for (11), $\lambda_{Y|X} = w^2$. This shows that $\tau'_{Y|X} = \sqrt{\tau_{Y|X}}$ should be used as an association measure rather than $\tau_{Y|X}$. In the case of W in (6), it is apparent that W is only approximately equal to w, but the approximation appears to be sufficiently close for W to be a competitive association measure.

About the Author

For biography *see* the entry ►Entropy.

Cross References

►Categorical Data Analysis
►Scales of Measurement
►Variables
►Variation for Categorical Variables

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken, NJ

Bishop YMM, Fienberg SE, Holland PW(1975) Discrete multivariate analysis. MIT, Cambridge, MA

Cramér H (1946) Mathematical methods for statistics. Princeton University Press, Princeton, NJ

Goodman LA, Kruskal WH (1979) Measures of association for cross-classifications. Springer, New York

Kendall M, Stuart A (1979) The advanced theory of statistics, vol.2, 4th edn. Charles Griffin, London

Kvålseth TO (1987) Entropy and correlation: some comments. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-17, 517–519

Liebetrau AM (1983) Measures of Association. Beverly Hills, CA: Sage Publications.

Astrostatistics

Eric D. Feigelson
Professor, Associate Director
Penn State Center for Astrostatistics Pennsylvania State University, University Park, PA, USA

Introduction

The term "astronomy" is best understood as short-hand for "astronomy and astrophysics." Astronomy is the observational study of matter beyond Earth: planets and other bodies in the Solar System, stars in the Milky Way Galaxy, galaxies in the Universe, and diffuse matter between these concentrations of mass. The perspective is rooted in our viewpoint on or near Earth, typically using telescopes or robotic satellites. Astrophysics is the study of the intrinsic nature of astronomical bodies and the processes by which they interact and evolve. This is an inferential intellectual effort based on the well-confirmed assumption that physical processes established to rule terrestrial phenomena – gravity, thermodynamics, electromagnetism, quantum mechanics, plasma physics, chemistry, and so forth – also apply to distant cosmic phenomena.

Statistical techniques play an important role in analyzing astronomical data and at the interface between astronomy and astrophysics. Astronomy encounters a huge range of statistical problems: samples selected with truncation; variables subject to censoring and heteroscedastic measurement errors; parameter estimation of complex models derived from astrophysical theory; anisotropic spatial clustering of galaxies; time series of periodic, stochastic, and explosive phenomena; image processing of both gray-scale and Poissonian images; ►data mining of terabyte-petabyte datasets; and much more. Thus, astrostatistics is not focused on a narrow suite of methods, but rather brings the insights from many fields of statistics to bear on problems arising in astronomical research.

History

As the oldest observational science, astronomy was the driver for statistical innovations over many centuries (Stigler 1986; Hald 1998). Hipparchus, Ptolemy, al-Biruni, and Galileo Galilei were among those who discussed methods for averaging discrepant astronomical measurements. The least squares method (see ►Least Squares) and its understanding in the context of the normal error distribution were developed to address problems in Newtonian celestial mechanics during the early nineteenth century by Pierre-Simon Laplace, Adrian Legendre, and Carl Friedrich Gauss. The links between astronomy and statistics considerably weakened during the first decades of the twentieth century as statistics turned its attention to social and biological sciences while astronomy focused on astrophysics. Maximum likelihood methods emerged slowly starting in the 1970s, and Bayesian methods are now gaining considerably popularity.

Modern astrostatistics has grown rapidly since the 1990s. Several cross-disciplinary research groups emerged to develop advanced methods and critique common practices (http://hea-www.harvard.edu/AstroStat; http://www.incagroup.org; http://astrostatistics.psu.edu). Monographs were written on astrostatistics (Babu and Feigelson 1996), galaxy clustering (Martinez and Saar 2002), image processing (Starck and Murtagh 2006), Bayesian analysis

(Gregory 2005), and Bayesian cosmology (Hobson et al. 2010). The *Statistical Challenges in Modern Astronomy* (Babu and Feigelson 2007) conferences bring astronomers and statisticians together to discuss methodological issues.

The astronomical community is devoting considerable resources to the construction and promulgation of large archival datasets, often based on well-designed surveys of large areas of the sky. These surveys can generate petabytes of images, spectra and time series. Reduced data products include tabular data with approximately ten variables measured for billions of astronomical objects. Major projects include the Sloan Digital Sky Survey, International Virtual Observatory, and planned Large Synoptic Survey Telescope (http://www.sdss.org, http://www.ivoa.net, http://www.lsst.org). Too large for traditional treatments, these datasets are spawning increased interest in computationally efficient data visualization, data mining, and statistical analysis. A nascent field of astroinformatics allied to astrostatistics is emerging.

Topics in Contemporary Astrostatistics

Given the vast range of astrostatistics, only a small portion of relevant issues can be outlined here. We outline three topics of contemporary interest (The astronomical research literature can be accessed online through the SAO/NASA Astrophysics Data System, http://adsabs.harvard.edu.).

Heteroscedastic Measurement Errors

Astronomical measurements at telescopes are made with carefully designed and calibrated instruments, and "background" levels in dark areas of the sky are examined to quantitatively determine the noise levels. Thus, unlike in social and biological science studies, heteroscedastic measurement error are directly obtained for each astronomical measurement. This produces unusual data structures. For example, a multivariate table of brightness of quasars in six photometric bands will have 12 columns of numbers giving the measured brightness and the associated measurement error in each band.

Unfortunately, few statistical techniques are available for this class of non-identically distributed data. Most errors-in-variables methods are designed to treat situations where the heteroscedasticity is not measured, and instead becomes part of the statistical model (Carroll et al. 2006). Methods are needed for density estimation, regression, multivariate analysis and classification, spatial processes, and time series analysis. Common estimation procedures in the astronomical literature weight each measurement by its associated error. For instance, in a functional regression model, the parameters $\hat{\boldsymbol{\theta}}$ in model M are estimated by minimizing the weighted sum of squared residuals $\sum_i (O_i - M_i(\hat{\boldsymbol{\theta}})^2/\sigma_i^2$ of the observed data O_i where σ_i^2 are the known variances of the measurement errors.

More sophisticated methods are being developed, but have not yet entered into common usage. Kelly (2007) treats structural regression as an extension of a normal mixture model, constructing a likelihood which can either be maximized with the EM Algorithm or used in ►Bayes' theorem. The Bayesian approach is more powerful, as it also can simultaneous incorporate censoring and truncation into the measurement error model. Delaigle and Meister (2008) describe a nonparametric kernel density estimator that takes into account the heteroscedastic errors. More methods (e.g., for multivariate clustering and time series modeling) are needed.

Censoring and Truncation

In the telescopic measurement of quasar brightnesses outlined above, some targeted quasars may be too faint to be seen above the background noise level in some photometric bands. These nondetections lead to censored data points. The situation is similar in some ways to censoring treated by standard survival analysis, but differs in other ways: the data are left-censored rather than right-censored; censoring can occur in any variable, not just a single response variable; and censoring levels are linked to measurement error levels. Survival techniques have come into common usage in astronomy since their introduction (Isobe et al. 1986). They treat some problems such as density estimation (with the Kaplan-Meier product-limit estimator), two-sample tests (with the Gehan, logrank and Peto-Prentice tests), correlation (using a generalization of Kendall's τ), and linear regression (using the Buckley-James line).

Consider a survey of quasars at a telescope with limited sensitivity where the quasar sample is not provided in advance, but is derived from the photometric colors of objects in the survey. Now quasars which are too faint for detection are missing entirely from the dataset. Recovery from this form of truncation is more difficult than recovery from censoring with a previously established sample. A major advance was the derivation of the nonparametric estimator for a randomly truncated dataset, analogous to the Kaplan-Meier estimator for censored data, by astrophysicist Lynden-Bell (1971). This solution was later recovered by statistician Woodroofe (1985), and bivariate extensions were developed by Efron and Petrosian (1992).

Periodicity Detection in Difficult Data

Stars exhibit a variety of periodic behaviors: binary star or planetary orbits; stellar rotation; and stellar oscillations. While Fourier analysis is often used to find and characterize such periodicities, the data often present problems such as non-sinusoidal repeating patterns, observations of limited duration, and unevenly-spaced observations. Non-sinusoidal periodicities occur in elliptical orbits, eclipses, and rotational modulation of surface features. Unevenly-spaced data arise from bad weather at the telescope, diurnal cycles for ground-based telescopes, Earth orbit cycles for satellite observatories, and inadequate observing time provided by telescope allocation committees.

Astronomers have developed a number of statistics to locate periodicities under these conditions. The Lomb-Scargle periodogram (Scargle 1982) generalizes the Schuster periodogram to treat unevenly-spaced data. Stellingwerf (1978) presents a widely used least-squared technique where the data are folded modulo trial periods, grouped into phase bins, and intra-bin variance is compared to inter-bin variance using χ^2. The method treats unevenly spaced data, measurement errors, and non-sinusoidal shapes. Dworetsky (1983) gives a similar method without binning suitable for sparse datasets. Gregory and Loredo (1992) develop a Bayesian approach for locating non-sinusoidal periodic behaviors from Poisson distributed event data. Research is now concentrating on methods for computationally efficient discovery of planets orbiting stars as they eclipse a small fraction during repeated transits across the stellar surface. These methods involve matched filters, Bayesian estimation, least-squares box-fitting, maximum likelihood, ►analysis of variance, and other approaches (e.g., Pontopappas et al. 2005).

About the Author

Dr. Eric Feigelson was trained in astronomy at Haverford College and Harvard University during the 1970s, and entered the Astronomy and Astrophysics faculty at Pennsylvania State University in 1983 where he received an NSF Presidential Young Investigator Award and is now Professor. In addition to X-ray astronomical studies of star and planet formation, he has a long-standing collaboration with statisticians. Working with G. Jogesh Babu at Penn State's Center for Astrostatistics, he organizes summer schools, conferences and other resources for advancing statistical methodology in astronomy. He serves as an Associate Editor of the *Astrophysical Journal*, on the Virtual Astronomical Observatory Science Council, and other organizations relating statistics to astronomy.

Cross References

►Chaotic Modelling
►False Discovery Rate
►Heteroscedasticity
►Linear Regression Models
►Statistics, History of

References and Further Reading

Babu GJ, Feigelson ED (1996) Astrostatistics. Chapman and Hall, London

Babu GJ, Feigelson ED (2007) Statistical challenges in modern astronomy IV. Astronomical Society of the Pacific, San Fransisco, California

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement errors in nonlinear models. Chapman and Hall/CRC, Boca Raton, FL

Delaigle A, Meister A (2008) Density estimation with heteroscedastic error. Bernoulli 14:562–579

Dworetsky MM (1983) A period-finding method for sparse randomly spaced observations of 'How long is a piece of string?'. Mon Not Royal Astro Soc 203:917–924

Efron B, Petrosian V (1992) A simple test of independence for truncated data with applications to redshift surveys. Astrophys J 399:345–352

Gregory PC (2005) Bayesian logical data analysis for the physical sciences. Cambridge University Press, Cambridge, UK

Gregory PC, Loredo TJ (1992) A new method for the detection of a periodic signal of unknown shape and period. Astrophys J 398:146–168

Hald A (1998) A history of mathematical statistics from 1750 to 1930. Wiley, New York

Hobson MP et al (eds) (2010) Bayesian methods in cosmology. Cambridge University Press, Cambridge

Isobe T, Feigelson ED, Nelson PI (1986) Statistical methods for astronomical data with upper limits. II–correlation and regression. Astrophys J 306:490–507

Kelly BC (2007) Some Aspects of Measurement Error in Linear Regression of Astronomical Data. Astrophys J 665:1489–1506

Lynden-Bell D (1971) A method of allowing for known observational selection in small samples applied to 3CR quasars. Mon Not R Astro Soc 155:95–118

Martinez VJ, Saar E (2002) Statistics of the galaxy distribution. CRC, Boca Raton, USA

Protopapas P, Jimenez R, Alcock C (2005) Fast identification of transits from light-curves. Mon Not R Astro Soc 362: 460–468

Scargle JD (1982) Studies in astronomical time series analysis. II Statistical aspects of spectral analysis of unevenly spaced data. Astrophys J 263:835–853

Starck J-L, Murtagh F (2006) Astronomical image and data analysis. Springer, New York

Stellingwerf RF (1978) Period determination using phase dispersion minimization. Astrophys J 224:953–960

Stigler SM (1986) The history of Statistics: the measurement of uncertainty before 1900. Harvard University Press, Cambridge, MA

Woodroofe MB (1985) Estimating a distribution function with truncated data. Ann Statist 13:163–177

A

Asymptotic Normality

John E. Kolassa
Professor
Rutgers University, Newark, NJ, USA

Consider a sequence of random variables T_n, whose distribution depends on a parameter n that generally represents sample size. The sequence is said to be asymptotically normal if there exists a sequences μ_n and σ_n such that $\lim_{n\to\infty} P\left[(T_n - \mu_n)/\sigma_n \le x\right] = \Phi(x)$ for all x, where $\Phi(x)$ is the standard Gaussian distribution function

$$\int_{-\infty}^{x} \exp(-y^2/2)(2\pi)^{-1/2}\, dy. \tag{1}$$

One often writes

$$T_n \sim AN(\mu_n, \sigma_n^2) \tag{2}$$

to express asymptotic normality. Note that μ_n generally depend on n, and furthermore may be data-dependent. Furthermore, in some cases T_n might be a sequence of random vectors; in this case, μ_n is a sequence of vectors, σ_n^2 is a sequence of matrices, and Φ the vector valued counterpart of (1). In the scalar case, for fixed n, the quantity σ_n is called the standard error of T_n.

Many frequentist statistical inferential procedures are performed by constructing a T_n so that (2) holds under a null hypothesis, with a dissimilar distribution under interesting alternative hypotheses, and reporting

$$2(1 - \Phi(|(T_n - \mu_n)/\sigma_n|)) \tag{3}$$

as a two-sided p-value; the application for one-sided p-values is similar, and there are also Bayesian applications of a similar flavor. Serfling (1980) provides further information.

Consider the following examples of quantities that are asymptotically normal:

- If T_n is the mean of n independent and identically distributed random variables, each with expectation μ and standard deviation σ, then

  $$T_n \sim AN(\mu, \sigma^2/n). \tag{4}$$

  Furthermore, if s_n is the traditional standard deviation of the contributors the the mean,

  $$T_n \sim AN(\mu, s_n^2/n); \tag{5}$$

  note that the standard error here is data-dependent, and it is incorrect to call $s_n/\sqrt{n}$ a standard deviation of T_n, even approximately. In the present case square root of the second argument to the AN operator estimates the standard deviation of T_n, but a further example shows that even this need not be true. In this case, the standard Z-test for a single sample mean follows from using (4) when σ is known, and when the components of T_n are binary, the standard standard Z-test for a single sample mean follows from using (4) with σ^2 the standard Bernoulli variance. When σ is unknown, (5) is often used instead, and for $n \le 30$, the t distribution function is generally used in place of Φ in (3) for greater accuracy.
- Many rank-based statistics are asymptotically normal; for example, if T_n is the Wilcoxon signed-rank statistic (see ▶Wilcoxon-Signed-Rank Test) for testing whether the expectation of n independent and identically distributed random variables takes on a particular null value, assuming symmetry and continuity of the underlying distribution. Without loss of generality, take this null mean to be zero. Then T_n is obtained by ranking the absolute values of the observations, and summing the ranks of those observations that are positive. Hettmansperger (1984) notes that (2) holds with $\mu_n = n(n+1)/2$ and $\sigma_n = \sqrt{n(n+1)(2n+1)/24}$, and the test against the two-sided alternative reports the p-value (3). In this case, T_n may be written as the sum of independent but not identically-distributed random variables, or as the sum of identically-distributed but not independent random variables.
- Many parameter estimates resulting from fitting models with independent observations are asymptotically normal. For example, consider independent Bernoulli observations Y_i with $P\left[Y_i = 1\right] = \exp(\beta_1 + \beta_2 x_i)/(1 + \exp(\beta_1 + \beta_2 x_i))$. Let

  $$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[Y_i\beta_1 + x_i Y_i \beta_2 - \log(1 + \exp(\beta_1 + \beta_2 x_i))\right], \tag{6}$$

  and let $\hat{\boldsymbol{\beta}}$ maximize ℓ; here $\hat{\boldsymbol{\beta}}$ implicitly depends on n. Then

  $$\hat{\boldsymbol{\beta}} \sim AN(\boldsymbol{\beta}, [-\ell''(\boldsymbol{\beta})]^{-1}), \tag{7}$$

  as one can see by heuristically expressing $\ell'(\boldsymbol{\beta}) + \ell''(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx \ell'(\hat{\boldsymbol{\beta}}) = 0$, and solving for $\hat{\boldsymbol{\beta}}$ to obtain $\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta} - [\ell''(\boldsymbol{\beta})]^{-1}\ell'(\boldsymbol{\beta})$, noting that $\ell''(\boldsymbol{\beta})$ is nonrandom, and noting that a variant of the Central Limit Theorem proves the asymptotic normality of $\ell'(\boldsymbol{\beta})$, and hence of $\hat{\boldsymbol{\beta}}$. This heuristic argument is easily made rigorous once one notes $\hat{\boldsymbol{\beta}}$ is consistent (i.e., for any $\epsilon > 0$, $\lim_{n\to\infty} P\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| > \epsilon\right] = 0$; see Cox and Hinkley 1974). In this example, the outcome $Y_i = 1\ \forall i$

has positive probability, and for such $\{Y_1, \ldots, Y_n\}$, $\hat{\beta}_1$ is infinite. A similar result holds for $Y_i = 0\ \forall i$. Hence the variance of $\hat{\boldsymbol{\beta}}$ does not exist.

About the Author

John Kolassa is Professor of Statistics and Biostatistics, and Director of the Graduate Programs in Statistics and Biostatistics at Rutgers University. John Kolassa was previously Assistant and Associate Professor in the Department of Biostatistics, and Graduate Student Advisor, at the University of Rochester. John Kolassa is an Elected Fellow of the ASA and IMS, and has received grants from and served on grant review panels for the National Science Foundation and the National Institutes of Health.

Cross References

►Approximations to Distributions
►Asymptotic Relative Efficiency in Testing
►Central Limit Theorems
►Chernoff-Savage Theorem
►Limit Theorems of Probability Theory
►Martingale Central Limit Theorem
►Properties of Estimators
►Sampling Problems for Stochastic Processes
►Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts

References and Further Reading

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, New York

Hettmansperger TP (1984) Statistical inference based on ranks. Krieger, Melbourne, FL

Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York

Asymptotic Relative Efficiency in Estimation

Robert Serfling
Professor
University of Texas at Dallas, Richardson, TX, USA

Asymptotic Relative Efficiency of Two Estimators

For statistical estimation problems, it is typical and even desirable that several reasonable estimators can arise for consideration. For example, the mean and median parameters of a symmetric distribution coincide, and so the *sample mean* and the *sample median* become competing estimators of the point of symmetry. *Which is preferred? By what criteria shall we make a choice?*

One natural and time-honored approach is simply to compare the sample sizes at which two competing estimators meet a given standard of performance. This depends upon the chosen measure of performance and upon the particular population distribution F.

To make the discussion of sample mean versus sample median more precise, consider a distribution function F with density function f symmetric about an unknown point θ to be estimated. For $\{X_1, \ldots, X_n\}$ a sample from F, put $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $\mathrm{Med}_n = \mathrm{median}\{X_1, \ldots, X_n\}$. Each of $\overline{X}_n$ and Med_n is a consistent estimator of θ in the sense of convergence in probability to θ as the sample size $n \to \infty$. To choose between these estimators we need to use further information about their performance. In this regard, one key aspect is *efficiency*, which answers: *How spread out about θ is the sampling distribution of the estimator?* The smaller the variance in its sampling distribution, the more "efficient" is that estimator.

Here we consider "large-sample" sampling distributions. For $\overline{X}_n$, the classical central limit theorem (see ►Central Limit Theorems) tells us: if F has finite variance σ_F^2, then the sampling distribution of $\overline{X}_n$ is approximately $N\left(\theta, \sigma_F^2/n\right)$, i.e., Normal with mean θ and variance σ_F^2/n. For Med_n, a similar classical result (Serfling 1980) tells us: if the density f is continuous and positive at θ, then the sampling distribution of Med_n is approximately $N(\theta, 1/4[f(\theta)]^2 n)$. On this basis, we consider $\overline{X}_n$ and Med_n to perform equivalently at respective sample sizes n_1 and n_2 if

$$\frac{\sigma_F^2}{n_1} = \frac{1}{4[f(\theta)]^2 n_2}.$$

Keeping in mind that these sampling distributions are only approximations assuming that n_1 and n_2 are "large," we define the *asymptotic relative efficiency (ARE)* of Med to $\overline{X}$ as the *large-sample limit* of the ratio n_1/n_2, i.e.,

$$\mathrm{ARE}\left(\mathrm{Med}, \overline{X}, F\right) = 4[f(\theta)]^2 \sigma_F^2. \tag{1}$$

Definition in the General Case

For any parameter η of a distribution F, and for estimators $\widehat{\eta}^{(1)}$ and $\widehat{\eta}^{(2)}$ approximately $N(\eta, V_1(F)/n)$ and $N(\eta, V_2(F)/n)$, respectively, the *ARE of $\widehat{\eta}^{(2)}$ to $\widehat{\eta}^{(1)}$* is given by

$$\mathrm{ARE}\left(\widehat{\eta}^{(2)}, \widehat{\eta}^{(1)}, F\right) = \frac{V_1(F)}{V_2(F)}. \tag{2}$$

Interpretation. If $\widehat{\eta}^{(2)}$ is used with a sample of size n, the number of observations needed for $\widehat{\eta}^{(1)}$ to perform equivalently is $\mathrm{ARE}(\widehat{\eta}^{(2)}, \widehat{\eta}^{(1)}, F) \times n$.

Extension to the case of multidimensional parameter. For a parameter $\boldsymbol{\eta}$ taking values in $\mathbb{R}^k$, and two estimators $\widehat{\boldsymbol{\eta}}^{(i)}$ which are k-variate Normal with mean $\boldsymbol{\eta}$ and nonsingular covariance matrices $\boldsymbol{\Sigma}_i(F)/n$, $i = 1, 2$, we use [see Serfling (1980)]

$$\mathrm{ARE}\left(\widehat{\boldsymbol{\eta}}^{(2)}, \widehat{\boldsymbol{\eta}}^{(1)}, F\right) = \left(\frac{|\boldsymbol{\Sigma}_1(F)|}{|\boldsymbol{\Sigma}_2(F)|}\right)^{1/k}, \tag{3}$$

the ratio of *generalized variances* (determinants of the covariance matrices), raised to the power $1/k$.

Connection with the Maximum Likelihood Estimator

Let F have density $f(x\,|\,\eta)$ parameterized by $\eta \in \mathbb{R}$ and satisfying some differentiability conditions with respect to η. Suppose also that $I(F) = E_\eta\left\{\left[\frac{\partial}{\partial\eta}\log f(x\,|\,\eta)\right]^2\right\}$ (the *Fisher information*) is positive and finite. Then (Lehmann and Casella 1988) it follows that (a) the *maximum likelihood estimator* $\widehat{\eta}^{(\mathrm{ML})}$ of η is approximately $N(\eta, 1/I(F)n)$, and (b) for a wide class of estimators $\widehat{\eta}$ that are approximately $N(\eta, V(\widehat{\eta}, F)/n)$, a *lower bound* to $V(\widehat{\eta}, F)$ is $1/I(F)$. In this situation, (2) yields

$$\mathrm{ARE}\left(\widehat{\eta}, \widehat{\eta}^{(\mathrm{ML})}, F\right) = \frac{1}{I(F)V(\widehat{\eta}, F)} \le 1, \tag{4}$$

making $\widehat{\eta}^{(\mathrm{ML})}$ (asymptotically) the most efficient among the given class of estimators $\widehat{\eta}$. We note, however, as will be discussed later, that (4) does not necessarily make $\widehat{\eta}^{(\mathrm{ML})}$ the estimator of choice, when certain other considerations are taken into account.

Detailed Discussion of Estimation of Point of Symmetry

Let us now discuss in detail the example treated above, with F a distribution with density f symmetric about an unknown point θ and $\{X_1, \ldots, X_n\}$ a sample from F. For estimation of θ, we will consider not only $\overline{X}_n$ and Med_n but also a third important estimator.

Mean versus Median

Let us now formally compare $\overline{X}_n$ and Med_n and see how the ARE differs with choice of F. Using (1) with $F = N\left(\theta, \sigma_F^2\right)$, it is seen that

$$\mathrm{ARE}\left(\mathrm{Med}, \overline{X}, N\left(\theta, \sigma_F^2\right)\right) = 2/\pi = 0.64.$$

Thus, for sampling from a *Normal* distribution, the sample mean performs as efficiently as the sample median using only 64% as many observations. (Since θ and σ_F are location and scale parameters of F, and since the estimators $\overline{X}_n$ and Med_n are location and scale equivariant, their ARE does not depend upon these parameters.) The superiority of $\overline{X}_n$ here is no surprise since it is the MLE of θ in the model $N\left(\theta, \sigma_F^2\right)$.

As noted above, *asymptotic* relative efficiencies pertain to large sample comparisons and need not reliably indicate small sample performance. In particular, for F *Normal*, the *exact* relative efficiency of Med to $\overline{X}$ for sample size $n = 5$ is a very high 95%, although this decreases quickly, to 80% for $n = 10$, to 70% for $n = 20$, and to 64% in the limit.

For sampling from a *double exponential* (or *Laplace*) distribution with density $f(x) = \lambda e^{-\lambda|x-\theta|}/2$, $-\infty < x < \infty$ (and thus variance $2/\lambda^2$), the above result favoring $\overline{X}_n$ over Med_n is reversed: (1) yields

$$\mathrm{ARE}\left(\mathrm{Med}, \overline{X}, \mathrm{Laplace}\right) = 2,$$

so that the sample mean requires 200% as many observations to perform equivalently to the sample median. Again, this is no surprise because for this model the MLE of θ is Med_n.

A Compromise: The Hodges–Lehmann Location Estimator

We see from the above that the ARE depends dramatically upon the shape of the density f and thus must be used cautiously as a benchmark. For Normal versus Laplace, $\overline{X}_n$ is either greatly superior or greatly inferior to Med_n. This is a rather unsatisfactory situation, since in practice we might not be quite sure whether F is Normal or Laplace or some other type. A very interesting solution to this dilemma is given by an estimator that has excellent *overall performance*, the so-called *Hodges–Lehmann location estimator* (Hodges and Lehmann 1963; see ▶Hodges-Lehmann Estimators):

$$\mathrm{HL}_n = \mathrm{Median}\left\{\frac{X_i + X_j}{2}\right\},$$

the median of all pairwise averages of the sample observations. (Some authors include the cases $i = j$, some not.) We have (Lehmann 1998a) that HL_n is asymptotically $N(\theta, 1/12[\int f^2(x)dx]^2 n)$, which yields that $\mathrm{ARE}\left(\mathrm{HL}, \overline{X}, N\left(\theta, \sigma_F^2\right)\right) = 3/\pi = 0.955$ and $\mathrm{ARE}(\mathrm{HL}, \overline{X}, \mathrm{Laplace}) = 1.5$. Also, for the ▶*Logistic* distribution with density $f(x) = \sigma^{-1}e^{(x-\theta)/\sigma}/[1 + e^{(x-\theta)/\sigma}]^2$, $-\infty < x < \infty$, for which HL_n is the MLE of θ and thus optimal, we have $\mathrm{ARE}(\mathrm{HL}, \overline{X}, \mathrm{Logistic}) = \pi^2/9 = 1.097$ [see Lehmann (1998b)]. Further, for $\mathcal{F}$ the class of all distributions symmetric about θ and having finite variance, we have $\inf_{\mathcal{F}} \mathrm{ARE}(\mathrm{HL}, \overline{X}, F) = 108/125 = 0.864$ [see Lehmann (1998a)]. The estimator HL_n is highly competitive with $\overline{X}$ at Normal distributions, can be infinitely more efficient at some other symmetric distributions F, and is never much

less efficient at any distribution F in $\mathcal{F}$. The computation of HL_n appears at first glance to require $O(n^2)$ steps, but a much more efficient $O(n \log n)$ algorithm is available [see Monohan (1984)].

Efficiency versus Robustness Trade-Off

Although the asymptotically most efficient estimator is given by the MLE, the particular MLE depends upon the shape of F and can be drastically inefficient when the actual F departs even a little bit from the nominal F. For example, if the assumed F is $N(\mu, 1)$ but the actual model differs by a small amount ε of "contamination," i.e., $F = (1-\varepsilon)N(\mu,1) + \varepsilon N(\mu, \sigma^2)$, then

$$\mathrm{ARE}(\mathrm{Med}, \overline{X}, F) = \frac{2}{\pi}\left(1 - \varepsilon + \varepsilon\sigma^{-1}\right)^2\left(1 - \varepsilon + \varepsilon\sigma^2\right),$$

which equals $2/\pi$ in the "ideal" case $\varepsilon = 0$ but otherwise $\to \infty$ as $\sigma \to \infty$. A small perturbation of the assumed model thus can destroy the superiority of the MLE.

One way around this issue is to take a *nonparametric* approach and seek an estimator with ARE satisfying a favorable lower bound. Above we saw how the estimator HL_n meets this need.

Another criterion by which to evaluate and compare estimators is *robustness*. Here let us use finite-sample *breakdown point (BP)*: the minimal fraction of sample points which may be taken to a limit L (e.g., $\pm\infty$) without the estimator also being taken to L. A *robust* estimator remains stable and effective when in fact the sample is only partly from the nominal distribution F and contains some non-F observations which might be relatively extreme contaminants.

A single observation taken to ∞ (with n fixed) takes $\overline{X}_n$ with it, so $\overline{X}_n$ has BP = 0. Its optimality at Normal distributions comes at the price of a complete sacrifice of robustness. In comparison, Med_n has extremely favorable BP = 0.5 but at the price of a considerable loss of efficiency at Normal models.

On the other hand, the estimator HL_n appeals broadly, possessing *both* quite high ARE over a wide class of F and relatively high BP $= 1 - 2^{-1/2} = 0.29$.

As another example, consider the problem of estimation of scale. Two classical scale estimators are the *sample standard deviation* s_n and the *sample MAD* (median absolute deviation about the median) MAD_n. They estimate scale in different ways but can be regarded as competitors in the problem of estimation of σ in the model $F = N(\mu, \sigma^2)$, as follows. With both μ and σ unknown, the estimator s_n is (essentially) the MLE of σ and is asymptotically most efficient. Also, for this F, the population MAD is equal to $\Phi^{-1}(3/4)\sigma$, so that the estimator $\widehat{\sigma}_n = \mathrm{MAD}_n/\Phi^{-1}(3/4) = 1.4826\,\mathrm{MAD}_n$ competes with s_n for estimation of σ. (Here Φ denotes the standard normal distribution function, and, for any F, $F^{-1}(p)$ denotes the pth quantile, $\inf\{x : F(x) \geq p\}$, for $0 < p < 1$.) To compare with respect to robustness, we note that a single observation taken to ∞ (with n fixed) takes s_n with it, s_n has BP = 0. On the other hand, MAD_n and thus $\widehat{\sigma}_n$ have BP = 0.5, like Med_n. However, $\mathrm{ARE}(\widehat{\sigma}_n, s_n, N(\mu, \sigma^2)) = 0.37$, even worse than the ARE of Med_n relative to $\overline{X}$. Clearly desired is a more balanced trade-off between efficiency and robustness than provided by either of s_n and $\widehat{\sigma}_n$. Alternative scale estimators having the same 0.5 BP as $\widehat{\sigma}_n$ but much higher ARE of 0.82 relative to s_n are developed in Rousseeuw and Croux (1993). Also, further competitors offering a range of trade-offs given by (BP, ARE) $= (0.29, 0.86)$ or $(0.13, 0.91)$ or $(0.07, 0.96)$, for example, are developed in Serfling (2002).

In general, efficiency and robustness trade off against each other. Thus ARE should be considered in conjunction with robustness, choosing the balance appropriate to the particular application context. This theme is prominent in the many examples treated in Staudte and Sheather (1990).

A Few Additional Aspects of ARE

Connections with Confidence Intervals

In view of the asymptotic normal distribution underlying the above formulation of ARE in estimation, we may also characterize the ARE given by (2) as the limiting ratio of sample sizes at which the *lengths of associated confidence intervals at approximate level* $100(1-\alpha)\%$,

$$\widehat{\eta}^{(i)} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{V_i(F)}{n_i}},\ i = 1, 2,$$

converge to 0 at the same rate, when holding fixed the coverage probability $1 - \alpha$. (In practice, of course, consistent estimates of $V_i(F)$, $i = 1, 2$, are used in forming the CI.)

Fixed Width Confidence Intervals and ARE

One may alternatively consider confidence intervals of *fixed length*, in which case (under typical conditions) the noncoverage probability depends on n and tends to 0 at an exponential rate, i.e., $n^{-1}\log\alpha_n \to c > 0$, as $n \to \infty$. For fixed width confidence intervals of the form

$$\widehat{\eta}^{(i)} \pm d\,\sigma_F,\ i = 1, 2,$$

we thus define the *fixed width asymptotic relative efficiency (FWARE)* of two estimators as the limiting ratio of sample sizes at which the respective *noncoverage probabilities* $\alpha_n^{(i)}$, $i = 1, 2$, of the associated fixed width confidence intervals

converge to zero at the same exponential rate. In particular, for Med versus $\overline{X}$, and letting $\eta = 0$ and $\sigma_F = 1$ without loss of generality, we obtain (Serfling and Wackerly 1976)

$$\mathrm{FWARE}(\mathrm{Med}, \overline{X}, F) = \frac{\log m(-d)}{\log[2(F(d) - F^2(d))^{1/2}]}, \quad (5)$$

where $m(-d)$ is a certain parameter of the ▶moment generating function of F. The FWARE is derived using *large deviation theory* instead of the central limit theorem. As $d \to 0$, the FWARE converges to the ARE. Indeed, for F a Normal distribution, this convergence (to $2/\pi = 0.64$) is quite rapid: the expression in (5) rounds to 0.60 for $d = 2$, to 0.63 for $d = 1$, and to 0.64 for $d \le 0.1$.

Confidence Ellipsoids and ARE

For an estimator $\widehat{\boldsymbol{\eta}}$ which is asymptotically k-variate Normal with mean $\boldsymbol{\eta}$ and covariance matrix $\boldsymbol{\Sigma}/n$, as the sample size $n \to \infty$, we may form (see Serfling 1980) an *associated ellipsoidal confidence region of approximate level* $100(1-\alpha)\%$ for the parameter $\boldsymbol{\eta}$,

$$E_{n,\alpha} = \{\boldsymbol{\eta} : n\,(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta})' \boldsymbol{\Sigma}^{-1} (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \le c_\alpha\},$$

with $P\left(\chi^2_k > c_\alpha\right) = \alpha$ and in practice using a consistent estimate of $\boldsymbol{\Sigma}$. The *volume* of the region $E_{n,\alpha}$ is

$$\frac{\pi^{k/2} (c_\alpha/n)^{k/2} |\boldsymbol{\Sigma}|^{1/2}}{\Gamma((k+1)/2)}.$$

Therefore, for two such estimators $\widehat{\boldsymbol{\eta}}^{(i)}$, $i = 1, 2$, the ARE given by (3) may be characterized as the limiting ratio of sample sizes at which the *volumes of associated ellipsoidal confidence regions at approximate level* $100(1-\alpha)\%$ converge to 0 at the same rate, when holding fixed the coverage probability $1 - \alpha$.

Under regularity conditions on the model, the maximum likelihood estimator $\widehat{\boldsymbol{\eta}}^{(\mathrm{ML})}$ has a confidence ellipsoid $E_{n,\alpha}$ attaining the *smallest possible volume* and, moreover, lying wholly within that for any other estimator $\widehat{\boldsymbol{\eta}}$.

Connections with Testing

Parallel to ARE in estimation as developed here is the notion of *Pitman ARE* for comparison of two hypothesis test procedures. Based on a different formulation, although the central limit theorem is used in common, the Pitman ARE agrees with (2) when the estimator and the hypothesis test statistic are linked, as for example $\overline{X}$ paired with the *t-test*, or Med_n paired with the ▶*sign test*, or HL_n paired with the ▶*Wilcoxon-signed-rank test*. See Lehmann 1998b, Nikitin 1995, Nikitin 2010, and Serfling 1980.

Other Notions of ARE

As illustrated above with FWARE, several other important approaches to ARE have been developed, typically using either moderate or large deviation theory. For example, instead of asymptotic variance parameters as the criterion, one may compare *probability concentrations* of the estimators in an ε-neighborhood of the target parameter η: $P\left(\left|\widehat{\eta}^{(i)} - \eta\right| > \varepsilon\right)$, $i = 1, 2$. When we have

$$\frac{\log P(|\widehat{\eta}^{(i)}_n - \eta| > \varepsilon)}{n} \to \gamma^{(i)}(\varepsilon, \eta), \; i = 1, 2,$$

as is typical, then the ratio of sample sizes n_1/n_2 at which these concentration probabilities converge to 0 at the same rate is given by $\gamma^{(1)}(\varepsilon, \eta)/\gamma^{(2)}(\varepsilon, \eta)$, which then represents another ARE measure for the efficiency of estimator $\widehat{\eta}^{(2)}_n$ relative to $\widehat{\eta}^{(1)}_n$. See Serfling (1980, 1.15.4) for discussion and Basu (1956) for illustration that the variance-based and concentration-based measures need not agree on which estimator is better. For general treatments, see Nikitin (1995), Puhalskii and Spokoiny (1998), Nikitin (2010), and Serfling (1980, Chap. 10), as well as the other references cited below. A comprehensive bibliography is beyond the present scope. However, very productive is *ad hoc* exploration of the literature using a modern search engine.

Acknowledgments

Support by NSF Grant DMS-0805786 and NSA Grant H98230-08-1-0106 is gratefully acknowledged.

About the Author

Robert Serfling is author of the classic textbook *Approximation Theorems of Mathematical Statistics*, Wiley, 1980, and has published extensively in statistical journals. He received a Humboldt-Preis, awarded by the Alexander von Humboldt Stiftung, Germany, "in recognition of accomplishments in research and teaching" (1985). He is a Fellow of the American Statistical Association and of Institute of Mathematical Statistics, and an Elected Member of International Statistical Institute. Professor Serfling was Editor of the IMS Lecture Notes Monograph Series (1988–1993) and currently is an Associate Editor for *Journal of Multivariate Analysis* (2007–) and for *Journal of Nonparametric Statistics* (2007–).

Cross References

▶Asymptotic Relative Efficiency in Testing
▶Estimation
▶Estimation: An Overview
▶Mean Median and Mode
▶Normality Tests

►Properties of Estimators
►Statistical Fallacies: Misconceptions, and Myths

References and Further Reading

Basu D (1956) On the concept of asymptotic relative efficiency. Sankhyā 17:193–196
Hodges JL, Lehmann EL (1963) Estimates of location based on rank tests. Ann Math Stat 34:598–611
Lehmann EL (1998a) Elements of large-sample theory. Springer, New York
Lehmann EL (1998b) Nonparametrics: statistical methods based on ranks. Prentice-Hall, Upper Saddle River, NJ
Lehmann EL, Casella G (1988) Theory of point estimation, 2nd edn. Springer, New York
Monohan JF (1984) Algorithm 616: fast computation of the Hodges–Lehmann location estimator. ACM T Math Software 10:265–270
Nikitin Y (1995) Asymptotic efficiency of nonparametric tests. Cambridge University Press, Cambridge
Nikitin Y (2010) Asymptotic relative efficiency in testing. International Encyclopedia of Statistical Sciences. Springer, New York
Puhalskii A, Spokoiny V (1998) On large-deviation efficiency in statistical inference. Bernoulli 4:203–272
Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. J Am Stat Assoc 88:1273–1283
Serfling R (1980) Approximation Theorems of Mathematical Statistics. Wiley, New York
Serfling R (2002) Efficient and robust fitting of lognormal distributions. N Am Actuarial J 4:95–109
Serfling R, Wackerly DD (1976) Asymptotic theory of sequential fixed-width confidence interval procedures. J Am Stat Assoc 71:949–955
Staudte RG, Sheather SJ (1990) Robust estimation and testing. Wiley, New York

Asymptotic Relative Efficiency in Testing

Yakov Nikitin
Professor, Chair of Probability and Statistics
St. Petersburg University, St. Petersburg, Russia

Asymptotic Relative Efficiency of Two Tests

Making a substantiated choice of the most efficient statistical test of several ones being at the disposal of the statistician is regarded as one of the basic problems of Statistics. This problem became especially important in the middle of XX century when appeared computationally simple but "inefficient" rank tests.

Asymptotic relative efficiency (ARE) is a notion which enables to implement in large samples the quantitative comparison of two different tests used for testing of the same statistical hypothesis. The notion of the asymptotic efficiency of *tests* is more complicated than that of asymptotic efficiency of *estimates*. Various approaches to this notion were identified only in late forties and early fifties, hence, 20–25 years later than in the estimation theory. We proceed now to their description.

Let $\{T_n\}$ and $\{V_n\}$ be two sequences of statistics based on n observations and assigned for testing the null-hypothesis H against the alternative A. We assume that the alternative is characterized by real parameter θ and for $\theta = \theta_0$ turns into H. Denote by $N_T(\alpha, \beta, \theta)$ the sample size necessary for the sequence $\{T_n\}$ in order to attain the power β under the level α and the alternative value of parameter θ. The number $N_V(\alpha, \beta, \theta)$ is defined in the same way.

It is natural to prefer the sequence with smaller N. Therefore the relative efficiency of the sequence $\{T_n\}$ with respect to the sequence $\{V_n\}$ is specified as the quantity

$$e_{T,V}(\alpha, \beta, \theta) = N_V(\alpha, \beta, \theta) \big/ N_T(\alpha, \beta, \theta),$$

so that it is the reciprocal ratio of sample sizes N_T and N_V.

The merits of the relative efficiency as means for comparing the tests are universally acknowledged. Unfortunately it is extremely difficult to explicitly compute $N_T(\alpha, \beta, \theta)$ even for the simplest sequences of statistics $\{T_n\}$. At present it is recognized that there is a possibility to avoid this difficulty by calculating the limiting values $e_{T,V}(\alpha, \beta, \theta)$ as $\theta \to \theta_0$, as $\alpha \to 0$ and as $\beta \to 1$ keeping two other parameters fixed. These limiting values $e^P_{T,V}$, $e^B_{T,V}$ and $e^{HL}_{T,V}$ are called respectively the Pitman, Bahadur and Hodges–Lehmann asymptotic relative efficiency (ARE), they were proposed correspondingly in Pitman (1949), Bahadur (1960) and Hodges and Lehmann (1956).

Only close alternatives, high powers and small levels are of the most interest from the practical point of view. It keeps one assured that the knowledge of these ARE types will facilitate comparing concurrent tests, thus producing well-founded application recommendations.

The calculation of the mentioned three basic types of efficiency is not easy, see the description of theory and many examples in Serfling (1980), Nikitin (1995) and Van der Vaart (1998). We only mention here, that Pitman efficiency is based on the central limit theorem (see ►Central Limit Theorems) for test statistics. On the contrary, Bahadur efficiency requires the large deviation asymptotics of test statistics under the null-hypothesis, while Hodges–Lehmann efficiency is connected with large deviation asymptotics under the alternative. Each type of efficiency has its own merits and drawbacks.

Pitman Efficiency

Pitman efficiency is the classical notion used most often for the asymptotic comparison of various tests. Under some regularity conditions assuming ▶asymptotic normality of test statistics under H and A, it is a number which has been gradually calculated for numerous pairs of tests.

We quote now as an example one of the first Pitman's results that stimulated the development of nonparametric statistics. Consider the two-sample problem when under the null-hypothesis both samples have the same continuous distribution and under the alternative differ only in location. Let $e^P_{W,t}$ be the Pitman ARE of the two-sample Wilcoxon rank sum test (see ▶Wilcoxon–Mann–Whitney Test) with respect to the corresponding Student test (see ▶Student's t-Tests). Pitman proved that for Gaussian samples $e^P_{W,t} = 3/\pi \approx 0.955$, and it shows that the ARE of the Wilcoxon test in the comparison with the Student test (being optimal in this problem) is unexpectedly high. Later Hodges and Lehmann (1956) proved that

$$0.864 \le e^P_{W,t} \le +\infty,$$

if one rejects the assumption of normality and, moreover, the lower bound is attained at the density

$$f(x) = \begin{cases} 3\left(5 - x^2\right)/\left(20\sqrt{5}\right) & \text{if } |x| \le \sqrt{5}, \\ 0 & \text{otherwise.} \end{cases}$$

Hence the Wilcoxon rank test can be infinitely better than the parametric test of Student but their ARE never falls below 0.864. See analogous results in Serfling (2010) where the calculation of ARE of related estimators is discussed.

Another example is the comparison of independence tests based on Spearman and Pearson correlation coefficients in bivariate normal samples. Then the value of Pitman efficiency is $9/\pi^2 \approx 0.912$.

In numerical comparisons, the Pitman efficiency appear to be more relevant for moderate sample sizes than other efficiencies Groeneboom and Oosterhoff (1981). On the other hand, Pitman ARE can be insufficient for the comparison of tests. Suppose, for instance, that we have a normally distributed sample with the mean θ and variance 1 and we are testing $H : \theta = 0$ against $A : \theta > 0$. Let compare two significance tests based on the sample mean $\bar{X}$ and the Student ratio t. As the t-test does not use the information on the known variance, it should be inferior to the optimal test using the sample mean. However, from the point of view of Pitman efficiency, these two tests are equivalent. On the contrary, Bahadur efficiency $e^B_{t,\bar{X}}(\theta)$ is strictly less than 1 for any $\theta > 0$.

If the condition of asymptotic normality fails, considerable difficulties arise when calculating the Pitman ARE as the latter may not at all exist or may depend on α and β. Usually one considers limiting Pitman ARE as $\alpha \to 0$. Wieand (1976) has established the correspondence between this kind of ARE and the limiting approximate Bahadur efficiency which is easy to calculate.

Bahadur Efficiency

The Bahadur approach proposed in Bahadur (1960; 1967) to measuring the ARE prescribes one to fix the power of tests and to compare the exponential rate of decrease of their sizes for the increasing number of observations and fixed alternative. This exponential rate for a sequence of statistics $\{T_n\}$ is usually proportional to some non-random function $c_T(\theta)$ depending on the alternative parameter θ which is called the *exact slope* of the sequence $\{T_n\}$. The Bahadur ARE $e^B_{V,T}(\theta)$ of two sequences of statistics $\{V_n\}$ and $\{T_n\}$ is defined by means of the formula

$$e^B_{V,T}(\theta) = c_V(\theta) \,/\, c_T(\theta).$$

It is known that for the calculation of exact slopes it is necessary to determine the large deviation asymptotics of a sequence $\{T_n\}$ under the null-hypothesis. This problem is always nontrivial, and the calculation of Bahadur efficiency heavily depends on advancements in large deviation theory, see Dembo and Zeitouni (1998) and Deuschel and Stroock (1989).

It is important to note that there exists an upper bound for exact slopes

$$c_T(\theta) \le 2K(\theta)$$

in terms of Kullback–Leibler information number $K(\theta)$ which measures the "statistical distance" between the alternative and the null-hypothesis. It is sometimes compared in the literature with the ▶Cramér–Rao inequality in the estimation theory. Therefore the absolute (nonrelative) Bahadur efficiency of the sequence $\{T_n\}$ can be defined as $e^B_T(\theta) = c_T(\theta)/2K(\theta)$.

It is proved that under some regularity conditions the likelihood ratio statistic is asymptotically optimal in Bahadur sense (Bahadur 1967; Van der Vaart 1998, Sect. 16.6; Arcones 2005).

Often the exact Bahadur ARE is uncomputable for any alternative θ but it is possible to calculate the limit of Bahadur ARE as θ approaches the null-hypothesis. Then one speaks about the *local* Bahadur efficiency.

The indisputable merit of Bahadur efficiency consists in that it can be calculated for statistics with non-normal asymptotic distribution such as Kolmogorov-Smirnov, omega-square, Watson and many other statistics.

Asymptotic Relative Efficiency in Testing. Table 1 Some local Bahadur efficiencies

Statistic	Distribution					
	Gauss	Logistic	Laplace	Hyperbolic cosine	Cauchy	Gumbel
D_n	0.637	0.750	1	0.811	0.811	0.541
ω_n^2	0.907	0.987	0.822	1	0.750	0.731

Consider, for instance, the sample with the distribution function (df) F and suppose we are testing the goodness-of-fit hypothesis $H_0 : F = F_0$ for some known continuous df F_0 against the alternative of location. Well-known distribution-free statistics for this hypothesis are the Kolmogorov statistic D_n and omega-square statistic ω_n^2. The following table presents their local absolute efficiency in case of six standard underlying distributions:

We see from Table 1 that the integral statistic ω_n^2 is in most cases preferable with respect to the supremum-type statistic D_n. However, in the case of Laplace distribution the Kolmogorov statistic is locally optimal, the same happens for the Cramér-von Mises statistic in the case of hyperbolic cosine distribution. This observation can be explained in the framework of Bahadur local optimality, see Nikitin (1995 Chap. 6).

See also Nikitin (1995) for the calculation of local Bahadur efficiencies in case of many other statistics.

Hodges–Lehmann efficiency

This type of the ARE proposed in Hodges and Lehmann (1956) is in the conformity with the classical Neyman-Pearson approach. In contrast with Bahadur efficiency, let us fix the level of tests and let compare the exponential rate of decrease of their second-kind errors for the increasing number of observations and fixed alternative. This exponential rate for a sequence of statistics $\{T_n\}$ is measured by some non-random function $d_T(\theta)$ which is called the Hodges–Lehmann index of the sequence $\{T_n\}$. For two such sequences the Hodges–Lehmann ARE is equal to the ratio of corresponding indices.

The computation of Hodges–Lehmann indices is difficult as requires large deviation asymptotics of test statistics under the alternative.

There exists an upper bound for the Hodges–Lehmann indices analogous to the upper bound for Bahadur exact slopes. As in the Bahadur theory the sequence of statistics $\{T_n\}$ is said to be *asymptotically optimal in the Hodges–Lehmann sense* if this upper bound is attained.

The drawback of Hodges–Lehmann efficiency is that most *two-sided* tests like Kolmogorov and Cramér-von Mises tests are asymptotically optimal, and hence this kind of efficiency cannot discriminate between them. On the other hand, under some regularity conditions the one-sided tests like linear rank tests can be compared on the basis of their indices, and their Hodges–Lehmann efficiency coincides locally with Bahadur efficiency, see details in Nikitin (1995).

Coupled with three "basic" approaches to the ARE calculation described above, intermediate approaches are also possible if the transition to the limit occurs simultaneously for two parameters at a controlled way. Thus emerged the Chernoff ARE introduced by Chernoff (1952), see also Kallenberg (1982); the intermediate, or the Kallenberg ARE introduced by Kallenberg (1983), and the Borovkov–Mogulskii ARE, proposed in Borovkov and Mogulskii (1993).

Large deviation approach to asymptotic efficiency of tests was applied in recent years to more general problems. For instance, the change-point, "signal plus white noise" and regression problems were treated in Puhalskii and Spokoiny (1998), the tests for spectral density of a stationary process were discussed in Kakizawa (2005), while Taniguchi (2001) deals with the time series problems, and Otsu (2010) studies the empirical likelihood for testing moment condition models.

About the Author

Professor Nikitin is Chair of Probability and Statistics of St. Petersburg University. He is an Associate editor of *Statistics and Probability Letters*, and member of the editorial Board, *Mathematical Methods of Statistics* and *Metron*. He is a Fellow of the Institute of Mathematical Statistics. Professor Nikitin is the author of the text *Asymptotic efficiency of nonparametric tests*, Cambridge University Press, NY, 1995, and has authored more than 100 papers, in many international journals, in the field of Asymptotic efficiency of statistical tests, large deviations of test statistics and nonparametric Statistics.

Cross References

►Asymptotic Relative Efficiency in Estimation
►Chernoff-Savage Theorem
►Nonparametric Statistical Inference
►Robust Inference

References and Further Reading

Arcones M (2005) Bahadur efficiency of the likelihood ratio test. Math Method Stat 14:163–179

Bahadur RR (1960) Stochastic comparison of tests. Ann Math Stat 31:276–295

Bahadur RR (1967) Rates of convergence of estimates and test statistics. Ann Math Stat 38:303–324

Borovkov A, Mogulskii A (1993) Large deviations and testing of statistical hypotheses. Siberian Adv Math 2(3, 4); 3(1, 2)
Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on sums of observations. Ann Math Stat 23:493–507
Dembo A, Zeitouni O (1998) Large deviations techniques and applications, 2nd edn. Springer, New York
Deuschel J-D, Stroock D (1989) Large deviations. Academic, Boston
Groeneboom P, Oosterhoff J (1981) Bahadur efficiency and small sample efficiency. Int Stat Rev 49:127–141
Hodges J, Lehmann EL (1956) The efficiency of some nonparametric competitors of the t-test. Ann Math Stat 26:324–335
Kakizawa Y (2005) Bahadur exact slopes of some tests for spectral densities. J Nonparametric Stat 17:745–764
Kallenberg WCM (1983) Intermediate efficiency, theory and examples. Ann Stat 11:170–182
Kallenberg WCM (1982) Chernoff efficiency and deficiency. Ann Stat 10:583–594
Nikitin Y (1995) Asymptotic efficiency of nonparametric tests. Cambridge University Press, Cambridge
Otsu T (2010) On Bahadur efficiency of empirical likelihood. J Econ 157:248–256
Pitman EJG (1949) Lecture notes on nonparametric statistical inference. Columbia University, Mimeographed
Puhalskii A, Spokoiny V (1998) On large-deviation efficiency in statistical inference. Bernoulli 4:203–272
Serfling R (1980) Approximation theorems of mathematical statistics. Wiley, New York
Serfling R (2010) Asymptotic relative efficiency in estimation. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer
Taniguchi M (2001) On large deviation asymptotics of some tests in time series. J Stat Plann Inf 97:191–200
Van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge
Wieand HS (1976) A condition under which the Pitman and Bahadur approaches to efficiency coincide. Ann Statist 4:1003–1011

Asymptotic, Higher Order

Juan Carlos Abril
President of the Argentinean Statistical Society, Professor Universidad Nacional de Tucumán and Consejo Nacional de Investigaciones Científicas y Técnicas, San Miguel de Tucumán, Argentina

Higher order asymptotic deals with two sorts of closely related things. First, there are questions of approximation. One is concerned with expansions or inequalities for a distribution function. Second, there are inferential issues. These involve, among other things, the application of the ideas of the study of higher order efficiency, admissibility and minimaxity. In the matter of expansions, it is as important to have usable, explicit formulas as a rigorous proof that the expansions are valid in the sense of truly approximating a target quantity up to the claimed degree of accuracy.

Classical asymptotics is based on the notion of asymptotic distribution, often derived from the central limit theorem (see ►Central Limit Theorems), and usually the approximations are correct up to $O(n^{-1/2})$, where n is the sample size. Higher order asymptotics provides refinements based on asymptotic expansions of the distribution or density function of an estimator of a parameter. They are rooted in the Edgeworth theory, which is itself a refinement of the central limit theorem. The theory of higher order asymptotic is very much related with the corresponding to *Approximations to distributions* treated as well in this Encyclopedia.

When higher order asymptotic is correct up to $o(n^{-1/2})$, it is second order asymptotic. When further terms are picked up, so that the asymptotic is correct up to $o(n^{-1})$, it is third order asymptotic. In his pioneering papers, C. R. Rao coined the term second order efficiency for a concept that would now is called third order efficiency. The new terminology is essentially owing to Pfanzagl and Takeuchi.

About the Author

Professor Abril is co-editor of the *Revista de la Sociedad Argentina de Estadística* (Journal of the Argentinean Statistical Society).

Cross References

►Approximations to Distributions
►Edgeworth Expansion

References and Further Reading

Abril JC (1985) Asymptotic expansions for time series problems with applications to moving average models. PhD thesis. The London School of Economics and Political Science, University of London, England
Barndorff-Nielsen O, Cox DR (1979) Edgeworth and saddle-point approximations with statistical applications. J R Stat Soc B 41:279–312
Daniels HE (1954) Saddlepoint approximations in statistics. Ann Math Stat 25:631–650
Durbin J (1980) Approximations for the densities of sufficient estimates. Biometrika 67:311–333
Feller W (1971) An introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York
Ghosh JK (1994) Higher order Asymptotic. NSF-CBMS Regional Conference Series in Probability and Statistics, 4. Hayward and Alexandria: Institute of Mathematical Statistics and American Statistical Association
Pfanzagl J (1979) Asymptotic expansions in parametric statistical theory. In: Krishnaiah PR (ed) Developments in statistics, vol. 3. Academic, New York, pp 1–97
Rao CR (1961) Asymptotic efficiency and limiting information. In Proceedings of Fourth Berkeley Symposium on Mathematical

Statistics and Probability. Berkeley: University of California Press, pp 531–546
Rao CR (1962) Efficient estimates and optimum inference procedure in large samples. J R Stat Soc B 24:46–63
Rao CR (1963) Criteria of estimation in large samples. Sankhya B 25:189–206
Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, New York
Wallace DL (1958) Asymptotic approximations to distributions. Ann Math Stat 29:635–654

Autocorrelation in Regression

Badi H. Baltagi
Distinguished Professor of Economics
Syracuse University, Syracuse, NY, USA

Linear regressions are a useful empirical tool for economists and social scientists and the standard ►least squares estimates are popular because they are the best linear unbiased estimators (BLUE) under some albeit strict assumptions. These assumptions require the regression disturbances not to be correlated with the regressors, also homoskedastic, i.e., with constant variance, and not autocorrelated. Violation of the no autocorrelation assumption on the disturbances, will lead to inefficiency of the least squares estimates, i.e., no longer having the smallest variance among all linear unbiased estimators. It also leads to wrong standard errors for the regression coefficient estimates. This in turn leads to wrong t-statistics on the significance of these regression coefficients and misleading statistical inference based on a wrong estimate of the variance–covariance matrix computed under the assumption of no autocorrelation. This is why standard regression packages have a robust heteroskedasticity and autocorrelation-consistent covariance matrix (HAC) option for these regressions which at least robustifies the standard errors of least squares and shows how sensitive they would be to such violations, see Newey and West (1987).

Autocorrelation is more likely to occur in time-series than in cross-section studies. Consider estimating the consumption function of a random sample of households. An unexpected event, like a visit of family members will increase the consumption of this household. However, this positive disturbance need not be correlated with the disturbances affecting consumption of other randomly drawn households. However, if we were estimating this consumption function using aggregate time-series data for the U.S., then it is very likely that a recession year affecting consumption negatively that year, may have a carry over effect to the next few years. A shock to the economy like an oil embargo in 1973 is likely to affect the economy for several years. A labor strike this year may affect production for the next few years. The simplest work horse for illustrating this autocorrelation in time series on the regression disturbances, say u_t is the first-order autoregressive process denoted by AR(1):

$$u_t = \rho u_{t-1} + \epsilon_t \qquad t = 1, 2, \ldots, T$$

where ϵ_t is independent and identically distributed with mean 0 and variance σ_ϵ^2. It is autoregressive because u_t is related to its lagged value u_{t-1}. One can show, see for example Baltagi (2008), that the correlation coefficient between u_t and u_{t-1} is ρ. Also, that the correlation coefficient between u_t and u_{t-r} is ρ^r. When $\rho = 0$, there is no autocorrelation and one test for this null hypothesis is the Durbin and Watson (1951) test discussed as a separate entry in this encyclopedia by Krämer. This AR(1) process is also stationary as long as $|\rho| < 1$. If $\rho = 1$, then this process has a unit root and it is called a ►random walk. See the entry by Dickey on testing for this unit root using the ►Dickey-Fuller tests. Note that if the process is stationary, then ρ is a fraction and the correlation for two disturbances r periods apart is ρ^r, i.e., a fraction raised to an integer power. This means that the correlation is decaying between the disturbances the further apart they are. This is reasonable in economics and may be the reason why this AR(1) form is so popular. One should note that this is not the only form that would correlate the disturbances across time. Other forms like the Moving Average (MA) process, and higher order Autoregressive Moving Average (ARMA) processes are popular, see Box and Jenkins (1970), but these are beyond the scope of this entry.

Since least squares is no longer BLUE under autocorrelation of the disturbances, Cochrane and Orcutt (1949) suggested a simple estimator that corrects for autocorrelation of the AR(1) type. This method starts with an initial estimate of ρ, the most convenient is 0, and minimizes the residual sum of squares of the regression. This gives us the least squares estimates of the regression coefficients and the corresponding least squares residuals which we denote by e_t. In the next step, one regresses e_t on e_{t-1} to get an estimate of ρ, say $\widehat{\rho}$. The second step of the Cochrane–Orcutt procedure (2SCO) is to perform the regression of $(Y_t - \widehat{\rho} Y_{t-1})$ on $(X_t - \widehat{\rho} X_{t-1})$ to get estimates of the regression coefficients. One can iterate this procedure (ITCO) until convergence. Both the 2SCO and the ITCO are asymptotically efficient as the sample size gets large. The argument for iterating must be justified in terms of small sample gains. Other methods of correcting for serial correlation include Prais and Winsten (1954), Durbin (1960), as well as maximum likelihood methods, all studied more extensively in Chap. 5 of Baltagi (2008). The Prais and

Winsten method recaptures the initial observation lost in the Cochrane–Orcutt method. Monte Carlo studies using an autoregressive regressor, and various values of ρ, found that least squares is still a viable estimator as long as $|\rho| < 0.3$, but if $|\rho| > 0.3$, then it pays to perform procedures that correct for serial correlation based on an estimator of ρ. For trended regressors, which is usually the case with economic data, least squares outperforms 2SCO, but not the Prais-Winsten procedure that recaptures the initial observation. In fact, Park and Mitchell (1980) who performed an extensive Monte Carlo using trended and untrended regressors recommend that one should not use regressions based on $(T - 1)$ observations as in the Cochrane and Orcutt procedure. They also found that test of hypotheses regarding the regression coefficients performed miserably for all estimators based on an estimator of ρ.

Correcting for serial correlation is not without its critics. Mizon (1995) argues this point forcefully in his article entitled "A simple message for autocorrelation correctors: Don't." The main point being that serial correlation is a symptom of dynamic misspecification which is better represented using a general unrestricted dynamic specification.

About the Author

Badi H. Baltagi is distinguished Professor of Economics, and Senior Research Associate at the Center for Policy Research, Syracuse University. He received his Ph.D. in Economics at the University of Pennsylvania in 1979. He served on the faculty at the University of Houston and Texas A&M University. He was a visiting Professor at the University of Arizona and the University of California, San Diego. He is the author of *Econometric Analysis of Panel Data* (Wiley, 4*th* edn. 2008); *Econometrics* (Springer, 4*th* edn. 2008), and editor of *A Companion to Theoretical Econometrics* (Wiley–Blackwell); *Recent Developments in the Econometrics of Panel Data*, Volumes I and II (Edward-Elgar); *Nonstationary Panels*, *Panel Cointegration*, and *Dynamic Panels* (Elsevier); *Panel Data Econometrics: Theoretical Contributions and Empirical Applications* (Physica-Verlag, 2004); *Spatial Econometrics: Methods and Applications*, (with Giuseppe Arbia), Physica-Verlag, 2009. He is author or co-author of over 100 publications, all in leading economics and statistics journals. Professor Baltagi was the holder of the George Summey, Jr. Professor Chair in Liberal Arts and was awarded the Distinguished Achievement Award in Research at Texas A&M University. He is co-editor of *Empirical Economics*, and Associate editor of *Journal of Econometrics* and *Econometric Reviews*. He is the replication editor of the *Journal of Applied Econometrics* and the series editor for *Contributions to Economic Analysis*. He is a fellow of the Journal of Econometrics and a recipient of the Multa and Plura Scripsit Awards from Econometric Theory. He is a founding fellow and member of the Board of Directors of the Spatial Econometrics Association.

Cross References

►Approximations to Distributions
►Box–Jenkins Time Series Models
►Correlation Coefficient
►Dickey-Fuller Tests
►Durbin–Watson Test
►Linear Regression Models
►Structural Time Series Models
►Tests of Independence
►Time Series
►Time Series Regression

References and Further Reading

Baltagi BH (2008) Econometrics, 4th edn., Springer, Berlin

Box GEP, Jenkins GM (1970) Time series analysis, forecasting and control. Holden Day, San Francisco

Cochrane D, Orcutt G (1949) Application of least squares regression to relationships containing autocorrelated error terms. J Am Stat Assoc 44:32–61

Durbin J (1960) Estimation of parameters in time-series regression model. J R Stat Soc, B, 22:139–153

Durbin J, Watson G (1951) Testing for serial correlation in least squares regression-II. Biometrika 38:159–178

Mizon GE (1995) A simple message for autocorrelation correctors: don't. J Econometrics 69:267–288

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55:703–708

Park RE, Mitchell BM (1980) Estimating the autocorrelated error model with trended data. J Econometrics 13:185–201

Prais S, Winsten C (1954) Trend estimation and serial correlation. Discussion Paper 383, Cowles Commission, Chicago

Axioms of Probability

Vincenzo Capasso
Professor of Probability and Mathematical Statistics
University of Milan, Milan, Italy

Ingredients of Probability Spaces

Definition 1 A collection $\mathcal{F}$ of subsets of a set Ω is called a *ring* on Ω if it satisfies the following conditions:

1. $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$,
2. $A, B \in \mathcal{F} \Rightarrow A \setminus B \in \mathcal{F}$.

A ring $\mathcal{F}$ is called an *algebra* if $\Omega \in \mathcal{F}$.

Definition 2 A ring $\mathcal{F}$ on Ω is called a *σ-ring* if it satisfies the following additional condition:

3. For every countable family $(A_n)_{n\in\mathbb{N}}$ of subsets of $\mathcal{F}$: $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

A σ-ring $\mathcal{F}$ on Ω is called a *σ-algebra* (or *σ-field*) if $\Omega \in \mathcal{F}$.

Proposition 1 The following properties hold:

1. If $\mathcal{F}$ is a σ-algebra of subsets of a set Ω, then it is an algebra.
2. If $\mathcal{F}$ is a σ-algebra of subsets of Ω, then
   - For any countable family $(E_n)_{n\in\mathbb{N}\setminus\{0\}}$ of elements of $\mathcal{F}$: $\bigcap_{n=1}^{\infty} E_n \in \mathcal{F}$,
   - For any finite family $(E_i)_{1\le i\le n}$ of elements of $\mathcal{F}$: $\bigcap_{i=1}^{n} E_i \in \mathcal{F}$,
   - $B \in \mathcal{F} \Rightarrow \Omega \setminus B \in \mathcal{F}$.

Definition 3 Every pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a σ-ring $\mathcal{F}$ of subsets of Ω is a *measurable space*. Furthermore, if $\mathcal{F}$ is a σ-algebra, then $(\Omega, \mathcal{F})$ is a *measurable space on which a probability measure can be built*.

Example 1

1. *Generated σ-algebra*. If $\mathcal{A}$ is a family of subsets of a set Ω, then there exists the smallest σ-algebra of subsets of Ω that contains $\mathcal{A}$. This is the σ-algebra *generated* by $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$. If, now, $\mathcal{G}$ is the set of all σ-algebras of subsets of Ω containing $\mathcal{A}$, then it is not empty because it has the set $\mathcal{P}(\Omega)$ of all subsets of Ω, among its elements, so that $\sigma(\mathcal{A}) = \bigcap_{\mathcal{C}\in\mathcal{G}} \mathcal{C}$.
2. *Borel σ-algebra*. Let Ω be a topological space. Then the *Borel σ-algebra* on Ω, denoted by $\mathcal{B}_\Omega$, is the σ-algebra generated by the set of all open subsets of Ω. Its elements are called Borel sets.
3. *Product σ-algebra*. Let $(\Omega_i, \mathcal{F}_i)_{1\le i\le n}$ be a family of measurable spaces, with all $\mathcal{F}_i, 1 \le i \le n$, σ-algebras, and let $\Omega = \prod_{i=1}^{n} \Omega_i$. Defining

$$\mathcal{R} = \left\{ E \subset \Omega \,|\, \forall i = 1, \ldots, n \; \exists E_i \in \mathcal{F}_i \text{ such that } E = \prod_{i=1}^{n} E_i \right\},$$

   the σ-algebra on Ω generated by $\mathcal{R}$ is called the *product σ-algebra* of the σ-algebras $(\mathcal{F}_i)_{1\le i\le n}$.

Proposition 2 Let $(\Omega_i)_{1\le i\le n}$ be a family of topological spaces with a countable base and let $\Omega = \prod_{i=1}^{n} \Omega_i$. Then the Borel σ-algebra $\mathcal{B}_\Omega$ is identical to the product σ-algebra of the family of Borel σ-algebras $(\mathcal{B}_{\Omega_i})_{1\le i\le n}$.

Axioms of Probability

We assume that the reader is already familiar with the basic motivations and notions of probability theory. We present the axioms of probability according to the Kolmogorov approach [see Kolmogorov (1956)].

Definition 4 Given a set Ω, and a σ-algebra $\mathcal{F}$ of subsets of Ω, a probability measure on $\mathcal{F}$ is any function $P : \mathcal{F} \to [0,1]$ such that

P_1. $P(\Omega) = 1$,

P_2. for any countable family $A_1, \ldots, A_n, \ldots$ of elements of $\mathcal{F}$ such that $A_i \cap A_j = \emptyset$, whenever $i \neq j$:

$$P\left(\bigcup_n A_n\right) = \sum_n P(A_n).$$

Definition 5 A *probability space* is an ordered triple $(\Omega, \mathcal{F}, P)$, where Ω is a set, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and $P : \mathcal{F} \to [0,1]$ is a probability measure on $\mathcal{F}$. The set Ω is called the *sample space*, the elements of $\mathcal{F}$ are called *events*.

Definition 6 A probability space $(\Omega, \mathcal{F}, P)$ is *finite* if Ω has finitely many elements.

Remark 1 If Ω is at most countable, then it is usual to assume that $\mathcal{F} = \mathcal{P}(\Omega)$, the σ-algebra of all subsets of Ω. In this case all sets $\{\omega\}$ reduced to sample points $\omega \in \Omega$ are events; they are called *elementary events*.

Remark 2 If the σ−algebra of events $\mathcal{F}$ is finite, then the requirement of countable additivity in the definition of the probability measure P can be reduced to finite additivity.

Remark 3 It is worth mentioning that an important alternative approach to probability theory is the so called *subjective probability*; this approach does not insist on Axiom P_2., and rather uses the finite version of it (De Finetti 1974–1975).

Definition 7 A finite probability space $(\Omega, \mathcal{F}, P)$ with $\mathcal{F} = \mathcal{P}(\Omega)$ is an *equiprobable* or *uniform* space, if

$$\forall\, \omega \in \Omega : \qquad P(\{\omega\}) = k \text{ (constant)};$$

i.e., its elementary events are equiprobable.

Remark 4 Following the axioms of a probability space and the definition of a uniform space, if $(\Omega, \mathcal{F}, P)$ is equiprobable, then

$$\forall\, \omega \in \Omega : \qquad P(\{\omega\}) = \frac{1}{|\Omega|},$$

where $|\cdot|$ denotes the cardinal number of elementary events in Ω, and

$$\forall A \in \mathcal{F} \equiv \mathcal{P}(\Omega) : \qquad P(A) = \frac{|A|}{|\Omega|}.$$

Intuitively, in this case we may say that $P(A)$ is the ratio of the number of favorable outcomes, divided by the number of all possible outcomes.

Example 2 Consider an urn that contains 100 balls, of which 80 are red and 20 are black but that are otherwise identical, from which a player draws a ball. Define the event

R: The first drawn ball is red.

Then

$$P(R) = \frac{|R|}{|\Omega|} = \frac{80}{100} = 0.8.$$

Definition 8 We shall call any event $F \in \mathcal{F}$ such that $P(F) = 0$, a *null event*.

Elementary consequences of the above definitions are the following ones.

Proposition 3 Let $(\Omega, \mathcal{F}, P)$ be is a probability space.

1. $P(A^c) = 1 - P(A)$, for any $A \in \mathcal{F}$;
2. $P(\emptyset) = 0$;
3. If $A, B \in \mathcal{F}$, $A \subseteq B$, then $P(B) = P(A) + P(B \setminus A)$;
4. If $A, B \in \mathcal{F}$, $A \subseteq B$, then $P(A) \leq P(B)$ (monotonicity);
5. If $A, B \in \mathcal{F}$, then

$$P(B \setminus A) = P(B) - P(B \cap A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B);$$

6. If $A, B \in \mathcal{F}$, $A \subseteq B$, then $P(B \setminus A) = P(B) - P(A)$;
7. (Principle of inclusion-exclusion) Let $A_1, \ldots, A_n \in \mathcal{F}$, then

$$P\left(\cup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n);$$

8. If $A_1, \ldots, A_n \in \mathcal{F}$, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

About the Author

Fellow of the European Academy of Sciences (2003–), President of ECMI (the European Consortium for Mathematics in Industry) (1999–2001); Founder (1991) and President of the European Society for Mathematical and Theoretical Biology (2000–2002); Founder (1985) and Director of SASIAM: "School for Advanced Studies in Industrial and Applied Mathematics", Tecnopolis, Bari, Italy (1985–1991), Founder and Director of MIRIAM (Milan Research Centre for Industrial and Applied Mathematics) (1999–2005) and later of ADAMSS (Research Centre for Advanced Applied Mathematical and Statistical Sciences) of the University of Milano (2005–2007), Director of CIMAB (InterUniversity Centre for Mathematics Applied to Biology, Medicine, Environment, etc.) (2008–).

Cross References

►Foundations of Probability
►Imprecise Probability
►Measure Theory in Probability
►Philosophy of Probability
►Probability Theory: An Outline

References and Further Reading

Ash RB (1972) Real analysis and probability. Academic, London
Bauer H (1981) Probability theory and elements of measure theory. Academic, London
Billingsley P (1995) Probability and measure. Wiley, New York
Breiman L (1968) Probability. Addison–Wesley, Reading, MA
Chung KL (1974) A course in probability theory, 2nd edn. Academic, New York
De Finetti B (1974–1975) Theory of probability, vols 1 and 2. Wiley, London
Dudley RM (2002) Real analysis and probability. Cambridge Studies in Advanced Mathematics 74, Cambridge University Press, Cambridge
Fristedt B, Gray L (1997) A modern approach to probability theory. Birkhäuser, Boston
Kolmogorov AN (1956) Foundations of the theory of probability. Chelsea, New York
Métivier M (1968) Notions fondamentales de la théorie des probabilités., Dunod, Paris

B

Balanced Sampling

Yves Tillé
Professor
University of Neuchâtel, Neuchâtel, Switzerland

Balanced sampling is a random method of selection of units from a population that provides a sample such that the Horvitz–Thompson estimators (see ►Horvitz–Thompson Estimator) of the totals are the same or almost the same as the true population totals for a set of control variables.

More precisely, let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population and $s \subset U$ a sample or a subset of U. A sampling design $p(s)$ is a probability distribution on the set of all the subsets $s \subset U$, i.e. $p(s) \geq 0$ and

$$\sum_{s \subset U} p(s) = 1.$$

The inclusion probability $\pi_k = pr(k \in s)$ of a unit k is its probability of being selected in the sample s.

Consider a variable of interest y that takes the value y_k on unit k. Let also Y be the total of the values taken by y on the units of the population

$$Y = \sum_{k \in U} y_k.$$

If $\pi_k > 0$, for all $k \in U$, the Horvitz–Thompson (1952) estimator

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

is unbiased for Y.

Let also $x_1, \ldots, x_j, \ldots, x_J$ be a sequence of auxiliary variables whose the values $x_{k1}, \ldots, x_{kj}, \ldots, x_{kJ}$ are known for each unit k of the population. According to the definition given in Tillé (2006), a sampling design is said to be balanced on the x variables if

$$\sum_{k \in S} \frac{x_{kj}}{\pi_k} \approx \sum_{k \in U} x_{kj}, \text{ for } j = 1, \ldots, J.$$

Balanced sampling involves sampling with fixed sample size and stratification. Indeed, a stratified design is balanced on the indicator variables of the strata, because the Horvitz–Thompson estimators of the sizes of the strata are equal to the population sizes of the strata. In the design-based inference, balanced sampling allows a strong improvement of the efficiency of the Horvitz–Thompson estimator when the auxiliary variables are correlated with the interest variable (see Deville and Tillé 2004). In the model-based inference, balanced samples are advocated to protect under miss-specification of the model (see Valliant et al. 2000).

Balanced sampling must not be confused with a representative sample. Representativity is a vague concept that usually means that some groups have the same proportions in the sample and in the population. This definition is fallacious because some groups can be over or under-represent in a sample to obtain a more accurate unbiased estimator. Moreover, balanced sampling implies that the sample is randomly selected and that predefined inclusion probabilities, that can be unequal, are satisfied at least approximately.

There exists a large family of methods for selecting balanced samples. The first one was probably proposed by Yates (1949) and consists of selecting a sample by a simple random sampling and next eventually changing some units to get a more balanced sample. Other methods, called rejective, consist of selecting several samples with an initial sampling design until obtaining a sample that is well balanced. Rejective methods however have the drawback that the inclusion probabilities of the rejective design are not the same as the ones of the initial design and are generally impossible to compute.

The cube method proposed by Deville and Tillé (2004) is a general multivariate algorithm for selecting balanced samples that can use several decades of auxiliary variables with equal of unequal inclusion probabilities. The cube method exactly satisfies the predefined inclusion probabilities and provides a sample that is balanced as well as possible. SAS and R language implementations are available.

About the Author

For biography *see* the entry ►Sampling Algorithms.

Cross References

►Horvitz–Thompson Estimator
►Representative Samples

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

►Sample Survey Methods
►Sampling Algorithms
►Stratified Sampling

References and Further Reading

Deville J-C, Tillé Y (2004) Efficient balanced sampling: the cube method. Biometrika 91:893–912
Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685
Tillé Y (2006) Sampling algorithms. Springer, New York
Valliant R, Dorfman AH, Royall RM (2000) Finite population sampling and inference: a prediction approach. Wiley, New York
Yates F (1949) Sampling methods for censuses and surveys. Griffin, London

Banking, Statistics in

Héctor Manuel Zárate Solano
Professor, Head of Statistics Section
Universidad Nacional de Colombia, Banco de la República Bogotá, Colombia

Statistics in banking is becoming increasingly important to the extent that modern banking at all levels would be impossible without the application of statistical methods (Hand 2006, p. 361). Particularly the financial crisis has underlined the importance of high quality and timely data to the banking sector. Indeed, "information shortages – directly related to a lack of timely and accurate data – with regard to complex credit derivatives lie at the heart of the current crisis and have complicated the crisis management efforts of central banks around the world." (Current challenges and the Future of Central Bank Statistics).

According to the Monetary and Financial Statistics (International Monetary Fund 2000), the banking sector consists of all resident corporations mainly engaged in financial intermediation in any given economy. These corporations consist of the central bank which is the national financial institution with the principal responsibility of ensuring monetary stability and financial sector soundness (Bangura 2005), and other depository corporations including commercial retail banks, merchant banks, savings banks, credit unions, credit cooperatives, rural and agricultural banks, etc.

The range of statistical tools and models in banking applications is vast, starting from purely descriptive models, risk assessment, statistics of extreme values, Markov chain approaches (including mover–stayer models), cluster analysis (see ►Cluster Analysis: An Introduction), statistical methods of fraud detection (see Bolton and Hand 2002; Hand 2007, 2010; Sudjianto et al. 2010), ►logistic regression, classification trees, etc.

Statistics in Banking: The Colombian Perspective

The Central Bank has focused significantly on the use of statistics to conduct monetary policy. Statistical methodologies have evolved in quantity, depth, and degree of sophistication as economic and financial systems have become more complex because of the generalized economic growth over time, increased market participants, and all sorts of financial transactions. Moreover, the adoption annual inflation-targeting, based on the consumer price index (CPI), to achieve a long-term GDP growth trend as the monetary policy rule since 1991 has set new challenges for our statistics section and practitioners as well. As a result, we have had to adapt to new demands for information. First, the inflation forecasting model of the Banco de la República requires monitoring the inflationary pressures continually, and second, assessing the economic activity. To accomplish these, econometric models based on observational data have been put into place, and the Bank has made efforts in using best practices with respect to surveys as an alternative to full reporting.

To elaborate on the above, we need to produce statistics in an efficient way, including a new form of data analysis provided by individual reports coming from financial and nonfinancial institutions mostly through electronic reporting. The degree of sophistication of statistical techniques depends on the data availability. A priority task involves data compilation, which is a statistical activity that requires the implementation of new ideas that combine statistical techniques, view points from economists, and the management of large databases. For instance, nominal short-term interest rates have remained as a main policy tool in this country, which depend on the reliability of the micro data reported daily to the Bank. We have to establish graphical and statistical tools to guarantee the quality and accuracy of the aggregate data. Aggregation also involves the incorporation of new methodologies, such as the index number theory, which explains the behavior of crucial indicators over time in carrying out monetary policy management. Consequently, we have worked on building core inflation indices, forward looking indicators, stock indices, real estate prices, and real exchange rates. Furthermore, one important part of the transmission mechanism is to know how monetary policy affects aggregate demand and it has

been important to identify shocks of trade terms, fiscal policy, and real world price of commodities.

With regard to surveys, we should note that the recent increased responses of financial markets, and others, to economic expectations. Thus, the Bank has made several surveys protecting statistical confidentiality: The Monthly Business Surveys, which provide clues on formation expectations in the private sector and fill the need for more timely indicators that can be used to gain insight into the economic climate before official statistics are published. Furthermore, they indicate signals of turning points in economic activity. The regional aspect of this survey also creates network contacts in the business community. Another survey is the Monthly Expert Inflation Forecasts, which summarize short macroeconomic forecasts from the experts. The statistics derived from the latter provide relevant information on the professional inflation consensus and allow us to assess the credibility of monetary policy through the dispersion of the forecast of the inflation. In addition, we developed Quarterly Market Expectation Surveys to ascertain the entrepreneurship perception of current and expected economic developments over the very short term in the main macroeconomic variables. We are also working on surveys involved with remittance flows and the cost of remitting money. Nevertheless, the realization of surveys has imposed central bank statisticians do serious efforts on issues related to statistical survey methodologies. For instance, introducing good questionnaire design, using probabilistic sampling techniques, and assuring adequate response rates. Also, we have to notice difficulty of survey implementation as there is a lack of incentives for respondent to cooperate, absence of proper legal mandate to collect information, and resources constraints. These factors are considered through the cost-benefit evaluation of the generation of this kind of information.

Statistics is decisive to monetary policy decisions by central bank policy makers, who face diverse forms of uncertainty, and it is important to rely on various information sources assessment of economic statistics and judgment. In fact, they need to analyze financial markets, watch carefully national accounts and labor force statistics. This statistical challenge has intensified cooperation with the national statistical institute (DANE) and producers of statistics in the private sector to try to improve data coherence, minimize the response burden to reporting agents, exchange micro data, share methodologies, and ameliorate data quality under international recognized standards. In addition, the experience from other countries has led to coordinate regional tasks in order to homogenize key economic indicators. It is also of paramount importance for the central bank's credibility to keep the public clearly informed about the actions that have been taken by the monetary authorities. As a result, new ways of presenting statistics have been implemented in different channels of communication such as Inflation Report, Report to Congress, and Press releases, among others. Statistics contribute to understand trends in the economy policy evaluations and help to design future policies. Therefore, the ability to research inflation issues and find recent statistical methodologies to present new economic indicators mainly where information is insufficient is a challenge for central bank statisticians. For example, it is actually required to understand the service sector trends, the behavior of flexible labor markets, and how the population's quality of life has improved.

Acknowledgment

Disclaimer: The views on the above do not necessarily reflect those of the Bank or the University.

Cross References

►Business Statistics
►Copulas in Finance
►Financial Return Distributions
►Quantitative Risk Management
►Risk Analysis
►Semi-Variance in Finance
►Statistical Modeling of Financial Markets

References and Further Reading

Bangura SAF (2005) Statistical information and the banking sector. In: Fourth meeting of the committee on development of information. United Nations Conference centre, Addis Ababa, Ethiopia

Bolton RJ, Hand DJ (2002) Statistical fraud detection: a review. Stat Sci 17(3):235–255

Current challenges and the future of central bank statistics. http://web.incisive-events.com/rma/2008/11/central-banking/rmcb_statistics-course—st.pdf

Hand DJ (2006) Statistics in banking. In: Encyclopedia of statistical sciences, 2nd edn., vol 1. Wiley, Hoboken, p 361

Hand DJ (2007) Statistical techniques for fraud detection, prevention, and evaluation, video lecture. http://videolectures.net/mmdss07handstf/

Hand DJ (2010) Fraud detection in telecommunications and banking: discussion of Becker, Volinsky, and Wilks (2010) and Sudjianto et al. Technometrics 52:34–38

International Monetary Fund (2000) Monetary and financial statistics manual. IMF, Washington

Sudjianto A, Nair S, Yuan M, Zhang A, Kern D, Cela-Díaz F (2010) Statistical methods for fighting financial crimes. Technometrics 52:5–19

Bartlett and Bartlett-Type Corrections

Gauss M. Cordeiro
Professor
Universidade Federal Rural de Pernambuco, Recife, Brazil

Bartlett Correction

The log-likelihood ratio (*LR*) statistic is one of the most commonly used statistics for inference in parametric models. Let w be the *LR* statistic for testing some composite or simple null hypothesis H_0 against an alternative hypothesis H. In regular problems, the null distribution of w is asymptotically χ^2_q, where q is the difference between the dimensions of the parameter spaces under the two hypotheses tested. However, as the samples sizes decreases, the use of such a statistic becomes less justifiable.

One way of improving the χ^2 approximation to the *LR* statistic is by multiplying w by a correction factor c known as the Bartlett correction (Lawley 1956; Hayakawa 1977; Cordeiro 1987). This idea was pioneered by Bartlett (1937) and later put into a general framework by Lawley (1956). Bartlett obtained a number of these corrections in the area of multivariate analysis in several papers published between 1938 and 1955, and these corrections became widely used for improving the large-sample χ^2 approximation to the null distribution of w.

Bartlett (1937) used the following approach to improve the χ^2 approximation to the null distribution of w. Suppose that, under the null hypothesis H_0, $E(w)$ is calculated up to order n^{-1}, say $E(w) = q + b + O(n^{-2})$, where b is a constant of order n^{-1} and n is the number of observations or some related quantity. Specifically, the Bartlett correction is determined by the relation $c = (1+b/q)$ and it represents an important tool for improving the χ^2 approximation for w. The corrected statistic $w^* = w/c$ has an expected value that comes closer to that of χ^2_q than does the expected value of w. Moreover, for continuous data, the distribution of w^* is, in general, closer to χ^2_q than is the distribution of w. Box (1949) used Bartlett's approach to investigate in detail the general expression for the moments of the statistic w in the following cases: the test of constancy of variance and covariance of k sets of p-variate samples and Wilk's test for the independence of k sets of residuals, the ith set having p_i variables. He showed, at least for these cases, that the modified statistic w^* follows a χ^2_q distribution more closely than does the unmodified statistic w. Box's results are applicable to all tests for which the Laplace transform of the test statistic can be explicitly written in terms of gamma functions and reciprocal gamma functions.

A general method to obtain Bartlett corrections for regular statistical models was developed in full generality by Lawley (1956), who gave a general formula for the correction c as function of covariant tensors. He derived expressions for the moments of certain derivatives of the log-likelihood function, and, via an exceedingly complicated calculation, obtained a general formula for the null expected value of w. Further, he showed that all cumulants of the corrected statistic w^* for testing composite hypotheses agree with those of the reference χ^2_q distribution with error of order $O(n^{-2})$ [see Hayakawa (1977) and Cordeiro (1987)]. Calculations of the Bartlett corrections via Lawley's approach are, however, notoriously cumbersome since they involve substantial effort into computing certain joint cumulants of log-likelihood derivatives. See also Eqs. 5.30–5.32 in Barndorff-Nielsen and Cox's (1994) book.

A further step on the improvement of the statistic w was taken by Hayakawa (1977), who derived an asymptotic expansion for the null distribution of w for testing a composite null hypothesis H_0 against a composite alternative hypothesis H. He showed that to order $O(n^{-1})$

$$Pr(w \le z) = F_q(z) + (24n)^{-1}[A_2 F_{q+4}(z) - (2A_2 - A_1)F_{q+2}(z) + (A_2 - A_1)F_q(z)], \quad (1)$$

where $F_s(\cdot)$ is the cumulative distribution function of a χ^2 random variable with s degrees of freedom. Here, A_1 is a function of expected values of the first four log-likelihood derivatives and of the first two derivatives of these expected values with respect to the parameters of the model and its expression holds for both simple and composite hypotheses, thus allowing for nuisance parameters. When nuisance parameters are present, A_1 can be calculated as the difference between two identical functions evaluated under the null and alternative hypotheses. The error in Equation 1 is $\mathcal{O}(n^{-2})$ and not $\mathcal{O}(n^{-3/2})$ as it is sometimes reported. However, the Bartlett correction is given by $c = 1 + (12nq)^{-1}A_1$, which differs from the one obtained from the above expansion, unless $A_2 = 0$. This points to a conflict between Hayakawa's and Lawley's results. The answer to this puzzle came with papers by Harris (1986) and Cordeiro (1987). Harris showed that A_2 should not be present in (1), whereas Cordeiro showed that the term A_2 is always equal to zero. The main contribution of Equation 1 with $A_2 = 0$ is that it provides a relatively simple demonstration that $w^* = w/c$ has a χ^2_q distribution with error $\mathcal{O}(n^{-2})$. In fact, Cordeiro (1987) demonstrated that the simple correction of the first moment of w to order $O(n^{-1})$, quite generally has the effect of eliminating the term of

this order in the asymptotic expansion of the corrected statistic w^*. This result was a starting point for numerous subsequent research efforts in the direction of establishing several expressions for Bartlett corrections in various classes of statistical models.

One difficulty encountered with the use of w^* rather than w is the fact that the required expectation may be very difficult or even impossible to compute. A general matrix formula for c was derived by Cordeiro (1993a), which has advantages for numerical and analytical purposes. For continuous data the effect of the Bartlett correction is amazingly good even for very small sample sizes. However, for discrete data, the Bartlett correction does not in general yield an improvement in the asymptotic error rate of the chi-squared approximation. Several papers have focused on deriving these corrections for special regression models by using matrix formulae for specific models, bypassing the traditional machinery of calculating these cumulants. We can always obtain these matrix formulae when the joint cumulants of log-likelihood derivatives are invariant under permutation of parameters. The matrix formulae in conjunction with computer algebra systems (Mathematica or Maple for example) and programming languages with support for matrix operations (Gauss, R and Ox) represent a computationally much simpler way of deriving Bartlett corrections in rather general classes of statistical models.

Several papers have focused on deriving matrix formulae for Bartlett corrections in general classes of regression models. Cordeiro (1983, 1987) described Bartlett's approach for univariate ▶generalized linear models. Corrected LR statistics for exponential family nonlinear models were obtained by Cordeiro and Paula (1989). They gave general matrix expressions for Bartlett corrections in these models involving an unpleasant looking quantity which may be regarded as a measure of nonlinearity of the systematic component of the model. Attfield (1991) and Cordeiro (1993b) showed how to correct LR statistics for heteroskedasticity. An algorithm for computing Bartlett corrections was given by Andrews and Stafford (1993). Cordeiro et al. (1994) proposed matrix formulae for Bartlett corrections in dispersion models. Cordeiro (1995) presented extensive simulation results on the performance of the corrected statistic w^* in generalized linear models focusing on gamma and log-linear models. For a detailed account of the applicability of Bartlett corrections, see Cribari-Neto and Cordeiro (1996).

Barndorff-Nielsen and Cox (1984) gave an indirect method for computing Bartlett corrections under rather general parametric models by establishing a simple connection between the correction term b and the norming constants of the general expression for the conditional distribution of the maximum likelihood estimator, namely $b = \left(\frac{A_0}{A}\right)^q \frac{n}{2\pi}$, where A and A_0 are the normalized constants of the general formula for the density of the maximum likelihood estimator conditional on an exact or approximate ancillary statistic when this formula is applied to the full and null models, respectively. It is usually easier to obtain the Bartlett correction for special cases using Lawley's formula than using Barndorff-Nielsen and Cox's expression, since the former involves only moments of log-likelihood derivatives whereas the latter requires exact or approximate computation of the conditional distribution of the maximum likelihood estimator. When there are many nuisance parameters, it may not be easy to obtain ancillary statistics for these parameters, and hence the evaluation of Barndorff-Nielsen and Cox's formula can be quite cumbersome. The constants A_0 and A are usually functions of the maximal ancillary statistic, although to the relevant order of magnitude, w^* is independent of the ancillary statistic selected. They have also obtained various expressions for these quantities and, in particular, an approximation which does not require integration over the sample space for the one-parameter case.

Since the statistic w is invariant under reparametrization, it is possible to obtain a large sample expansion for this statistic and its expectation in terms of invariants. McCullagh and Cox (1986) used this fact to represent the Bartlett correction as a function of invariant combinations of cumulants of the first two log-likelihood derivatives and gave it a geometric interpretation in full generality in terms of the model curvature. It is also noteworthy that McCullagh and Cox's (1986) general formula coincides with Lawley's (1956) formula. The advantage of McCullagh and Cox's formula is its geometric interpretation, whereas the main advantage of Lawley's result is that it can be more easily implemented to obtain Bartlett corrections for special models.

Bartlett-Type Correction

The problem of developing a correction similar to the Bartlett correction to other test statistics, such as the score (S) and Wald (W) statistics, was posed by Cox (1988) and addressed three years later in full generality by Cordeiro and Ferrari (1991), and by Chandra and Mukerjee (1991) and Taniguchi (1991) for certain special cases. We shall focus on Cordeiro and Ferrari's results since they are more general in the sense that they allow for nuisance parameters.

The score test has been widely used in statistics because it has an advantage over other large sample tests such as the LR and Wald tests. Thus, while these tests involve

estimation of the parameters under the alternative hypothesis, the score test requires estimation only under the null hypothesis. The *LR*, Wald and score statistics have the same chi-squared distribution asymptotically. Harris (1985) obtained an asymptotic expansion for the null distribution of the score statistic S to order $O(n^{-1})$ as

$$Pr(S \le z) = F_q(z) + (24n)^{-1}[A_3F_{q+6}(z) + (A_2 - 3A_3)F_{q+4}(z) + (3A_3 - 2A_2 + A_1)F_{q+2}(z) + (A_2 - A_1 + A_3)F_q(z)], \quad (2)$$

where A_1, A_2 and A_3 are functions of some cumulants of log-likelihood derivatives. The general expressions for these coefficients are given in Harris' paper. He showed that the first three cumulants of the score statistic are (to order n^{-1}) given by $\kappa_1 = q + A_1/(12n)$, $\kappa_2 = 2q + (A_1 + A_2)/(3n)$ and $\kappa_3 = 8q + (A_1 + 2A_2 + A_3)/n$.

Equation 2 holds for both simple and composite hypotheses. More importantly, this result implies that there exists no scalar transformation based on the score statistic which corrects all cumulants to a certain order of precision, as it is the case with the Bartlett correction to the *LR* statistic. The coefficients A's can be used to obtain corrections for models based on independent, but not necessarily identically distributed observations, thus covering a number of linear and nonlinear regression models (see ►Nonlinear Regression).

From Eq. 2, Cordeiro and Ferrari (1991) proposed a Bartlett-type correction to improve the score statistic. They defined a modified score statistic having a chi-squared distribution to order $O(n^{-1})$ under the null hypothesis

$$S^* = S\left(1 - \frac{1}{n}\sum_{j=1}^{3}\gamma_j S^{j-1}\right), \quad (3)$$

where $\gamma_1 = (A_1 - A_2 + A_3)/(12q)$, $\gamma_2 = (A_2 - 2A_3)/\{12q(q+2)\}$ and $\gamma_3 = A_3/\{12q(q+2)(q+4)\}$. They demonstrated that S^* is distributed as χ^2_q when terms of order smaller than $O(n^{-1})$ are neglected. When the A's involve unknown parameters they should be replaced by their maximum likelihood estimates under the null hypothesis but this does not affect the order of approximation of the correction. The correction factor in (3) is a function of the unmodified statistic, and hence this correction is not a "Bartlett correction" in the classical sense. Given its similarity with the Bartlett correction, however, it is termed "Bartlett-type correction."

Cordeiro and Ferrari (1991) obtained a more general result to be applied to any test statistic which converges to χ^2 which can be described as follows. Let S be a test statistic which is asymptotically distributed as χ^2_q. Chandra (1985) showed, under mild regularity conditions, that

$$Pr(S \le z) = F_q(z) + \frac{1}{n}\sum_{i=0}^{k} a_i F_{q+2i}(z) \quad (4)$$

when terms of order $O(n^{-2})$ or smaller are neglected. Equation 4 implies that the distribution function to $O(n^{-1})$ of a test statistic asymptotically distributed as chi-squared is, under certain conditions, a linear combination of chi-squareds with degrees of freedom $q, q+2, \ldots, q+2k$. The coefficients a's are linear functions of cumulants of log-likelihood derivatives for a general test statistic T. For *LR* and S_R, $k = 1$ and $k = 3$, respectively, where the a's are linear functions of the A's in Eqs. 1 and 2.

Let $\mu'_i = 2^i\Gamma(i + q/2)/\{\Gamma(q/2)\}$ be the ith moment about zero of the χ^2_q distribution, where $\Gamma(\cdot)$ is the gamma function. Cordeiro and Ferrari (1991) demonstrated that the modified test statistic

$$S^* = S\left\{1 - 2\sum_{i=1}^{k}\left(\sum_{j=1}^{k} a_j\right)(\mu'_i)^{-1} S^{i-1}\right\}$$

is distributed as χ^2_q to order $O(n^{-1})$. This is a very general result which can be used to improve many important tests in econometrics and statistics.

Building upon the result described above, Cordeiro et al. (1993) and Cribari-Neto and Ferrari (1995) obtained Bartlett-type corrections to improve score tests in generalized linear models with known and unknown dispersion parameter, respectively. Cordeiro and Ferrari (1991) demonstrated gave matrix formula for computing Bartlett-type corrections to improve score tests in a general statistical model and in exponential family nonlinear models, respectively. Finally, Cordeiro and Ferrari (1998) derived Bartlett-type corrections for chi-squared statistics based on the calculation of their moments.

About the Author

Gauss M. Cordeiro was President of the Brazilian Statistical Association (2000–2002). He has published more than 130 research articles in international journals. He was one of the founding editors of the *Brazilian Journal of Probability and Statistics* and the Editor in Chief of this journal (1995–2000). He was awarded the "Prêmio ABE" award from the Brazilian Statistical Association for his contributions to statistics.

Cross References

►Bias Correction

References and Further Reading

Andrews D, Stafford JE (1993) Tools for the symbolic computation of asymptotic expansions. J R Stat Soc B 55:613–627

Attfield CLF (1991) A Bartlett-adjustment to the likelihood ratio test for homoskedasticity in the linear model. Econ Lett 37: 119–123
Barndorff-Nielsen OE, Cox DR (1984) Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. J R Stat Soc B 46:484–495
Barndorff-Nielsen OE, Cox DR (1994) Inference and asymptotics. Chapman and Hall, London
Bartlett MS (1937) Properties of sufficiency and statistical tests. Proc R Soc A 160:268–282
Box GEP (1949) A general distribution theory for a class of likelihood criteria. Biometrika 36:317–346
Chandra TK (1985) Asymptotic expansions of perturbed chi-square variables. Sankhyā A 47:100–110
Chandra TK, Mukerjee R (1991) Bartlett-type modification for Rao's efficient score statistic. J Multivariate Anal 36:103–112
Cordeiro GM (1983) Improved likelihood ratio statistics for generalized linear models. J R Stat Soc B 45:404–413
Cordeiro GM (1987) On the corrections to the likelihood ratio statistics. Biometrika 74:265–274
Cordeiro GM (1993a) General matrix formula for computing Bartlett corrections. Stat Probab Lett 16:11–18
Cordeiro GM (1993b) Bartlett corrections and bias correction for two heteroscedastic regression models. Comm Stat Theor 22:169–188
Cordeiro GM (1995) Performance of a Bartlett-type modification for the deviance. J Stat Comput Sim 51:385–403
Cordeiro GM, Ferrari SLP (1991) A modified score statistic having chi-squared distribution to order $n-1$. Biometrika 78:573–582
Cordeiro GM, Ferrari SLP (1998) A note on Bartlett-type corrections for the first few moments of test statistics. J Stat Plan Infer 71:261–269
Cordeiro GM, Paula GA (1989) Improved likelihood ratio statistics for exponential family nonlinear models. Biometrika 76:93–100
Cordeiro GM, Ferrari SLP, Paula GA (1993) Improved score tests for generalized linear models. J R Stat Soc B 55:661–674
Cordeiro GM, Paula GA, Botter DA (1994) Improved likelihood ratio tests for dispersion models. Int Stat Rev 62:257–276
Cox DR (1988) Some aspects of conditional and asymptotic inference: a review. Sankhyā A 50:314–337
Cribari-Neto F, Cordeiro GM (1996) On Bartlett and Bartlett-type corrections. Econom Rev 15:339–367
Cribari-Neto F, Ferrari SLP (1995) Second order asymptotics for score tests in generalized linear models. Biometrika 82:426–432
Ferrari SLP, Cordeiro GM (1994) Matrix formulae for computing improved score tests. J Stat Comput Sim 49:196–206
Ferrari SLP, Cordeiro GM (1996) Corrected score tests for exponential family nonlinear models. Stat Probab Lett 26:7–12
Harris P (1985) An asymptotic expansion for the null distribution of the efficient score statistic. Biometrika 72:653–659
Harris P (1986) A note on Bartlett adjustments to likelihood ratio tests. Biometrika 73:735–737
Hayakawa T (1977) The likelihood ratio criterion and the asymptotic expansion of its distribution. Ann Inst Stat Math A 29:359–378
Lawley DN (1956) A general method for approximating to the distribution of the likelihood ratio criteria. Biometrika 71:233–244
McCullagh P, Cox DR (1986) Invariants and likelihood ratio statistics. Ann Stat 14:1419–1430
Taniguchi M (1991) Third-order asymptotic properties of a class of test statistics under a local alternative. J Multivariate Anal 37:223–238

Bartlett's Test

Hossein Arsham[1], Miodrag Lovric[2]
[1]Harry Wright Distinguished Research Professor of Statistics and Management Science
University of Baltimore, Baltimore, MD, USA
[2]Professor
University of Kragujevac, Kragujevac, Serbia

Bartlett's test (introduced in 1937 by Maurice Barlett (1910–2002)) is an inferential procedure used to assess the equality of variance in different populations (*not in samples* as sometimes can be found, since there is no point in testing whether the samples have equal variances – we can always easily calculate and compare their values). Some common statistical methods assume that variances of the populations from which different samples are drawn are equal. Bartlett's test assesses this assumption. It tests the null hypothesis that the population variances are equal.

All statistical procedures have underlying assumptions. In some cases, violation of these assumptions will not change substantive research conclusions. In other cases, violation of assumptions is critical to meaningful research. Establishing that one's data meet the assumptions of the procedure one is using is an expected component of all quantitatively based journal articles, theses, and dissertations. The following are the two general areas where Bartlett's test is applicable:

1. In regression analysis and time series analyses, the residuals should have a constant variance (i.e., homoskedastic condition). One may check this condition by dividing the residuals data into two or more groups, then using the Bartlett's test.
2. Another area of application is the *F*-test in ANOVA that requires the assumption that each underlying population has the same variance (i.e., homogeneity condition).

Bartlett's test is derived from the likelihood ratio test under the normal distribution. Therefore, it is dependent on meeting the assumption of normality.

Bartlett's test of homogeneity of variance is based on a chi-square statistic with $(k-1)$ degrees of freedom, where k is the number of categories (or groups) in the independent variable. In other words, Bartlett's test is used to test if k populations have equal variances.

We wish to test the null hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

against the alternative that at least two population variances are not equal. The following briefly explains the procedure employed by Bartlett's test.

To investigate the significance of the differences between the variances of k normally distributed populations, independent samples are drawn from each of the populations. Let S_j^2 denote the variance of a sample of n_j items from the jth population ($j = 1, \ldots, k$).

The test statistic has the following expression:

$$B = \frac{(N-k)\ln\left(\dfrac{\sum_{i=1}^{k}(n_i-1)s_i^2}{N-k}\right) - \sum_{i=1}^{k}(n_i-1)\ln\left(s_i^2\right)}{1+\dfrac{1}{3(k-1)}\left[\left(\sum_{i=1}^{k}\dfrac{1}{n_i-1}\right)-\dfrac{1}{N-k}\right]}$$

where N corresponds to the sum of all sample sizes. It is asymptotically distributed as a χ^2 distribution with $(k-1)$ degrees of freedom. The null hypothesis of equal population variances is rejected if test statistics is larger than the critical value. One may also use online tools to perform this test, under condition that each sample contains at least five observations.

Bartlett's test is known to be powerful if the underlying populations are normal. According to some recent results based on simulation (Legendre and Borcard), Bartlett's test and Box's test are the best overall methods to test the homogeneity of variances. With non-normal data, Bartlett's and Box's tests can be used if the samples are fairly large.

It was shown that Bartlett's test is unbiased (Pitman 1939) and consistent (Brown 1939). One of its major drawbacks, however, is that it is extremely sensitive (that is non-robust) to the departures from normality (Box 1953). Since in reality heterogeneity and non-normality frequently simultaneously occur, when the null hypothesis is rejected, we cannot know whether this is due to unequal population variances or non-normality, or both. This test is so sensitive to departures from normality that Box commented that it well may be used as a good test of the population normality. Box also remarked that using Bartlett's test to check whether it is appropriate to apply ANOVA would be rather like "putting a rowing boat to sea to find out whether conditions are sufficiently calm for an ocean liner to leave a port" (Box 1953, p 333).

Finally, for the unequal sample-sizes case, several approaches have been advised for finding exact critical values (see, for example, Chao and Glaser (1978) and Manoukian et al. (1986)).

About the Author

Dr. Hossein Arsham is The Wright Distinguished Research Professor in Decision Science and Statistics. He has created one of the most comprehensive sites on statistical analysis on the Internet. Currently, he is Associate editor for *International Journal of Ecological Economics and Statistics*, *International Journal of Statistics and Systems*, and *Journal of Interdisciplinary Mathematics*. He is a Fellow of The Royal Statistical Society (1984), The Operational Research Society (1985), the Institute of Combinatorics and Its Applications (1992), and The World Innovation Foundation: Scientific Discovery (1998). Professor Arsham has received numerous awards for his teaching.

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Tests for Homogeneity of Variance

References and Further Reading

Arsham H (2010) The *P*-values for Chi-square distribution. http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/pvalues.htm#rchdist

Arsham H (2010) Homogeneity of Multi-variances: the Bartlett's test. http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/BartletTest.htm

Box GEP (1953) Non-normality and tests on variances. Biometrika 40:318–335

Brown GW (1939) On the power of the L1 test for equality of several variances. Ann Math Stat 10:119–128

Chao M-T, Glaser RE (1978) The exact distribution of Bartlett's test statistic for homogeneity of variances with unequal sample sizes. JASA 73(352):422–426

Legendre P, Borcard D. Statistical comparison of univariate tests of homogeneity of variances (Submitted to the Journal of Statistical Computation and Simulation)

Lemeshko B, Yu EM (2004) Bartlett test in measurements with probability laws different from normal. Meas Tech 47(10): 960–968

Levin I (1999) Relating statistics and experimental design. Sage, Thousand Oaks

Manoukian EB, Maurais J, Ouimet R (1986) Exact critical values of Bartlett's test of homogeneity of variances for unequal sample sizes for two populations and power of the test. Metrika 33(1):275–289

Pitman EJG (1939) Tests of hypotheses concerning location and scale parameters. Biometrika 31:200–215

Ramsey PH (1994) Testing variances in psychological and educational research. J Educ Stat 19(1):23–42

Shoemaker LH (2003) Fixing the *F* test for equal variances. Am Stat 57(2):105–114

Wu J, Wong A (2003) A note on determining the *p*-value of Bartlett's test of homogeneity of variances. Commun Stat Theory 32(1):91–101

Bayes' Theorem

Joseph B. Kadane
Leonard J. Savage University Professor of Statistics, Emeritus
Carnegie Mellon University, Pittsburg, PA, USA

The conditional probability $P\{A \mid B\}$ of event A given event B, is commonly defined as follows:

$$P\{A \mid B\} = \frac{P\{AB\}}{P\{B\}} \tag{1}$$

provided $P\{B\} > 0$. Alternatively, (1) can be reexpressed as

$$P\{AB\} = P\{A \mid B\}P\{B\}. \tag{2}$$

The left-hand side of (2) is symmetric in A and B, while the right-hand side does not appear to be. Therefore we have

$$P\{A \mid B\}P\{B\} = P\{B \mid A\}P\{A\}, \tag{3}$$

or

$$P\{A \mid B\} = \frac{P\{B \mid A\}P\{A\}}{P\{B\}}, \tag{4}$$

which is the first form of Bayes' Theorem.

Note that the event B can be reexpressed as

$$B = AB \cup \overline{A}B. \tag{5}$$

Because AB and $\overline{A}B$ are disjoint, we have

$$P\{B\} = P\{AB\} + P\{\overline{A}B\} = P\{B|A\}P\{A\} + P\{B|\overline{A}\}P\{\overline{A}\}. \tag{6}$$

Substituting (6) into (5) yields the second form of Bayes' Theorem:

$$P\{A \mid B\} = \frac{P\{B \mid A\}P\{A\}}{P\{B \mid A\}P\{A\} + P\{B \mid \overline{A}\}P\{\overline{A}\}}. \tag{7}$$

Finally, let $A_1, A_2, \ldots, A_k$ be disjoint sets whose union is the whole space S. Then, generalizing (5),

$$B = \cup_{i=1}^{k} A_i B \tag{8}$$

and the sets $A_i B$ are disjoint. Then

$$P\{B\} = \sum_{i=1}^{k} P\{A_i B\} = \sum_{i=1}^{k} P\{B \mid A_i\}P(A_i\}. \tag{9}$$

Then generalizing (7), we have

$$P\{A_j \mid B\} = \frac{P\{B \mid A_j\}P\{A_j\}}{\sum_{i=1}^{k} P\{B \mid A_i\}P_i\{A_i\}}, \tag{10}$$

the third form of Bayes' Theorem.

As an example, let A be the event that a particular person has HIV, and suppose $P\{A\} = 0.001$. Suppose there is a test for the presence of HIV in a patient. If B is the event of a positive result from the test, suppose

$$P\{B \mid A\} = 0.95 \text{ and } P\{B \mid \overline{A}\} = 0.05,$$

which means that if the person has HIV, the test is 95% likely to find it, and if the person does not have HIV, the test is 5% likely to report a positive result.

Then

$$\begin{aligned} P\{A \mid B\} &= \frac{(0.95)(0.001)}{(0.95)(0.001) + (0.05)(0.999)} \\ &= \frac{0.00095}{0.00095 + 0.04995} \\ &= \frac{0.00095}{0.0509} \\ &= 0.0187 \end{aligned}$$

Thus a positive test result, on a respectable test, does not imply a high probability that the patient actually has HIV.

Bayes' Theorem has an aura of being controversial. Since everything above is a consequence of the laws of probability and the definition of conditional probability, the correctness of the formulas above is conceded by all the various schools of interpretation of probability.

The Bayesian school uses Bayes' Theorem as a way to understand the extent to which knowing that the event B (test result in the example) has changed the probability of the event A, resulting in a map from $P\{A\}$ to $P\{A \mid B\}$. In the example, $P\{A\} = 0.001$ is the probability the person has HIV given that person's risk factors. $P\{A \mid B\} = 0.0187$ is the updated probability that takes into account the positive test result B.

The controversy comes from the fact that some statisticians (Fisher 1959) take the view that the only events that have probabilities are those that can be regarded as elements of an infinite (or large) sequence of independent and identically distributed events. Then the probability of A is taken to be the relative frequency of the event A in this sequence. Thus the controversy has to do not with the legitimacy of Bayes' Theorem itself, but rather with the applications of it.

About the Author

Joseph B. Kadane received his Ph.D. in Statistics, Stanford University, 1966. He was Head, Department of Statistics, Carnegie Mellon University (1972–1981). He was Associate editor of *Journal of the American Statistical Association* (1968–1973) and Applications and Coordinating Editor (1983–1985), Associate Editor of *Annals of Statistics*

(1974–1976) and *Journal of Business and Economic Statistics* (1987–1998). He was also Elected Member, Board of Directors, International Society for Bayesian Analysis (1996–2000). Professor Kadane was named the Pittsburgh Statistician of the Year (1980), chosen by the Pittsburgh Chapter of the American Statistical Association. He was Co-winner, Frank Wilcoxon Award for the Best Applied Paper in *Technometrics*, 1993. He has (co-)authored over 250 papers. (23 in Economics and Econometrics; 17 in Sociology and Demography; 5 in Political Science; 19 in Operations Research and Computer Science; 74 in Mathematics and Statistical Theory; 40 in Law; 32 in Management Science; 27 in Psychology, Medicine, and Biology and 20 in Environmental and Physical Sciences.)

Cross References

►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Conditional Expectation and Probability
►Inversion of Bayes' Formula for Events
►Philosophy of Probability
►Probability, History of
►Statistical Inference: An Overview
►Statistics, History of
►Statistics: An Overview

References and Further Reading

Fisher R (1959) Statistical methods and scientific inference, 2nd edn. Oliver and Boyd, Edinburgh and London

Bayesian Analysis or Evidence Based Statistics?

D. A. S. Fraser
Professor
University of Toronto, Toronto, ON, Canada

Introduction

The original Bayes proposal leads to likelihood and confidence for many simple examples. More generally it gives approximate confidence but to achieve exact confidence reliability it needs refinement of the argument and needs more than just the usual minimum of the likelihood function from observed data. A general Bayes approach provides a flexible and fruitful methodology that has blossomed in contrast to the widely-based long-standing frequentist testing with focus on the 5% level. We examine some key events in the evolution of the Bayes approach promoted as an alternative to the present likelihood based frequentist analysis of data with model, the evidence-based approach of central statistics. And we are led to focus on the bane of Bayes: parameter curvature.

Bayes 1763

Bayes (1763) examined the Binomial model $f(y;\theta) = \binom{n}{\theta}$ $\theta^y(1-\theta)^{n-y}$ and proposed the flat prior $\pi(\theta) = 1$ on $[0,1]$. Then with data y^0 he used a lemma from probability calculus to derive the posterior $\pi(\theta|y^0) = c\theta^{y^0}(1-\theta)^{n-y^0}$ on $[0,1]$. And then for an interval say $(\theta,1)$ he calculated the integral of the posterior,

$$s(\theta) = \int_\theta^1 \theta^{y^0}(1-\theta)^{n-y^0}d\theta \Big/ \int_0^1 \theta^{y^0}(1-\theta)^{n-y^0}d\theta$$

and referred to it as probability that the parameter belonged to the interval $(\theta,1)$. Many endorsed the proposed calculation and many disputed it.

As part of his presentation he used an analogy. A ball was rolled on a level table, perhaps an available billiard table, and was viewed as having equal probability of stopping in any equal sized area. The table was then divided conceptually by a North-South line through the position where the ball stopped, with area θ to the West and $(1-\theta)$ to the East. The ball was then rolled n further times and the number y^0 of time that it stopped left of the line observed. In the analogy itself, the posterior probability calculation given data seems entirely appropriate.

The Economist 2000

In an article entitled "In praise of Bayes," the Economist (2000) speaks of an "increasingly popular approach to statistics (but) not everyone is persuaded of its validity." The article mentions many areas of recent application of the Bayesian approach, and cites "the essence …is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence." The indicated areas of application are wide spread and there is emphasis on attaining definitive answers. And this is set in full contrast to "traditional ways of presenting results" indicated to be the mid-twentieth-century decision theoretic view of accepting a null view 95–5% on some departure scale. The article does offer some caution for "when used indiscriminently" in the form of a quotation from Larry Wasserman that it can become "more a religion than a science."

The mathematical rule cited as the essence of the Bayesian approach is a very broad expansion from Bayes original proposal where a statistical model $f(y;\theta)$ evaluated at an observed data value y^0 giving $f(y^0;\theta)$ is combined with a constant mathematical prior $\pi(\theta) = 1$ and

treated as a conditional density. The force of the rule is that with new model-data information the new likelihood would be folded with the old. But this is of course standard practice in statistics: use the up-to-date likelihood, and possibly refine such a procedure with meta-analysis. What is different is that the Bayesian method essentially overlooks evidence beyond the observed likelihood function and does so on principle.

Validity or Analogy

Bayes considered a uniform prior and a Binomial (n,p) model, and used analogy to justify combining them by a standard lemma from probability calculus. For the analogy involving balls on a billiard table, the calculations seem entirely proper and appropriate. The more generally interpreted Bayes approach has a statistical model $f(y;\theta)$ with data y^0 coupled with a mathematical prior $\pi(\theta)$ representing symmetries or other properties of the model or context. Analogies can be great for explaining an argument but not to be the argument itself: there is no billiard table equivalent in the typical binomial or more general context.

There is an explicit time line: There is a context with a true value θ_* for the parameter θ; there is an investigation $f(y;\theta)$ yielding an observed y^0 from the true value model $f(y;\theta_*)$; and possibilities for θ are then to be assessed. Thus in order: θ_* is realized but unknown; y^0 is observed; then assess θ. The values θ_* and y^0 are realized and are in the past. And the issue is what can be said about θ given the model $f(y;\theta)$ and data y^0.

If θ is understood in fact to come from an objective source $\pi(\theta)$ with realized value θ_*; then the time line is longer. Accordingly: $\pi(\theta)$ produces θ_*; $f(y;\theta_*)$ produces y^0; and the issue is to assess θ. In this situation $\pi(\theta)$ is properly an objective prior. And an option is of course is to examine and present the composite model $\pi(\theta)f(y;\theta)$ with observed y^0. But an even more compelling option is to examine and present $\pi(\theta)$ and to separately examine and present $f(y;\theta)$ with y^0.

Now consider the model $f(y;\theta)$ with data y^0; and the mathematical prior $\pi(\theta)$ as proposed by Bayes. The lemma from the probability calculus has two probability inputs say $\pi(x)$ and $f(y|x)$ and it has one probability output $\pi(x|y^0)$; the output records the behavior of x that is associated with the observed value $y = y^0$. For the Bayes case $\pi(x)$ would be $\pi(\theta)$ and $\pi(x|y^0)$ would be $\pi(\theta|y^0)$. Is the lemma applicable or relevant in the Bayes case? In the Bayes case there is just one probability input $f(y;\theta)$; and the other nominal input is $\pi(\theta)$, a mathematical object that refers to symmetry or patterns in the model and has no probability status whatsoever. Thus the assumptions of the lemma do not hold, and consequently the output of the lemma is ... by analogy ... not by derivation. The usage of the lemma in the proposed argument is not proper and can be viewed as fraudulent logic.

The standard frequentist would refer to $f(y^0;\theta)$ as likelihood $L(\theta;y^0) = L^0(\theta)$. An exploration with weighted likelihood $\pi(\theta)L^0(\theta)$ can be a very natural, obvious and sensible procedure ... for just that, for exploring possibilities for θ. But for obtaining probabilities, perhaps a pipe dream!

Likelihood and Confidence

Bayes' (1763) original approach suggested a density $c\pi(\theta)f(y^0;\theta)$ as a description of an unknown θ in the presence of observed data y^0. As such it records likelihood $L^0(\theta)$ or weighted likelihood. And this was long before the formal introduction (Fisher 1922) of likelihood. Both viewpoints record the same formal information concerning the parameter; the differences are in the color or flavor associated with the particular argument; and with properties attributed to the output.

Bayes (1763) also offered a distribution as a summary of information concerning the parameter θ; the distribution had density $c\pi(\theta)f(y^0;\theta) = c\pi(\theta)L^0(\theta)$. The majority of models at that time had least-squares location structure and for such models the posterior $\pi(\theta)L^0(\theta)$ using a natural prior just reproduces what is now called confidence (Fisher 1930, 1935).

It thus seems appropriate to acknowledge that Bayes introduced the primary concepts of likelihood and confidence long before Fisher (1922, 1930) and long before the refinement offered by Neyman (1937). For likelihood he offered the extra flexibility of the weight function but for confidence he did not have the statistical refinement that later provided the logical extension to non-location models; this latter can be viewed as a matter of reasonable fine tuning of the argument, of intellectual evolution, and of the familiar iterative processes of science.

Laplace and Venn

Laplace (1812) seems to have fully endorsed the proposals arising from Bayes (1763). And Venn (1886) seems equally to have rejected them. Certainly asserting the conclusions of a theorem or lemma when one of the premises does not hold is unacceptable from a mathematical or logical viewpoint. Nonetheless the results were impressive and filled a substantial need, but indeed with downstream risks. And it does have, as is now becoming apparent, the support of approximate confidence (Fraser 2010). At present Bayes and confidence lead a coexistence, perhaps an uneasy unstable coexistence!

Priors and Priors

The original Bayes prior was a flat prior $\pi(\theta) = 1$ for a probability θ that then in sequence becomes the parameter in a Binomial (n, θ) model; the resulting posterior is $\pi(\theta)L^0(\theta)$, which focally uses the observed likelihood from the Binomial context. Some aspects of invariance were invoked to support the particular choice. The possible plausible extensions are immense.

For a location model $f\{y - \beta(\theta)\}$ the natural prior would be $\pi(\theta)d\theta = d\beta(\theta) = \beta'(\theta)d\theta$. Thus for $f(y - X\beta)$ we would have $\pi(\beta)d\beta = dX\beta = cd\beta$, giving a flat prior for the regression coefficients. Motivation would come by noting that $y - \beta(\theta)$ has a fixed θ-free distribution.

Extensions are possible by seeking approximate θ-free distributions. This was initiated by Jeffreys (1939) and then fine-tuned to acknowledge various types of parameters (Jeffreys 1946). These extensions use expected information $i(\theta) = E\{-\ell_{\theta\theta}(\theta; y); \theta\}$ in the model to calibrate the scale for θ; here $-\ell_{\theta\theta}(\theta)$ is the negative second derivative of likelihood and the initial Jeffreys prior is $\pi(\theta)d\theta = |i(\theta)|^{1/2}d\theta$, and it is parameterization invariant. For the regression model, where $y = X\beta + \sigma z$ with $N(0,1)$ error, the Jeffreys (1939) prior is $\pi(\theta)d\theta = d\beta d\sigma/\sigma^{r+1}$ where r is the column rank of X. The second or modified Jeffreys (1946) is $\pi(\theta)d\theta = d\beta d\sigma/\sigma$ and gives generally more acceptable results, often in agreement with confidence.

The approximate approach can be modified (Fraser et al. 2010) to work more closely with the location invariance indicated by the initial Bayes (1763) approach. In many regular problems continuity within the model leads to a relationship $d\hat{\theta} = W(\theta)d\theta$ where $W(\theta)$ is a $p \times p$ matrix; the $d\hat{\theta}$ refers to an increment at the data y^0 and the $d\theta$ refers to an increment $d\theta$ at θ. This immediately indicates the prior $\pi(\theta)d\theta = |W(\theta)|d\theta$ based on simple extension of the translation invariance $dy = \beta'(\theta)d\theta$ for the model $f(y - \beta(\theta))d\theta$; and it widely agrees with preferred priors in many problems. But the parameter must not have curvature: the bane of Bayes!

The approximate approach can also be modified to make use of an asymptotic result that to second order the statistical model can be treated as an exponential model (Reid and Fraser 2010; Fraser et al. 2010). This uses continuity to obtain a nominal reparameterization $\varphi(\theta)$ that yields second and third order inference by acting as if the model were just $g(s; \theta) = \exp\{\ell(\theta) + \varphi(\theta)s\}h(s)$ with data $s^0 = 0$. This allows information to be calculated within the approximating model using the information function $j_{\varphi\varphi}(\theta; s) = -\ell_{\varphi\varphi}\{\theta(\varphi)\}$; this draws attention to marginalization and to curvature effects that are not usually apparent in the search for default priors (Fraser et al. 2010).

The preceding can also be viewed as a somewhat natural evolution from the original Bayes proposal with some reference to location invariance. The evolution has been assisted by the fact that many posterior distributions have appealing and sensible properties. It is our view here that these sensible properties are precisely the approximate confidence properties that have become evident quite separately. In any case the priors just described can all be classified as default priors, priors that one might choose to use as a default without strong arguments for something different.

The Bayesian approach is committed to using a weight function applied to an observed likelihood and thus to formally omitting other properties of the model. Within this approach the default priors are widely called objective priors. But the term objective means objective reference, and this as a property is specifically absent here; there is a strong flavor of deception. Thus using the term objective for default priors seems highly inappropriate, but could be viewed as just a seeking for a wider area of application. The author was present at the Bayesian convention when the term was being chosen and did not register an objection, being perhaps somewhat of an outsider, not a good defense! We will however explicitly refer to them as default priors, and keep the term objective for contexts where the prior does have an explicit reference in context.

A difficulty with the use of default priors is that a posterior probability obtained by marginalization from a full posterior distribution may not be equal the posterior probability calculated directly from the appropriate marginal model; this was given prominence by Dawid et al. (1973) and applies equally to confidence distributions and other attempts to present model-data information as a distribution for the parameter. The complication in any of these cases derives from parameter curvature: for some discussion see Fraser et al. (2010) and Fraser and Sun (2010).

The wealth of possibilities available from a weight-function combined with likelihood is well documented in the development of the Bayesian methods as just described. Its success can amply be supported as "approximate confidence" but derived by a route that is typically much easier. Approximate confidence provides full support for the acceptable, often meritorious behavior of Bayes posterior probabilities. We address later whether there can be anything beyond approximate confidence in support of the Bayesian approach.

Another approach, somewhat different from the original Bayes way of obtaining a weight function is derived from Kullback-Leibler distance on measure spaces (Bernardo 1971): this chooses a prior to maximize the

statistical distance from prior to posterior. Modifications of this distance approach have been developed to obtain specialized priors for different component parameters of interest, often parameters that have a statistical curvature.

The richness available from using just a likelihood function is clearly evident to Bayesians if not to frequentists; but is not widely acknowledged. Much of recent likelihood theory divides on whether or not to use more than the observed likelihood, specifically sampling properties that are associated with likelihood characteristics but are not widely or extensively available. In many ways central statistics has ignored the extra in going beyond likelihood, and indeed has ignored the wealth available from just likelihood alone.

Meanwhile those committed to using just the weighted likelihoods, those associated with the development of the Bayes approach as we have just described, have aggressively sought to use the weighted likelihood approach as a general approach to updating information and to producing decisions. Central to this direction is the subjective approach with a major initiative coming from Savage (1972). This takes a prior to represent the views, the understanding, the personal probabilities concerning the true value of the parameter; these might come from highly personal thoughts, from detailed elicitation from knowledgeable people, from gut feelings as one approaches a game at a casino; and they can have the benefit of intuition or the merits of a seasoned gambler, with or without insider information. But who should use it? Certainly the chronic gambler will. But from the statistical perspective here there is nothing of substance to say that such prior "information" $\pi(\theta)$ should be combined with likelihood. With due respect it can be presented as $\pi(\theta)$ alongside a presentation of the evidence-based well calculated confidence. If a user would like to combine them, it would certainly be plausible for him to do so but it would not be an imperative despite Bayesian persuasion. Certainly place them both to be seen and available. In wide generality combining them is not a necessary statistical step, although sometimes expedient.

Lindley and Territory

Fisher's (1930, 1935) proposal for confidence with effective support from Neyman (1937) offered strong alternatives to a prominent sympathy for the Bayesian approach. Then Jeffreys (1939, 1946) with great prominence in geophysics provided reinforcement for the use of the Bayesian approach in the physical sciences. Meanwhile the confidence approach gained strength both in mathematics departments and in scientific applications. Both approaches lead from model and data to a distribution for the parameter, but the results were often in conflict. Both sides clearly felt threatened, and each side in a practical sense had territory to defend.

Lindley (1958) focused on the very basic case, a scalar parameter and a scalar variable, say with distribution function $F(y;\theta)$. The Bayesian answer with prior $\pi(\theta)$ is given by the posterior distribution $c\pi(\theta)F_y(y;\theta)d\theta$ where the subscript y denotes differentiation with respect to the argument y thus giving the density or likelihood function. By contrast the Fisher (1930, 1935) approach gives the confidence distribution $|F_\theta(y;\theta)|d\theta$. Lindley examined when these would be equal and solved for $\pi(\theta)$:

$$\pi(\theta) = c\frac{F_{;\theta}(y;\theta)}{F_y(y;\theta)} = c\frac{\partial}{\partial\theta}y(u;\theta);$$

the right hand expression records the derivative of the quantile function for fixed p-value $u = f(y;\theta)$ as pursued in Fraser et al. (2010). The equation is actually a differential equation that asserts that the model must be a location model, the form of model actually found in section "►Likelihood and Confidence" to have good Bayesian answers. In Fraser et al. (ibid) the equation is used to determine the data dependent priors that give posterior probabilities having objective validation.

Lindley was concerned that the confidence approach did not follow the primal Bayesian concept that a probability statement concerning a parameter should be updated by multiplication by new likelihood and his criticism had a profound effect suggesting that the confidence distribution approach was defective. We now know that the defect is the attempt to use a distribution as the summary, either by Bayes or by confidence. And if there were to be a lesser of two evils then calling confidence probability and calling Bayes approximate confidence would be safer.

Another view might be that this was just a territorial dispute as to who had the rights to provide a distributional description of the parameter in the model data context. But the social conflict aspects were not in evidence. Rather there was a wide spread perception that giving a distribution of confidence was wrong. Neyman (1937) of course had provided a route around. But nonetheless, the judgment stuck: a confidence distribution was wrong and a Bayesian analysis was all right. Of course, in Dawid et al. (1973), there is a clear message that neither approach can handle vector parameters without special fine-tuning. Clearly Lindley had focused on a substantive issue but the arguments invoked had not quite attained the point of acknowledging that an effective prior must in general be data dependent; for some current discussion see Fraser et al. (2010).

Bayesian Analysis and Imperatives

Bayesian (1763) analysis has been around for a long time, but alternative views perhaps now identified as frequentist are perhaps older although somewhat less formalized. These approaches have cross dialogued and often been in open conflict. Each has made various appeals to holding the truth. And they have actively sought territorial advantage. In particular Fisher's 1930 initial steps towards confidence were directly to provide an alternative to inverse probability, the name at the time attached to the Bayesian approach. So it is not surprising that there would be a very focal reverse criticism (Lindley 1958) of the confidence approach.

Those favoring the Bayesian approach have frequently felt they were underdogs, often for example having their articles rejected by journals for just being Bayesian. It thus seems rather natural that the Bayesian supporters would seek to broaden their methodology and their community. The subjective approach as strongly initiated by Savage (1954) has led to a powerful following in an area where prior probabilities are extended to include personal feelings, elicited feelings, and betting view points. Certainly such extensions are a guide for gambling and much more. But there is nothing of substance to assert that they should be ... the imperative ... used for the analysis. The prior subjective assessment and the objective evidence-based assessment can be placed side by side for anyone to see and to use as deemed appropriate. And the Bayes combination of these can also be presented for anyone to use if so inclined. Perhaps the Bayesian expansion was ill advised to promote the imperative: that the proper analysis was that of the Bayes paradigm.

What is perhaps even more dangerous is the widely promoted hierarchical model where each parameter is given a prior distribution, and then parameters in the prior distributions are themselves given priors, perhaps then multilevel. Often an impressive edifice that seems only equaled by the lack of evidence for the various introduced elements and the impressive resort to McMC. The resort to multilevel Bayes modeling would seemingly be best viewed as one of expediency, to extend the base of Bayes without supporting evidence.

And then of course there are model data situations where the true parameter has come from a source with a known frequency distribution. In such cases the obvious name for the prior would be objective prior. But assembled Bayesians as mentioned earlier have adopted that name for the opposite situation, where there is in fact no objective reference, and the prior is purely a mathematical construct. But what about the multitude of cases where there is an identified source for the true parameter value? These can arise widely when the entity being examined has been obtained by sampling from some identified subpopulation; or they can arise by genetics or by Mendel or perhaps by updated genetic laws. Or much more. In reality this is just a modeling issue: what aspect of the context, of the immediate environment, or the more extended environment should be modeled. It is a modeling issue. It is perhaps only natural that Bayesian promotion should seek to subsume wider and wider contexts as part of the evolution of the approach. Especially when traditional statistics has been widely immersed in technical criteria connected with some global optimization or with decision rules to reject at some 5% level or accept at some 19/20 level, even when it was becoming abundantly apparent that these rules for scientific publication have serious defects.

But if there is an objective source $\pi(\theta)$ for a true value in a model-data context, there is nothing that says it should be folded into a combined model for analysis. The prior source $\pi(\theta)$ can be set in parallel with the more directly evidence-based analysis of the model-data combination. And of course even the combined model-data-prior analysis presented. But again there is no substantive precept that says the combined analysis is the statistical inference. Such a step would be purely an assertion of extended territory for Bayesian analysis.

Curvature: The Bane of Bayes

Contours of a parameter can have obvious curvature. A simple example can throw light on the effects of such curvature.

Consider (y_1, y_2) with a Normal $\{(\theta_1, \theta_2); I\}$ distribution on the plane. With data $\left(y_1^0, y_2^0\right)$ the basic original Bayes approach would say that (θ_1, θ_2) was Normal $\left\{\left(y_1^0, y_2^0\right); I\right\}$. First we examine an obviously linear parameter $\psi = \theta_1$ and assess say the value $\psi = 0$ on the basis of data, say $\left(y_1^0, y_2^0\right) = (0, 0)$.

In an obvious way y_1 measures ψ and has the Normal$(\psi; 1)$ distribution. Accordingly the p-value for ψ from the observed data is

$$p(\psi) = \Phi\left\{\left(y_1^0 - \psi\right)/1\right\} = \Phi(-\psi).$$

And for assessing the value $\psi = 0$ we have $p(0) = 50\%$.

Now consider the Bayesian assessment of the value ψ. the marginal posterior distribution of ψ is $N\left(y_1^0, 1\right)$ and the corresponding posterior survivor value is

$$s(\psi) = 1 - \Phi\left(\left(\psi - y_1^0\right)/1\right) = 1 - \Phi(\psi)$$

at the observed data. In particular for assessing $\psi = 0$ we would have $s(0) = 50\%$. The Bayesian and frequentist values are equal for the special $\psi = 0$ and also for general ψ.

Now consider a clearly curved parameter, the distance ψ on the parameter space from the point $(-1, 0)$ to the parameter value (θ_1, θ_2),

$$\psi = \left\{(\theta_1 + 1)^2 + \theta_2^2\right\}^{1/2}.$$

An obvious way to measure this parameter is by using the distance r from the point $(-1, 0)$; thus $r = \left\{(y_1 + 1)^2 + y_2^2\right\}^{1/2}$. The distribution of r^2 is noncentral Chi-square with two degrees of freedom and noncentrality $\delta^2 = \psi^2$. The indicated p-value for assessing ψ is then

$$p(\psi) = H_2(r^2; \psi^2)$$

where H_2 is the noncentral Chi-square distribution function with two degrees of freedem and noncentrality $\delta^2 = \psi^2$. This is readily available in *R*. In particular for assessing $\psi = 1$ we would have

$$p(1) = H_2(1; 1) = 26.7\%$$

which is substantially less than 50%.

Now consider the Bayesian assessment of the curved parameter ψ. The posterior distribution of ψ^2 from the observed data is noncentral Chi-square with two degrees of freedom and noncentrality $\delta^2 = 1$. If follows that the posterior survivor value for assessing $\psi = 1$ is

$$s(1) = 1 - H_2(1; 1) = 73.3\%$$

which is substantially larger than 50%.

For this simple example we have seen that the p-value and the survivor value are equal for a linear parameter. This happens generally for linear parameters (Fraser and Reid 2002). And with the introduction of a curvature change to the parameter, the Bayesian and frequentist values go in opposite directions. This happens widely with curved parameters: as a parameter contour is changed from linear to curved, the Bayesian survivor changes in the opposite direction from the frequentist. Thus the Bayesian can be viewed as correcting negativity, that is making an adjustment opposite to what is appropriate in a context. For some recent discussion see Fraser (2010). The example above suggests that curvature is precisely the reason that Bayes fails to correctly assess parameters.

Consider y with a Normal $\{\theta, \sigma^2(\theta)\}$ distribution where the variance $\sigma^2(\theta)$ depends weakly on the mean θ. Precise p-values are available for assessing θ:

$$p(\theta) = \Phi\{(y - \theta)/\sigma(\theta)\}$$

with a clear frequency interpretations. The confidence inversion is well established (Fisher 1930, 1935). The Bayesian inversion does not seem to have an obvious prior that targets the parameter θ.

How does one assess the merits of a proposed distribution for a parameter? The use of two-sided intervals provides a slippery slope. Strange tradeoffs can be made between two interval bounds; see for example Fraser et al. (2004) on statistics for discovering new particles in High Energy Physics. A more direct approach is to examine a particular quantile of a proposed distribution, say the β-th quantile $\hat{\theta}_\beta$ which has posterior probability β to the right and $(1 - \beta)$ to the left. One can certainly simulate or have an oracle and determine what proportion of the time the true value is larger than the particular quantile being considered; and determine whether the true proportion bears a sensible relation to the alleged value β. This has been addressed at length in Fraser (2010).

In particular for the Normal$\{\theta, \sigma^2(\theta)\}$ example there is no determination of a prior that will give the third order accuracy that is available from the confidence approach unless the prior is directly specific to the observed data value. This result holds in wide generality: the use of a default or Bayesian prior cannot lead to the third order accuracy readily available from the evidence-based procedures of frequentist inference. And parameter curvature is the number one culprit.

Why Bayes?

Linear approximations are widely used throughout statistics, mathematics, physics, the sciences generally, and much more. They provide a local replica of something that might be intangible otherwise and when used iteratively can provide exploration of something unknown otherwise. There is substantial evidence that the Bayes procedure provides an excellent first order approximation for the analysis of a statistical model. There are also ample warnings that global acceptance of Bayes results can be extremely hazardous. Use but be cautious!

The Bayes calculus asserts that the posterior results are probabilities. And the name itself is assertive. The Bayesian supporters have also been vocal, asserting that confidence results do not have the status of probabilities calculated by the Bayes paradigm; some indication of this is implicit in Lindley (1958); and further indication is found in the active broadening of the application area for Bayesian analysis. From an evidence-based approach it

is clear that the direct use of the likelihood function provides substantial information, first order information. And higher order results are available with the careful choice of prior. But beyond that, the Bayes procedure comes up short, unless the priors become data dependent and the calculations are carefully targeted using an evidence-based formulation.

Thus linear approximations can be hugely useful but they can carry substantial risks. The assertion of probability status is directly contradicted by reality! And no indications seem available that a Bayesian calculation could yield more than just approximate confidence. The promotional assertiveness that accompanies current Bayes development is misleading and misleading to the extent of being fraudulent.

Of course there can be contexts where there is an objective prior $\pi(\theta)$ that records how the true value was generated. The Bayes paradigm can be applied but it is inappropriate; the direct approach is a matter of modeling, of what aspect of the context is appropriate to include. From this viewpoint the indicated methodology predates Bayes; it is just probability analysis. Even then it allows the prior information and the evidence based information to be presented separately, thus of course allowing the end user to combine if needed or wanted.

There is no imperative that says the prior and the evidence-based should be combined. It is an option. And it is an option with risks!

About the Author

Donald A.S. Fraser was born in Toronto in 1925. He obtained his Ph.D. in 1949 under the supervision of Samuel Wilks at Princeton University. He was the first Chair of the Department of Statistics, University of Toronto (1977–1983). Among many awards, Professor Fraser was the first recipient of the Gold Medal of the Statistical Society of Canada, inaugurated in 1985. He received the R.A. Fisher Award and Prize, American Statistical Association (1990), and Gold Medal, Islamic Statistical Society (2000). He was the first statistician to be named a Fellow of the Royal Society of Canada (1967). He has supervised 55 Ph.D. students. Professor Fraser has (co-)authored over 250 papers and authored five books, including *Nonparametric Methods in Statistics* (Wiley 1957), *The Structure of Inference* (Wiley 1968) and *Inference and Linear Models* (McGraw Hill 1979). In 2002, he was awarded a degree of Doctor of Science, *honoris causa*, by the University of Toronto. In 2010 (April 30–May 1), the third in a series of conferences, titled Data Analysis and Statistical Foundations III, was held to honor the accomplishments of Professor Fraser.

Cross References

►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Foundations of Probability
►Frequentist Hypothesis Testing: A Defense
►Likelihood
►Model Selection
►Philosophical Foundations of Statistics
►Prior Bayes: Rubin's View of Statistics
►Probability Theory: An Outline
►Probability, History of
►Significance Tests: A Critique
►Statistical Inference
►Statistical Inference: An Overview
►Statistics: An Overview
►Statistics: Nelder's view

References and Further Reading

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Phil Trans R Soc London 53:370–418; 54:296–325. Reprinted in (1958) Biometrika 45:293–315

Bernardo JM (1971) Reference posterior distributions for Bayesian inference (with discussion). J R Stat Soc B 41: 113–147

Dawid AP, Stone M, Zidek JV (1973) Marginalization paradoxes in Bayesian and structural inference. J R Stat Soc B 35: 189–233

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Phil Trans R Soc London A 222:309–368

Fisher RA (1930) Inverse probability. Proc Camb Phil Soc 26: 528–535

Fisher RA (1935) The fiducial argument in statistical inference. Ann Eugenic B 6:391–398

Fraser DAS (2010) Is Bayes posterior just quick and dirty confidence? Stat Sci, in review

Fraser DAS, Reid N (2002) Strong matching of frequentist and Bayesian parametric inference. J Stat Plann Infer 103: 263–285

Fraser DAS, Reid N, Wong A (2004) Inference for bounded parameters. Physics Rev D 69:033002

Fraser DAS, Reid N, Marras E, Yi GY (2010) Default prior for Bayesian and frequentist inference. J R Stat Soc B to appear.

Fraser DAS, Sun Y (2010) Some corrections for Bayes curvature. Pak J Statist, 25:351–370

Jeffreys H (1939) Theory of probability, 3rd edn. Oxford University Press, Oxford

Jeffreys H (1946) An invariant form by the prior probabilities in estimation problem. Proc Roy Soc A 186:453–461

Laplace PS (1812) Théorie analytique des probabilités. Courcier, Paris

Lindley DV (1958) Fiducial distribution and Bayes theorem. J R Stat Soc B 20:102–107

Neyman J (1937) Outline of a theory of statistical estimation based on the classical theory of probability. Phil T R Soc A 237: 333–380

Reid N, Fraser DAS (2010) Mean likelihood and higher order inference. Biometrica, 97:159–170
Savage LJ (1972) The Foundations of Statistics. Wiley, New York

Bayesian Approach of the Unit Root Test

Hocine Fellag[1], Lynda Atil[2]
[1]Professor in Statistics
University of Tizi-Ouzou, Tizi-Ouzou, Algeria
[2]University of Tizi-Ouzou, Tizi-Ouzou, Algeria

In statistical inference, two approaches of hypothesis testing can be used. The first one, called classical approach, is based on calculating probabilities of the data in hand under certain conditions (which are encompassed by the null and alternative hypotheses). The second one, named Bayesian approach, looks at probabilities of competing conditions (which are the hypotheses being compared) given the data in hand. This approach integrates prior probabilities associated with competing conditions into the assessment of which condition is the most likely explanation for the data in hand. Also, it allows to evaluate the likelihood of competing conditions by evaluating the change in the odds associated with these conditions, a change produced by assessing the data in hand. If the odds change sufficiently when the data are examined, then the scientist may alter his opinion about which competing condition is the most likely. One of the main papers published on this topic is the paper of Schotman and Van Dijk (1991a), where a very important problem in Bayesian analysis is tackled, namely, the Bayesian approach for unit root testing. Several economists, Dejong and Whiteman (1991), Koop (1992), and in particular, Sims (1988), and Sims and Uhlig (1991), have advocated forcefully for Bayesian alternatives over the more traditional classical approach such as the ADF tests (Dickey and Fuller 1981) in unit root testing. Despite the apparent advantages of the Bayesian approach over the classical approach in unit root testing, a relatively small number of studies have used the Bayesian approach. The reasons may be that the Bayesian approach requires a likelihood function and the use of prior information.

The modeling objective of the Bayesian approach is not to reject a hypothesis based on a predetermined level of significance, but to determine how probable a hypothesis is relative to other competing hypotheses. Schotman and Van Dijk (1991a) propose a posterior odds analysis of the hypothesis of a unit root in real exchange rates because nominal and real exchange rates behave almost like random walks (see ▶Random Walk).

Now, suppose that we have a sample of T consecutive observations on a time series y_t generated by

$$y_t = \rho y_{t-1} + \mu_t \tag{1}$$

where

1. y_0 is a known constant.
2. μ_t are identically and independently (i.i.d) normally distributed with mean zero and unknown variance σ^2.
3. $\rho \in S \cup \{1\}$, $S = \{\rho / -1 < a \leq \rho < 1\}$.

The econometric analysis aims at discriminating between a stationary model (here defined as $a \leq \rho < 1$) and the nonstationary model with $\rho = 1$. The lower bound a in assumption (3) largely determines the specification of the prior for ρ. Recall that the principal Bayesian tool to compare a sharp null hypothesis with a composite alternative hypothesis is the posterior odds ratio, which is defined as

$$K_1 = K_0 \frac{\int_0^\infty p(\sigma) L(y \mid \rho = 1, \sigma, y_0) d\sigma}{\int_S \int_0^\infty p(\sigma) p(\rho) L(y \mid \rho, \sigma, y_0) d\sigma d\rho} = \frac{p(\rho = 1 \mid y)}{p(\rho \in S \mid y)} \tag{2}$$

K_0 and K_1 are the prior odds and the posterior odds in favor of the hypothesis $\rho = 1$, respectively. $p(\rho)$ represents the prior density of $\rho \in S$, $p(\sigma)$ the prior density of σ.

$L(y \mid .)$ is the likelihood function of the observed data $y = (y_1 \ldots y_T)'$ and $Y = (y_0, y')'$ is all observed data.

The Bayes factor is defined as the ratio of the marginal posterior density of ρ under the null hypothesis $\rho = 1$ over a weighted average of the marginal posterior under the alternative using the prior density of ρ as a weight function. Then, one can notice that the posterior odds K_1 is equal to the prior odds K_0 times the Bayes factor. The prior odds express the special weight given to the null hypothesis, the point $\rho = 1$ is given the discrete prior probability $\vartheta = K_0/(1 + K_0)$. From the posterior odds, one can compute the posterior probability of the null hypothesis as $K_1/(1 + K_1)$.

For the complete specification of the marginal prior of ρ and σ, we assume that

$$Pr(\rho = 1) = \nu \tag{3}$$

$$p(\rho \mid \rho \in S) = \frac{1}{1 - a} \tag{4}$$

$$p(\sigma) \propto \frac{1}{\sigma} \tag{5}$$

The prior of ρ is uniform on S but has a discrete probability ϑ that $\rho = 1$. The likelihood function for the vector of T observations y is

$$L(y \mid \rho, \sigma, y_0) = (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}\mu'\mu\right\} \quad (6)$$

where $\mu = y - y_{-1}\rho$, and $y_{-1} = (y_0, \ldots, y_{T-1})'$

Having computed the relevant integrals in (2), the posterior odds ratio becomes

$$K_1 = \frac{C_T^{-1}}{(T-1)^{1/2}} \frac{\nu}{1-\nu} \left(\frac{\sigma_0^2}{\hat{\sigma}^2}\right)^{-T/2} \left(\frac{1-a}{s_{\hat{\rho}}}\right) \left[F\left(\frac{1-\hat{\rho}}{s_{\hat{\rho}}}\right) - F\left(\frac{a-\hat{\rho}}{s_{\hat{\rho}}}\right)\right]^{-1} \quad (7)$$

where

$$\sigma_0^2 = \frac{1}{T-1}(y-y_{-1})'(y-y_{-1}) \quad \hat{\sigma}^2 = \frac{1}{T-1}\left(y'y - \frac{(y'_{-1}y)^2}{(y'_{-1}y_{-1})}\right)$$

$$s_{\hat{\rho}}^2 = \hat{\sigma}^2 \left(y'_{-1}y_{-1}\right)^{-1} \quad \hat{\rho} = \frac{y'_{-1}y}{y'_{-1}y_{-1}} \quad C_T = \frac{\Gamma((T-1)/2)\Gamma(1/2)}{\Gamma(T/2)} \quad (8)$$

The empirical lower bound a^* is given by $a^* = \hat{\rho} + s_{\hat{\rho}} F^{-1}(\alpha F(-\hat{\tau}))$. $F(.)$ is the cumulative t-distribution with $(T-1)$ degrees of freedom and $\hat{\tau} = \frac{\hat{\rho}-1}{s_{\hat{\rho}}}$ is the Dickey–Fuller test statistic. The unit root model is preferred if $K_1 > 1$ or $P(\rho = 1|y, y_0) \geq 0.50$, thus treating the null and the alternative in a symmetric way.

After fixing numerical values for ϑ and α, the posterior odds is just a function of the data like any other test statistic. Due to a specific way that the lower bound has been constructed, the posterior odds are directly related to the ►Dickey–Fuller test. Setting the prior odds equal to one and for large T, Schotman & Van Dijk approximate $F(.)$ to the cumulative normal distribution. The posterior odds become a function of the Dickey–Fuller statistic $\hat{\tau}$.

$$\ln K_1 = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\hat{\tau}^2 + \ln\left(\frac{-\hat{\tau} - F^{-1}(\alpha F(-\hat{\tau}))}{F(-\hat{\tau})}\right) \quad (9)$$

Since the posterior odds is a function of the Dickey–Fuller test statistic, its sampling properties correspond exactly to those of the Dickey–Fuller test. In literature, there is a great attention to the nature of suitable noninformative priors for the autoregressive coefficients. For example, Sims (1988) and Sims and Uhlig (1991) advocate the use of flat priors. Phillips (1991a) proved that flat priors bias the inference toward stationary models, and suggests to use Jeffrey priors derived from conditional likelihood functions. Also, Uhlig (1994a) determines the Jeffreys priors for an AR(1) process from the exact likelihoods and justifies the use of flat priors in some specific cases only. Uhlig (1994b) summarizes the Bayesian contribution to the unit root problem and discusses the sensitivity of the tails of the predictive densities on the prior treatment of explosive roots. Schotman and Van Dijk (1991a) stress the sensitivity of the posterior odds to the size of the stationary region and suggest restricting the later's size. Berger and Yang (1994) consider a reference prior approach for the AR(1) model. It is particularly interesting to note that Marriott and Newbold (1998) criticized the use of priors, such as the uniform or the Jeffreys prior, for the autoregressive coefficients in this context and advocate the use of sharp informative prior distributions. However, for the simple problem of testing for a unit root in a first-order autoregressive process, they find that the prior distribution for the autoregressive coefficient has a substantial impact on the posterior odds, so that, a very sharp beta prior performs extremely well when the generating process is stationary autoregressive, but the uniform prior is preferable when the true model is nonstationary. Marriott and Newbold (1998) explore the use of the ►beta distribution as a prior specification. They have explained how Bayesian calculations can be carried out, noting the importance of the analyst, giving careful thought to the question of what might be an appropriate prior.

Conclusions

Generally, authors agree with the idea that the Bayesian approach offers an alternative and a more useful way than the classical approach in empirical modeling. In unit root testing, Sims (1988), Sims and Uhlig (1991), and Koop (1992, 1994) have advocated the Bayesian approach over the classical ADF tests. In some papers, various opinions were expressed saying that the Bayesian solution is clear and simple when the classical approach is logically unsound. In a series of empirical applications, using a Bayesian approach, Dejong and Whiteman (1989, 1991a, b) obtained results rejecting the presence of a unit root in various economic series, which contradicted those of Nelson and Plosser (1982). Ahking (2004) studied the power of Koop's "objective" unit root test. In particular, he was interested in whether or not it provides a better alternative to the classical ADF unit root test, and whether or not the use of "objective" priors are appropriate. However, there is no evidence to suggest that the "objective" Bayesian test is better than the classical ADF tests in unit root tests. Moreover, the "objective" priors do not seem to be appropriate since they tend to produce results that are biased in favor of the trend stationary hypothesis. Thus, unfortunately, while there is a need for more objective analysis of Bayesian time series, Koop's "objective" Bayesian test does not appear to move us closer to that goal. So, one can ask, what is the best use in unit root tests, classical or Bayesian procedure? Intuitively,

it is very hazardous to discriminate between two competing economic theories on the basis of a univariate model. However, unit root tests may be helpful when they are used in a more complete modeling strategy as a protection against gross errors, as well as misspecification tests.

About the Author

Professor Fellag was Vice chancellor of Mouloud Mammeri University of Tizi-Ouzou, Algeria, in charge of international relations from 2005 until 2010. He is also a honorary professor of Rey Juan Carlos University of Madrid (Spain) in 2010. He is Director of the national Algerian doctoral school of statistics since 2007. Dr. Fellag has supervised more than 20 doctoral students in statistics and has published around 20 papers in international journals.

Cross References

►Bayesian Statistics
►Box–Jenkins Time Series Models
►Dickey-Fuller Tests
►Seasonal Integration and Cointegration in Economic Time Series
►Seasonality

References and Further Reading

Ahking FW (2004) The power of the 'Objective' Bayesian unit root test. Working Paper 2004–2014, University of Connecticut
Berger JO, Yang RY (1994) Noninformative priors and Bayesian testing for the AR(1) model. Economet Theor 10:461–482
Dejong DN, Whiteman CH (1989) Trends and cycles as unobserved components in US real GNP: a Bayesian perspective. J Am Stat Assoc Pap proc 63–70
Dejong DN, Whiteman CH (1991a) The temporal stability of dividends and stock prices: Evidence from the likelihood function. Am Econ Rev 81:600–617
Dejong DN, Whiteman CH (1991b) Reconsidering'Trends and random walk in macroeconomic time series. J Monetary Econ 28:221–254
Dickey DA, Fuller WA (1981) Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49: 1057–1072
Koop G (1992) Objective' Bayesian unit root tests. J Appl Econ 7: 65–82
Koop G (1994) Recent progress in applied Bayesian econometrics. J Econ Surv 8:1–34
Marriott J, Newbold P (1998) Bayesian comparison of ARIMA and stationary ARMA models. Int Stat Rev 66(3):323–336
Nelson CR, Plosser CI (1982) Trends and random walk in macroeconomic time series: Some evidence and implication. J Monetary Econ 10:139–162
Phillips PCB (1991a) To criticize the critics: An objective Bayesian analysis of stochastics trends. J Appl Econ 6:333–364
Phillips PCB (1991b) Bayesian routes and unit roots: De rebus priorbus semper est disputandum. J Appl Econ 6:435–474
Schotman P, Van Dijk HK (1991a) A Bayesian analysis of the unit root in real exchange rates. J Econ 49:195–238
Schotman P, Van Dijk HK (1991b) On Bayesian routes to unit roots. J Appl Econ 6:461–464
Sims CA (1988) Bayesian skepticism on unit root econometrics. J Econ Dyn Control 12:463–474
Sims CA, Uhlig H (1991) Understanding unit rooters: A helicopter tour. Econometrica 59:1591–1599
Uhlig H (1994a) On Jeffreys' prior when using the exact likelihood function. Econom Theor 10:633–644
Uhlig H (1994b) What macroeconomists should know about unit roots: A Bayesian perspective. Econom Theor 10:645–671

Bayesian Nonparametric Statistics

Jaeyong Lee
Associate Professor
Seoul National University, Seoul, Korea

Bayesian nonparametric statistics covers Bayesian analysis of nonparametric models, statistical models whose parameter space is not finite-dimensional, and allows more flexible modeling than the parameteric alternatives. Bayesian analysis of a statistical model consists of three ingredients: prior, model (or likelihood), and posterior. The prior $\pi(\theta)$ is a probability measure on the parameter space Θ that reflects the analyst's knowledge about the unknown parameter θ before he or she observes the data. The model describes the random mechanism by which the observation X is generated given the parameter θ, that is, $X|\theta \sim f(x|\theta)$. The posterior is the conditional distribution of θ given X, $\pi(\theta|X)$, which reflects the analyst's knowledge about θ after X is observed. The prior and posterior of nonparametric models are, thus, probability measures of infinite dimensional parameter spaces. Examples of such parameter spaces include the space of all probability measures on the real line, the space of all probability density functions on the real line, the space of all smooth functions, and many more.

The most widely used nonparametric prior is the Dirichlet process (Ferguson 1973), a probability measure on the space of all probability measures. Let $\mathcal{X}$ be a measurable space with a σ-field $\mathcal{A}$ and α be a nonnull finite measure on $\mathcal{X}$. A random probability measure P on $\mathcal{X}$ is said to follow a Dirichlet process with parameter α, denoted by $DP(\alpha)$, if for every measurable partition $(A_1, A_2, \ldots, A_k)$ of $\mathcal{X}$, $(P(A_1), P(A_2), \ldots, P(A_k))$ follows

(finite-dimensional) Dirichlet distribution with parameter $(\alpha(A_1), \alpha(A_2), \ldots, \alpha(A_k))$.

The Dirichlet process has many important properties that have theoretical and practical consequences. The class of Dirichlet processes is a conjugate prior class in the following sense. Suppose

$$P \sim DP(\alpha), \quad X_1, X_2, \ldots | P \sim P. \tag{1}$$

Then, the posterior P given $X_1, X_2, \ldots, X_n$ also follows Dirichlet process, that is, $P|X_1, \ldots, X_n \sim DP\left(\alpha + \sum_{i=1}^{n} \delta_{X_i}\right)$, where δ_x is a degenerate probability measure at x. Under model (1), the sequence $X_1, X_2, \ldots$ form, marginally, the Pölya urn sequence, that is, $X_1 \sim \alpha/\alpha(\mathcal{X})$ and for $n \geq 1$,

$$X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha + \sum_{i=1}^{n} \delta_{X_i}}{\alpha(\mathcal{X}) + n}. \tag{2}$$

This property is the key to the posterior computation of the mixtures of Dirichlet process models. Sethuraman (1994) derived an alternative constructive definition of the Dirichlet process. Suppose $Y_1, Y_2, \ldots$ are iid $\alpha/\alpha(\mathcal{X})$, $\theta_1, \theta_2 \ldots$ are iid $Beta(1, \alpha(\mathcal{X}))$ independently of $Y_i's$. Let

$$p_1 = \theta_1, \quad p_n = \theta_n \prod_{i=1}^{n-1} (1 - \theta_i), n \geq 2. \tag{3}$$

Then, $P = \sum_{i=1}^{\infty} p_i \delta_{Y_i} \sim DP(\alpha)$. This property shows P is discrete with probability 1 if $P \sim DP(\alpha)$, which was initially thought to be a shortcoming of the Dirichlet process.

To remedy the discreteness of the Dirichlet process, the mixtures of Dirichlet process model has been proposed, which turns out to be the most successful model with Dirichlet process. It has the following structure:

$$P \sim DP(\alpha)$$
$$X_1, X_2, \ldots, X_n | P \overset{iid}{\sim} \int h(x|\theta) dP(\theta),$$

where $h(x|\theta)$ is a probability density function with parameter θ, or equivalently,

$$P \sim DP(\alpha)$$
$$\theta_1, \theta_2, \ldots, \theta_n | P \overset{iid}{\sim} P$$
$$X_i | \theta_i \overset{ind}{\sim} h(x|\theta_i), \quad 1 \leq i \leq n.$$

The mixtures of Dirichlet processes have been used for different problems, for example, Bayesian density estimation, cluster analysis (see ►Cluster Analysis: An Introduction), etc. Especially, mixtures of Dirichlet processes have been successful in cluster analysis. Since the random probability measure P is discrete, the random sample θs from P naturally have ties, and clusters of Xs are based on ties in θs.

Since the Dirichlet process was proposed, many other nonparametric priors have appeared. Among them are neutral to the right process (Doksum 1974), Gaussian process (O'Hagan 1978), beta process (Hjort 1990; Lo 1993), Polya tree process (Lavine 1992), and species sampling model (Pitman 1996), all of which are priors for distribution except beta process and gaussian process. The beta process and gaussian process are priors for cumulative hazard function and regression function, respectively.

The posterior computation with nonparametric models can be complicated. There are algorithms specialized for specific priors (e.g., MacEachern 1994; MacEachern and Müller 1998; Lee 2007). Recently, DPpackage (Jara 2007), an R package that automates the posterior computation of some nonparametric models, has been built and lessens computational effort of practical users of nonparametric models.

Unlike parametric models, whose posteriors behave asymptotically optimal in the frequentist sense, nonparametric posteriors can behave suboptimally. Diaconis and Freedman (1986) have shown that even an innocent-looking prior may generate inconsistent posterior. This observation spurs the research effort to obtain conditions for posterior consistency, posterior convergence rate, and ►asymptotic normality of the posterior (Bernstein-von Mises theorem). There is now a large body of literature on the asymptotic issue of ►Bayesian statistics (e.g., Ghosal et al. 2000; Shen and Wasserman 2001; Kim and Lee 2001; Freedman 1999).

Acknowledgment

I thank Professor Fernando Quintana for his helpful discussion and comments on the paper. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (20090075171).

Cross References

►Bayesian Statistics
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Posterior Consistency in Bayesian Nonparametrics
►Statistical Inference: An Overview

References and Further Reading

Diaconis P, Freedman D (1986) On inconsistent Bayes estimates of location. Ann Stat 14(1):68–87

Doksum K (1974) Tailfree and neutral random probabilities and their posterior distributions. Ann Probab 2:183–201
Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230
Freedman D (1999) On the Bernstein-von Mises theorem with infinitedimensional parameters. Ann Stat 27(4):1119–1140
Ghosal S, Ghosh JK, van der Vaart AadW (2000) Convergence rates of posterior distributions. Ann Stat 28(2):500–531
Hjort NL (1990) Nonparametric Bayes estimators based on beta processes in models for life history data. Ann Stat 18(3): 1259–1294
Jara A (2007) Applied Bayesian non- and semi-parametric inference using dppackage. Rnews 7(3):17–26
Kim Y, Lee J (2001) On posterior consistency of survival models. Ann Stat 29(3):666–686
Lavine M (1992) Some aspects of Pólya tree distributions for statistical modelling. Ann Stat 20(3):1222–1235
Lee J (2007) Sampling methods of neutral to the right processes. J Comput Graph Stat 16(3):656–671
Lo AY (1993) A Bayesian bootstrap for censored data. Ann Stat 21(1):100–123
MacEachern SN (1994) Estimating normal means with a conjugate style Dirichlet process prior. Comm Stat Sim Comput 23(3): 727–741
MacEachern SN, Müller P (1998) Estimating mixture of dirichlet process models. J Comput Graph Stat 7(2):223–338
O'Hagan A (1978) Curve fitting and optimal design for prediction. J R Stat Soc Series B 40(1):1–42
Pitman J (1996) Some developments of the Blackwell-MacQueen urn scheme. In: Statistics, probability and game theory, vol 30 of IMS Lecture Notes Monogr. Ser., pp 245–267. Inst. Math. Statist., Hayward, CA
Sethuraman J (1994) A constructive definition of Dirichlet priors. Stat Sinica, 4(2):639–650
Shen X, Wasserman L (2001) Rates of convergence of posterior distributions. Ann Stat 29(3):687–714

Bayesian P-Values

Jayanta K. Ghosh[1], Mohan Delampady[2]
[1]Professor of Statistics
Purdue University, West Lafayette, IN, USA
[2]Professor
Indian Statistical Institute, Bangalore, India

While Bayesians do not like classical ▶P-values and prefer measuring evidence in data through posterior probabilities of parameters or models, some problems like testing or exploration of goodness of fit of a single given model have led to the introduction of P-values. We confine ourselves to this particular context of goodness of fit in the brief discussion of Bayesian P-values. Most of this material is taken from Ghosh, Delampady and Samanta (2006) and Ghosh, Purkayastha and Samanta (2005).

Suppose that we have a single model M that specifies a density $f(x|\theta), \theta \in \Theta$ for the observable X and the Bayesian has a prior $\pi(\theta)$. The Bayesian wishes to examine how well the model M fits the data x_{obs} on the basis of a statistic $T(X)$ which measures the goodness of fit of data and model. Of course, T is also chosen by the Bayesian even though it is not part of the usual paradigm for Bayesian inference.

Let

$$m_\pi(x) = \int_\Theta f(x|\theta)\pi(\theta)\, d\theta$$

be the prior predictive density. Box (1980) defines

$$p = \int_{\{T(x)>T(x_{obs})\}} m_\pi(x)\, dx$$

as a prior predictive P-value. Of course, this depends on the prior π of the Bayesian.

To reduce the dependence on π and also to make it possible to use an improper non-informative prior π with proper posterior $\pi(\theta|x_{obs})$, Gutman (1967), Rubin (1984), Meng (1996) and Gelman et al. (1996) propose a posterior predictive P-value p^* defined as follows.

Let

$$m^*(x|x_{obs}) = \int_\Theta f(x|\theta)\pi(\theta|x_{obs})\, d\theta,$$

$$p^* = \int_{\{T(x)>T(x_{obs})\}} m^*(x|x_{obs})\, dx.$$

However, as pointed out by Bayarri and Berger (1998), p^* involves a double use of the data in both the integrand and the tail area of the integrand defining p^*. In order to remove this undesirable feature, Bayarri and Berger (1998) introduce what they call a conditional predictive P-value which is defined as follows.

Identify a statistic $U(X)$ which is not a function of $T(X)$ and let $m(t|u)$ be the conditional predictive density of T given U. Then the conditional predictive P-value is

$$p_c = \int_{\{T(x)>T(x_{obs})\}} m(t|u_{obs})\, dt.$$

Bayarri and Berger (1998) also define a partial posterior predictive P-value in the same vein. This alternative P-value does not require the choice of the auxiliary statistic U and to that extent is less arbitrary.

When the model is simple, meaning that the distribution of the random observable has no unknown parameters, then under this model, all these P-values reduce to the tail area probability under this distribution (which is exactly the classical P-value). In this case, the P-value treated as random (i.e., $p = p(X)$) has the $U(0,1)$ distribution, a desirable property as far as its interpretation is concerned. This property is desirable even when the model is composite and the nuisance parameters are eliminated

in some way. However, this is not always likely and what can be expected is that they be asymptotically uniformly distributed. Robins et al. (2000) argue that in the absence of the $U(0,1)$ property, one should at least require that these P-values not suffer from serious conservative or anti-conservative behavior; $p(X)$ is defined to be conservative (anti-conservative) when $P(p(X) < u)$ is smaller (larger) than u for all $u < 1/2$, with P denoting the distribution of X under (any θ from) the model.

Assuming certain regularity conditions on the model as well as the prior, Robins et al. (2000) prove that the conditional predictive and partial predictive P-values are asymptotically uniformly distributed, whereas the posterior predictive P-value is often conservative.

The different P-values are illustrated below with two examples from Bayarri and Berger (1998, 2000).

Example 1 Suppose $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from some distribution, and we want to check if it is $N(\mu_0, \sigma^2)$, σ^2 unknown and μ_0 is a specified value for its mean. The natural discrepancy statistic is $T(\mathbf{X}) = (\bar{X} - \mu_0)$. Consider the usual non-informative prior $\pi(\sigma^2) \propto 1/\sigma^2$. Then the prior predictive P-value doesn't exist since this prior is improper. Let $s^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$. We obtain

$$\pi(\sigma^2|\mathbf{x}_{obs}) \propto (\sigma^2)^{-n/2-1} \exp\left(-n\left(s_{obs}^2 + t_{obs}^2\right)/(2\sigma^2)\right),$$

and thus the posterior predictive density of T is

$$m(t|\mathbf{x}_{obs}) \propto \left(1 + \frac{1}{n}\frac{nt^2}{s_{obs}^2 + t_{obs}^2}\right)^{-(n+1)/2}.$$

therefore the posterior predictive P-value is

$$p^* = 2\left\{1 - \mathcal{T}_n\left(\frac{\sqrt{n}t_{obs}}{\sqrt{s_{obs}^2 + t_{obs}^2}}\right)\right\},$$

with $\mathcal{T}_\nu$ denoting the c.d.f. of Student's t_ν.

Choose $U(X) \equiv s^2$, and note that $nU|\sigma^2 \sim \sigma^2\chi_{n-1}^2$. Therefore,

$$\pi(\sigma^2|U = s^2) \propto (\sigma^2)^{-(n-1)/2-1} \exp(-ns^2/(2\sigma^2)),$$

and consequently, the conditional predictive density of T given $U = s_{obs}^2$ is

$$\begin{aligned} m(t|s_{obs}^2) &= \int_0^\infty f_T(t|\sigma^2)\pi\left(\sigma^2|s_{obs}^2\right) d\sigma^2 \\ &\propto \int_0^\infty (\sigma^2)^{-1/2} \exp\left(-\frac{nt^2}{2\sigma^2}\right)(\sigma^2)^{-(n-1)/2} \\ &\quad \exp\left(-\frac{n}{2\sigma^2}s_{obs}^2\right)\frac{d\sigma^2}{\sigma^2} \\ &\propto \int_0^\infty \exp\left(-nv\left\{s_{obs}^2 + t^2\right\}\right) v^{n/2}\frac{dv}{v} \\ &\propto \left(1 + \frac{1}{n-1}\frac{(n-1)t^2}{s_{obs}^2}\right)^{-n/2}. \end{aligned}$$

This implies, under the conditional predictive distribution,

$$\sqrt{n-1}\frac{T}{s_{obs}} \sim t_{n-1}.$$

The conditional predictive P-value, therefore, is

$$p_c = 2\left\{1 - \mathcal{T}_{n-1}\left(\frac{\sqrt{n-1}t_{obs}}{s_{obs}}\right)\right\}.$$

Bayarri and Berger (1998, 2000) show that in this example the partial predictive P-value and the conditional predictive P-value coincide.

Listed in Table 1 are values of p^*, $t_{obs} = \sqrt{n-1}(\bar{x}_{obs} - \mu_0)/s_{obs}$ and $t^* = \sqrt{n/(n-1)}t_{obs}/\sqrt{1 + t_{obs}^2/(n-1)}$ corresponding to different values of n when p_c is fixed at 0.01, 0.05 and 0.10, respectively.

Example 2 Suppose, as in the previous example, X_1, $X_2, \ldots, X_n$ is a random sample from some population. The target model now is Exponential(λ). Consider $\pi(\lambda) \propto 1/\lambda$. Let $T = X_{(1)}$ be the model checking statistic and let $S = \sum_{i=1}^n X_i$. The posterior density of λ given $S = s_{obs}$ is proportional to $\lambda^{n-1}\exp(-s_{obs}\lambda)$ so that the posterior predictive density of T given $S = s_{obs}$ is

$$n^2 s_{obs}^n (nt + s_{obs})^{-(n+1)}.$$

Bayesian P-Values. Table 1 P-values: p^* versus p_c for the normal model

		n						
		2	3	5	10	20	50	100
$p_c = 0.01$	t_{obs}	63.66	9.92	4.60	3.25	2.86	2.68	2.63
$p_c = 0.01$	t^*	1.414	1.715	2.051	2.324	2.454	2.528	2.552
$p_c = 0.01$	p^*	0.293	0.185	0.096	0.042	0.023	0.015	0.012
$p_c = 0.05$	t_{obs}	12.71	4.30	2.78	2.26	2.09	2.01	1.98
$p_c = 0.05$	t^*	1.410	1.645	1.814	1.904	1.936	1.951	1.956
$p_c = 0.05$	p^*	0.293	0.198	0.129	0.086	0.067	0.057	0.053
$p_c = 0.10$	t_{obs}	6.31	2.92	2.13	1.83	1.73	1.68	1.66
$p_c = 0.10$	t^*	1.397	1.559	1.631	1.649	1.649	1.647	1.646
$p_c = 0.10$	p^*	0.297	0.217	0.164	0.130	0.115	0.106	0.103

Bayesian P-Values. Table 2 P-values: p^* versus p_c for the exponential model

		n				
		2	3	5	10	20
$t^* = 1.00$	p_c	0	0	0	0	0
$t^* = 1.00$	p^*	0.25	0.125	0.031	0	0
$t^* = 0.75$	p_c	0.25	0.063	0.004	0	0
$t^* = 0.75$	p^*	0.327	0.187	0.061	0.004	0
$t^* = 0.60$	p_c	0.4	0.16	0.026	0	0
$t^* = 0.60$	p^*	0.391	0.244	0.095	0.009	0
$t^* = 0.50$	p_c	0.5	0.25	0.063	0.002	0
$t^* = 0.50$	p^*	0.444	0.296	0.132	0.017	0
$t^* = 0.40$	p_c	0.6	0.36	0.13	0.01	0
$t^* = 0.40$	p^*	0.51	0.364	0.186	0.035	0.001
$t^* = 0.25$	p_c	0.75	0.563	0.316	0.075	0.004
$t^* = 0.25$	p^*	0.64	0.512	0.328	0.107	0.012

This yields the posterior predictive P-value of

$$p^* = \left(1 + \frac{nt_{obs}}{s_{obs}}\right)^{-n}.$$

A direct calculation gives

$$f(\mathbf{x}|T = t, \lambda) \propto \lambda^{n-1} \exp(-\lambda(s - nt)),$$

from which the partial posterior density for λ is seen to be

$$\frac{\lambda^{n-2} \exp(-\lambda(s_{obs} - nt_{obs}))}{\Gamma(n-1)(s_{obs} - nt_{obs})^{-(n-1)}};$$

This is called partial posterior because it is obtained from the partial likelihood, $f(\mathbf{x}_{obs}|T = t_{obs}, \lambda)$ instead of the full likelihood $f(\mathbf{x}_{obs}|\lambda)$. Then the partial posterior predictive density of T is obtained as

$$\frac{n(n-1)(s_{obs} - nt_{obs})^{n-1}}{(nt + s_{obs} - nt_{obs})^n}.$$

This leads to the following expression for the partial posterior predictive P-value:

$$\left(1 - \frac{nt_{obs}}{s_{obs}}\right)^{n-1}.$$

Bayarri and Berger (2000) go on to show that this coincides with the conditional predictive P-value p_c upon taking the conditioning (auxiliary) statistic to be the MLE of λ from $f(\mathbf{x}|T = t, \lambda)$ (which is given by $(n-1)/(S-nT)$.

In Table 2, values of p^* and p_c corresponding to some values of $t^* = nt_{obs}/s_{obs}$ and n are displayed.

About the Authors

Dr Jayanta Kumar Ghosh (born May 22, 1937) is Professor of Statistics, Department of Statistics, Purdue University. He graduated from the University of Calcutta. He has spent a substantial part of his career in the Indian Statistical Institute, Calcutta. He served as the Director of ISI and later held the Jawaharlal Nehru Professorship. "Professor Jayanta Kumar Ghosh, or J. K. Ghosh, as he is commonly known, has been a prominent contributor to the discipline of statistics for five decades. The spectrum of his contributions encompasses sequential analysis, the foundations of statistics, Finite populations, Edgeworth expansions, second order efficiency, Bartlett corrections, noninformative, and especially matching priors, semiparametric inference, posterior limit theorems, Bayesian non-parametrics, model selection, Bayesian hypothesis testing and high dimensional data analysis, as well as some applied work in reliability theory, statistical quality control, modeling hydrocarbon discoveries, geological mapping and DNA Fingerprinting." [B. Clarke and S. Ghosal (2008). J. K. Ghosh's contribution to statistics: A brief outline, In: *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, IMS Collections, Institute of Mathematical Statistics, Beachwood, Ohio (3), p. 1.] He has (co-)authored over 150 papers and several books, including: *Higher Order Asymptotics* (published jointly by Institute of Mathematical Statistics and American Statistical Association, 1994), *Bayesian Nonparametrics* (with R.V. Ramamoorthi, Springer, 2003), and *An Introduction to Bayesian Analysis, Theory and Methods* (with M. Delampady and T. Samanta, Springer, 2006). He has supervised more than 30 PhD students. He was awarded the Shanti Swarup Bhatnagar Award for Mathematical Science (1981), Mahalanobis Gold Medal of Indian Science Congress Association (1998) and P.V. Sukhatme Prize for Statistics (2000). Professor Jayanta Ghosh is a Fellow of the Indian Academy of Sciences, Fellow of the Indian National Science Academy, and Vice-President of the Calcutta Statistical Association. He is Past President, International Statistical Institute (1993).

Mohan Delampady is Professor, Theoretical Statistics and Mathematics Division, Indian Statistical Institute, Bangalore. He was awarded the Professor L.J. Savage prize for excellence in research towards a Ph.D. thesis in Bayesian econometrics and statistics by NBER and NSF in 1987.

Cross References

►Bayesian Statistics
►P-Values
►P-Values, Combining of

References and Further Reading

Bayarri MJ, Berger J (1998) Quantifying surprise in the data and model verification. In Bernardo JM et al. (eds) Bayesian statistics 6, Oxford University Press, Oxford, pp 53–82

Bayarri MJ, Berger JO (2000) P values for composite null models. J Am Stat Assoc 95:1127–1142, 1157–1170 (discussion)

Box GEP (1980) Sampling and Bayes inference in scientific modeling and robustness. J R Stat Soc (Series A) 143:383–430

Gelman A, Meng X, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. Stat Sinica 6:733–807 (with discussion)

Ghosh JK, Delampady M, Samanta T (2006) An introduction to Bayesian analysis: theory and methods. Springer, New York

Ghosh JK, Purkayastha S, Samanta T (2005) Role of P-values and other measures of evidence in Bayesian analysis. In: Dey DK, Rao CR (eds) Handbook of statistics, 25, Bayesian thinking: modeling and computation. pp 151–170

Guttman I (1967) The use of the concept of a future observation in goodness-of-fit problems. J R Stat Soc (Series B) 29:104–109

Meng XL (1994) Posterior predictive p-values. Ann Stat 22:1142–1160

Robins JM, van der Vaart A, Ventura V (2000) Asymptotic distribution of P Values for composite null models. J Am Stat *Assoc* 95:1143–1157, 1157–1170 (discussion)

Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Stat 12:1151–1172

Bayesian Reliability Modeling

Renkuan Guo
Professor
University of Cape Town, Cape Town, South Africa

Bayesian reliability modeling is an application of rigorous Bayesian statistical inference theory, one of the frontiers of modern statistics. It is particularly useful in the context of the scarcity of system failure (or quality index) data. Bayesian decision theory can guide reliability engineers to utilize both "soft" and "hard" evidence relevant to the reliability index under investigation. Soft evidence includes expert knowledge, design and performance of similar products, and their mathematical treatment, etc. In contrast, hard evidence includes the direct failure (or quality testing) data, any other transformable evidence having functional relationship with failure rate, such as test data from proving ground sources, factors in a product operating environment and any partially relevant evidence, say, warranty data, customer research surveys, etc.

The basic form of Bayesian reliability modeling can be illustrated by following formulation and example. When evidence of system performance denoted by $\underline{x}$, is scarce, a reliability (or quality) index of the system, denoted by $s\,(0 \le s \le 1)$, we may use the soft evidence on s in the form of a prior density $\rho(s)$ in terms of ►Bayes' theorem to calculate the posterior density of s,

$$f(s|\underline{x}) = \frac{l(s|\underline{x})\rho(s)}{\int_0^1 l(s|\underline{x})\rho(s)ds}. \tag{1}$$

where $l(s|\underline{x}) \triangleq f(\underline{x}|s)$ is called the likelihood function, obtained from the joint density of the sample evidence $\underline{x}$.

For example, N electronic devices are under testing until a preset time T, by assuming that the failure time of a random individual device follows an exponential distribution with the density:

$$f(t|\lambda) = \lambda e^{-\lambda t}, t \ge 0, \lambda > 0. \tag{2}$$

Suppose that in testing period $[0, T]$, x units failed and the failure times are recorded as $t_1, t_2, \cdots, t_x$, thus the sample evidence as $\underline{t} = (t_1, t_2, \cdots, t_x, T, T, \cdots, T)$, and the likelihood function is then given by

$$l(\lambda|\underline{t}) = \left[\prod_{i=1}^{x}\left(\lambda e^{-\lambda t_i}\right)\right]\left[1-\left(1-e^{-\lambda T}\right)\right]^{n-x} = \lambda^x e^{-\lambda\eta}, \tag{3}$$

where $\eta = (n-x)\,T + \sum_{i=1}^{x} t_i$. Then a gamma prior density on the failure rate λ, with priori parameters α and β, may represent the expert knowledge:

$$\rho(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\beta-1}e^{-\alpha\lambda}, \lambda > 0, \alpha \ge 0, \beta \ge 0. \tag{4}$$

The posterior density of the failure rate λ is evaluated in terms of Eq. 1:

$$f(\lambda|\underline{x}) = \frac{l(\lambda|\underline{x})\rho(\lambda)}{\int_0^1 l(\lambda|\underline{x})\rho(\lambda)d\lambda} = \frac{(\alpha+\eta)^{\beta+x}}{\Gamma(\alpha+x)}\lambda^{\beta+x+1}e^{-(\alpha+\eta)\lambda}, \tag{5}$$

which is also a gamma density because the prior density takes the form of the conjugate family. Further analysis or inference will be based on the posterior density.

It is necessary to emphasize that in reliability engineering reality, the formations and applications of Bayesian reliability modeling are far more complicated and diversified than that shown as the basic form in Eq. 1 because the concrete formation of an individual reliability problem is heavily dependent upon the compositional form of Bayesian decision criterion engaged, the form of distribution of

quality index and hence the form of the likelihood function, which is inevitably linked to a host of factors which may have serious impacts. The factors include any sample censoring mechanism, the form of prior and the design or physical structure of the system under study, the manner of collecting evidence (e.g., sequential or otherwise) and even research progress of the relevant mathematical branches.

For example, Dai et al. (2007), demonstrated that graph theory related fault tree analysis is a common classical reliability analysis model, while the Bayesian counterpart, Bayesian networks (abbreviated as BN), is just a combination of fault trees and appropriate arrangement of conditional probability and on probability assessments which can combine the soft and hard evidence under a Bayesian risk criterion. Bayesian networks can model the complicated physical structure as well as failure dependence structures of a complex system and predict the system failure behavior. A dynamic version of BN (DBN) was also developed for large system reliability modeling.

Artificial ▶Neural Networks form a powerful regression modeling tool, permitting a non-linear analysis. Combination of a Bayesian theoretical frame and Standard Neural Networks structure creates the Bayesian Neural Networks (abbreviated as BNNs), which allow active Bayesian reliability modeling for complex systems with complicated sample evidence. Mathematically, BNNs are nothing but probabilistic networks. For details, see Waszczyszyn et al. (2008). Markov chain Monte Carlo (abbreviated as MCMC) simulation (see ▶Markov Chain Monte Carlo) is another frontier of modern computational statistics. Merging of BNN and MCMC has created an extremely powerful but convenient computational reliability engineering model type.

It is well-known that the conventional Bayesian risk criterion is based on quadratic loss function and use of a conjugate family. Maximum Entropy modeling is an important Bayesian inference. The reliability engineering community has made efforts in this direction. However, Maximum entropy (abbreviated as Maxent) modeling software development and application in environmental context may improve from attention to Bayesian reliability modeling efforts. See Phillips et al. (2006).

Therefore, Bayesian reliability modeling is widely applied in business and industries, particularly in complex systems, nano-electronics and software industry, say, accelerating life testing models, reliability growth models, testing models for new product design, etc. See Blischke and Murthy (2000), Kuo (2006), and Garg et al. (2007).

As we pointed out at the beginning, Bayesian reliability modeling is a small-sample inference in its mathematical nature. Hence there is no reason to ignore the other developments in small-sample inference. It will be beneficial to combine some elements in small-sample inference with developments in the Bayesian reliability modeling. Jin et al. (2009) illustrated how to use a "grey differential equation" approach to improve BNNs in software reliability modeling, although the "grey" concept is not yet well-accepted in mathematical and statistical societies. Guo et al. (2009) proposed a differential equation associated regression (abbreviated as DEAR) model, which is small-sample based inference with a rigorous mathematical foundations and resolves the problem in "grey differential equations."

Ushakov and Harrison (1994) offered a systematic treatment in Chap. 16. Singpurwalla's book (2006) is authoritative in Bayesian reliability modeling, and reflects the author's state-of-art modeling experiences and judgments.

About the Author

Dr Renkuan Guo is a Professor in the Department of Statistical Sciences, University of Cape Town, South Africa. He served as Regional Editor for *Economic Quality Control*. His research work and publications (159 papers and four book chapters) can be divided into three aspects: (1) from 1990 to 2000, he focused on reliability modeling, including Cox PH models, and Kijima age models; (2) from 2001 to 2008, he concentrated on fuzzy reliability and quality modeling, including credibilistic fuzzy reliability models, grey differential equation models, random fuzzy models, particularly, the DEAR (Differential Equation Associated Regression) models; (3) from 2009 up to now, he starts initiating the uncertainty statistics on Professor Baoding Liu's axiomatic uncertain measure foundation, including uncertain Bayesian decision model, uncertain reliability models and hybrid reliability models with uncertain parameter.

Cross References

▶Degradation Models in Reliability and Survival Analysis
▶Imprecise Reliability
▶Industrial Statistics
▶Parametric and Nonparametric Reliability Analysis

References and Further Reading

Blischke WR, Murthy DNP (2000) Reliability – modeling, prediction, and optimization. Wiley, New York

Dai YS, Xie M, Long Q, Ng SH (2007) Uncertainty analysis in software reliability modeling by Bayesian analysis with maximum-entropy principle. IEEE Trans Software Eng 33(11):781–795

Garg RK, Gupta VK, Agrawal VP (2007) Reliability modelling and analysis of a tribo-mechanical system. Int J Model Simulat 27(3):288–294.

Guo R, Guo D, Tim Dunne T, Thiart C (2009) DEAR Model – The theoretical foundation. J Uncertain Syst 3(1):36–51
Jin A, Jiang JH, Lou JG, Zhang R (2009) Software reliability modeling based on grey system theory. J Comput Appl 29(3):690–694
Kuo W (2006) Challenges related to reliability in nano electronics. IEEE Trans Reliab 55(4):569–670
Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190:231–259
Singpurwalla ND (2006) Reliability and risk - a Bayesian perspective. Wiley, New York
Ushakov IA, Harrison RA (1994) Handbook of reliability engineering. Wiley, New York
Waszczyszyn Z, Slonski M, Miller B, Piatkowski G (2008) Bayesian neural networks in the regression analysis of structural mechanics problems. In: 8th World congress on computational mechanics (WCCM8), 5th European congress on computational methods in applied sciences and engineering (ECCOMAS 2008), Venice, Italy, 30 June–5 July, 2008

Bayesian Semiparametric Regression

Ludwig Fahrmeir
Professor
Ludwig-Maximilians-University, Munich, Germany

Linear or ►generalized linear models assume that the (conditional) mean $\mu = E(y|\mathbf{x})$, of the response y, given the covariate vector $\mathbf{x}$, is linked to a linear predictor μ by

$$\mu = h(\eta), \quad \eta = \mathbf{x}'\boldsymbol{\beta}.$$

Here, h is a known response function and $\boldsymbol{\beta}$ is an unknown vector of regression parameters. More generally, other characteristics of the response distribution, such as ►variance or ►skewness may be related to covariates in similar manner. Another example is the Cox model for ►survival data, where the hazard rate is assumed to have the form

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})$$

with $\lambda_0(t)$ as an (unspecified) baseline hazard rate. In most practical regression situations, however, we are facing at least one of the following problems.

(a) For the continuous covariates in the data set, the assumption of a strictly linear effect on the predictor may not be appropriate.
(b) Observations may be spatially correlated.
(c) Heterogeneity among individuals or units may be insufficiently described by covariates. Hence, unobserved unit- or cluster specific heterogeneity must be considered appropriately.
(d) Interactions between covariates may be of complex, nonlinear form.

►Semiparametric regression models extend models with linear predictors by incorporating additional non- and semiparametric components. Bayesian semiparametric regression regularizes the resulting high-dimensional inferential problem by imposing appropriate priors.

Observation Models

We consider some semiparametric regression models that may be considered as special classes of structured additive regression (STAR) models (Fahrmeir et al. 2004). Generalized additive models (GAMs) extend the linear predictor of GLMs to

$$\eta_i = x_i'\beta + f_1(z_1) + \ldots + f_p(z_{ip}), \quad i = 1, \ldots, n \qquad (1)$$

where f_j are smooth functions of continuous covariates $z_1, \ldots, z_p$. Most semiparametric regression approaches assume that unknown functions are represented or approximated through a linear combination of basis functions, i.e.,

$$f(z) = \sum_{k=1}^{K} \gamma_k B_k(z)$$

for a typical function f. The most popular basis function representations are spline functions, with truncated power series or B-spline basis functions $B_k(z)$, a relatively large number of knots, and a correspondingly high-dimensional vector $\boldsymbol{\gamma}$ of basis function coefficients $\gamma_1, ..., \gamma_K$.

Collecting all predictors η_i in the predictor vector $\boldsymbol{\eta}$, and constructing appropriate design matrix, the predictor (1) can be rewritten in matrix notation as a high-dimensional linear predictor

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \ldots + \mathbf{Z}_p\boldsymbol{\gamma}_p. \qquad (2)$$

GAMs can be extended by additively incorporating e.g., interaction terms $f_{1|2}(z_1, z_2)$ of two continuous covariates, varying coefficient terms $g(z)x$, where the effect of x varies with z, a spatial effect $f_{\text{spat}}(s)$ where s denotes spatial location and individual – or group specific i.i.d. random effects $\boldsymbol{\alpha}_g$, $g = 1, \ldots, G$. Combining these different types of effects in additive form leads to STAR models with generalized additive mixed models (GAMMs), varying coefficient models and geoadditive models as important subclasses. The Cox model can be extended in quite similar fashion, see Hennerfeind et al. (2006) and Kneib and Fahrmeir (2006). It turns out that after appropriate definition of

design matrices and coefficient vectors, the predictor still is of additive structure as in (2).

Priors and Inference

For Bayesian inference, flat priors $p(\beta) \propto const$ or weakly informative Gaussian priors are usually assumed for linear effects. In a Gaussian smoothness prior approach, it turns out that all priors for parameters γ_j representing smooth functions, Gaussian Markov random fields for spatial effects, i.i.d. random effects, etc., have the same generic conditionally Gaussian form

$$p\left(\boldsymbol{\gamma}_j|\tau_j^2\right) \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\gamma}_j'\mathbf{K}_j\boldsymbol{\gamma}_j\right). \tag{3}$$

The precision matrix $\mathbf{K}_j$ depends on the specific effect and acts as a penalty matrix to enforce smoothness, and τ_j^2 is an inverse smoothing parameter, controlling the amount of smoothness. In full Bayesian inference, a hyperprior is assigned to τ_j^2, and regression and smoothness parameters are estimated jointly through MCMC techniques (Brezger and Lang 2006). Another possibility is to look at (2) and (3) as a mixed model with correlated random effects, enabling empirical Bayes inference with mixed model technology, see Fahrmeir et al. (2004) and Ruppert et al. (2003). A recent review on (Bayesian) semiparametric regression is Ruppert et al. (2009). A forthcoming book (Fahrmeir and Kneib 2010) provides details on all issues. Approximate full Bayesian inference avoiding MCMC has been recently proposed by Rue et al. (2009).

A somewhat different approach to Bayesian inference in semiparametric regression is based on adaptive selection of knots of *B*-splines or coefficients of basis functions through concepts of Bayesian variable selection, see for example the book by Denison et al. (2002) or Smith et al. (2000), Kohn et al. (2001).

About the Author

Ludwig Fahrmeir is a Professor, Department of Statistics, Ludwig-Maximilians-University Munich, Germany. He was Chairman of the Collaborative Research Centre "Statistical Analysis of Discrete Structures with Applications in Econometrics and Biometrics" (1995–2006) and is currently coordinator of the project "Analysis and Modelling of Complex Systems in Biology and Medicine" at the University of Munich. He is an Elected Fellow of the International Statistical Institute. He has authored and coauthored more than 100 papers and seven books, including *Multivariate statistical modelling based on generalized linear models* (with Tutz, G., 2nd enlarged edition, Springer Series in Statistics, New York, 2001).

Cross References

►Bayesian Statistics
►Semiparametric Regression Models

References and Further Reading

Brezger A, Lang S (2006) Generalized structured additive regression based on Bayesian *P*-splines. Comput Stat Data Anal 50: 967–991

Denison DGT, Holmes CC, Mallick BK, Smith AFM (2002) Bayesian methods for nonlinear classification and regression. Wiley, Chichester

Fahrmeir L, Kneib T (2010) Bayesian smoothing and regression of longitudinal, spatial and event history data. Oxford University Press, to appear

Fahrmeir L, Kneib T, Lang S (2004) Penalized structured additive regression for space-time data: a Bayesian perspective. Stat Sinica 14:731–761

Hennerfeind A, Brezger A, Fahrmeir L (2006) Geoadditive survival models. J Am Stat Assoc 101:1065–1075

Kneib T, Fahrmeir L (2006) Structured additive regression for multi-categorical space-time data: a mixed model approach. Biometrics 62:109–118

Kohn R, Smith M, Chan D (2001) Nonparametric regression using linear combinations of basis functions. Stat Comput 11: 313–322

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. J R Stat Soc B 71:1–35

Ruppert D, Wand M, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge

Ruppert D, Wand MP, Carroll RJ (2009) Semiparametric regression during 2003–2007. Electron J Stat 3:1193–1256

Smith M, Kohn R, Yau P (2000) Nonparametric Bayesian bivariate surface estimation. In: Schimek G (ed) Smoothing and regression, Ch 19. Wiley, New York

Bayesian Statistics

José M. Bernardo
Professor of Statistics, Facultad de Matemáticas
Universitat de València, Burjassot, Spain

Introduction

Available observations generally consist of (possibly many) sets of data of the general form $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, where the $\boldsymbol{x}_i$'s are somewhat "homogeneous" (possibly multidimensional) observations $\boldsymbol{x}_i$. Statistical methods are then typically used to derive conclusions on both the nature of the process which has produced those observations, and on the expected behavior at future instances of the same process. A central element of *any* statistical analysis is the specification of a *probability model* which is assumed to describe

the mechanism which has generated the observed data D as a function of a (possibly multidimensional) parameter (vector) $\boldsymbol{\omega} \in \boldsymbol{\Omega}$, sometimes referred to as the *state of nature*, about whose value only limited information (if any) is available. All derived statistical conclusions are obviously conditional on the assumed probability model.

Unlike most other branches of mathematics, conventional methods of statistical inference suffer from the lack of an axiomatic basis; as a consequence, their proposed desiderata are often mutually incompatible, and the analysis of the same data may well lead to incompatible results when different, apparently intuitive procedures are tried (see Lindley (1970) and Jaynes (1976) for many instructive examples). In marked contrast, the Bayesian approach to statistical inference is firmly based on axiomatic foundations which provide a unifying logical structure, and guarantee the mutual consistency of the methods proposed. Bayesian methods constitute a *complete* paradigm to statistical inference, a scientific revolution in Kuhn's sense.

Bayesian statistics only require the *mathematics* of probability theory and the *interpretation* of probability which most closely corresponds to the standard use of this word in everyday language: it is no accident that some of the more important seminal books on Bayesian statistics, such as the works of Laplace (1812), Jeffreys (1961) or de Finetti (1970) are actually entitled "Probability Theory." The practical consequences of adopting the Bayesian paradigm are far reaching. Indeed, Bayesian methods (1) reduce statistical inference to problems in probability theory, thereby minimizing the need for completely new concepts, and (2) serve to discriminate among conventional statistical techniques, by either providing a logical justification to some (and making explicit the conditions under which they are valid), or proving the logical inconsistency of others.

The main consequence of these foundations is the mathematical *need* to describe by means of probability distributions all uncertainties present in the problem. In particular, unknown parameters in probability models *must* have a joint probability distribution which describes the available information about their values; this is often regarded as *the* characteristic element of a Bayesian approach. Notice that (in sharp contrast to conventional statistics) *parameters are treated as random variables* within the Bayesian paradigm. This is not a description of their variability (parameters are typically *fixed unknown* quantities) but a description of the *uncertainty* about their true values.

An important particular case arises when either no relevant prior information is readily available, or that information is subjective and an "objective" analysis is desired, one that is exclusively based on accepted model assumptions and well-documented data. This is addressed by *reference analysis*, which uses information-theoretic concepts to derive appropriate reference posterior distributions, defined to encapsulate inferential conclusions on the quantities of interest solely based on the supposed model and the observed data.

In this article it is assumed that probability distributions may be described through their probability density functions, and no distinction is made between a random quantity and the particular values that it may take. Bold italic roman fonts are used for *observable* random vectors (typically data) and bold italic greek fonts are used for unobservable random vectors (typically parameters); lower case is used for variables and upper case for their dominion sets. Moreover, the standard mathematical convention of referring to *functions*, say f and g of $\boldsymbol{x} \in \boldsymbol{\chi}$, respectively by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$, will be used throughout. Thus, $p(\boldsymbol{\theta} \mid C)$ and $p(\boldsymbol{x} \mid C)$ respectively represent general *probability densities* of the random vectors $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{x} \in \boldsymbol{\chi}$ under conditions C, so that $p(\boldsymbol{\theta} \mid C) \geq 0$, $\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta} \mid C)\, d\boldsymbol{\theta} = 1$, and $p(\boldsymbol{x} \mid C) \geq 0$, $\int_{\boldsymbol{\chi}} p(\boldsymbol{x} \mid C)\, d\boldsymbol{x} = 1$. This admittedly imprecise notation will greatly simplify the exposition. If the random vectors are discrete, these functions naturally become probability mass functions, and integrals over their values become sums.

Density functions of specific distributions are denoted by appropriate names. Thus, if x is a random quantity with a normal distribution of mean μ and standard deviation σ, its probability density function will be denoted $\mathrm{N}(x \mid \mu, \sigma)$.

Bayesian methods make frequent use of the concept of logarithmic divergence, a very general measure of the goodness of the approximation of a probability density $p(\boldsymbol{x})$ by another density $\hat{p}(\boldsymbol{x})$. The Kullback-Leibler, or *logarithmic divergence* of a probability density $\hat{p}(\boldsymbol{x})$ of the random vector $\boldsymbol{x} \in \boldsymbol{\chi}$ from its true probability density $p(\boldsymbol{x})$, is defined as $\delta\{\hat{p}(\boldsymbol{x}) \mid p(\boldsymbol{x})\} = \int_{\boldsymbol{\chi}} p(\boldsymbol{x}) \log\{p(\boldsymbol{x})/\hat{p}(\boldsymbol{x})\}\, d\boldsymbol{x}$. It may be shown that (1) the logarithmic divergence is non-negative (and it is zero if, and only if, $\hat{p}(\boldsymbol{x}) = p(\boldsymbol{x})$ almost everywhere), and (2) that $\delta\{\hat{p}(\boldsymbol{x}) \mid p(\boldsymbol{x})\}$ is invariant under one-to-one transformations of $\boldsymbol{x}$.

This article contains a brief summary of the mathematical foundations of Bayesian statistical methods (section "►Foundations"), an overview of the paradigm (section "►The Bayesian Paradigm"), a description of useful inference summaries, including both point an region estimation and hypothesis testing (section "►Inference Summaries"), an explicit discussion of objective Bayesian methods (section "►Reference Analysis"), and a final

discussion which includes pointers to further issues not addressed here (section "►Discussion").

Foundations

A central element of the Bayesian paradigm is the use of probability distributions to describe all relevant unknown quantities, interpreting the probability of an event as a conditional measure of uncertainty, on a $[0,1]$ scale, about the occurrence of the event in some specific conditions. The limiting extreme values 0 and 1, which are typically inaccessible in applications, respectively describe impossibility and certainty of the occurrence of the event. This interpretation of probability includes and extends all other probability interpretations. There are two independent arguments which prove the mathematical inevitability of the use of probability distributions to describe uncertainties; these are summarized later in this section.

Probability as a Measure of Conditional Uncertainty

Bayesian statistics uses the word *probability* in precisely the same sense in which this word is used in everyday language, as a *conditional measure of uncertainty* associated with the occurrence of a particular event, given the available information and the accepted assumptions. Thus, $\Pr(E \mid C)$ is a measure of (presumably rational) belief in the occurrence of the *event* E under *conditions* C. It is important to stress that probability is *always* a function of two arguments, the event E whose uncertainty is being measured, and the conditions C under which the measurement takes place; "absolute" probabilities do not exist. In typical applications, one is interested in the probability of some event E given the available *data* D, the set of *assumptions* A which one is prepared to make about the mechanism which has generated the data, and the relevant contextual *knowledge* K which might be available. Thus, $\Pr(E \mid D, A, K)$ is to be interpreted as a measure of (presumably rational) belief in the occurrence of the *event* E, given data D, assumptions A and any other available knowledge K, as a measure of how "likely" is the occurrence of E in these conditions. Sometimes, but certainly not always, the probability of an event under given conditions may be associated with the relative frequency of "similar" events in "similar" conditions. The following examples are intended to illustrate the use of probability as a conditional measure of uncertainty.

Probabilistic diagnosis. A human population is known to contain 0.2% of people infected by a particular virus. A person, *randomly selected* from that population, is subject to a test which, from laboratory data, is known to yield positive results in 98% of infected people and in 1% of non-infected, so that, if V denotes the event that a person carries the virus and + denotes a positive result, $\Pr(+ \mid V) = 0.98$ and $\Pr(+ \mid \overline{V}) = 0.01$. Suppose that the result of the test turns out to be positive. Clearly, one is then interested in $\Pr(V \mid +, A, K)$, the *probability* that the person carries the virus, given the positive result, the assumptions A about the probability mechanism generating the test results, and the available knowledge K of the prevalence of the infection in the population under study (described here by $\Pr(V \mid K) = 0.002$). An elementary exercise in probability algebra, which involves ►Bayes' theorem in its simplest form (see section "►The Bayesian Paradigm"), yields $\Pr(V \mid +, A, K) = 0.164$. Notice that the four probabilities involved in the problem have *the same interpretation*: they are all conditional measures of uncertainty. Besides, $\Pr(V \mid +, A, K)$ is *both* a measure of the uncertainty associated with the event that the particular person who tested positive is actually infected, *and* an *estimate* of the proportion of people in that population (about 16.4%) that would eventually prove to be infected among those which yielded a positive test.

Estimation of a proportion. A survey is conducted to estimate the proportion θ of individuals in a population who share a given property. A random sample of n elements is analyzed, r of which are found to possess that property. One is then typically interested in using the results from the sample to establish regions of $[0,1]$ where the unknown value of θ may plausibly be expected to lie; this information is provided by *probabilities* of the form $\Pr(a < \theta < b \mid r, n, A, K)$, a conditional measure of the uncertainty about the event that θ belongs to (a, b) *given* the information provided by the data (r, n), the assumptions A made on the behavior of the mechanism which has generated the data (a random sample of n Bernoulli trials), and any relevant knowledge K on the values of θ which might be available. For example, after a political survey in which 720 citizens out of a random sample of 1500 have declared their support to a particular political measure, one may conclude that $\Pr(\theta < 0.5 \mid 720, 1,500, A, K) = 0.933$, indicating a probability of about 93% that a referendum of that issue would be lost. Similarly, after a screening test for an infection where 100 people have been tested, none of which has turned out to be infected, one may conclude that $\Pr(\theta < 0.01 \mid 0, 100, A, K) = 0.844$, or a probability of about 84% that the proportion of infected people is smaller than 1%.

Measurement of a physical constant. A team of scientists, intending to establish the unknown value of a physical constant μ, obtain data $D = \{x_1, \ldots, x_n\}$ which are considered to be measurements of μ subject to error. The probabilities of interest are then typically of the form

$\Pr(a < \mu < b \,|\, x_1, \ldots, x_n, A, K)$, the *probability* that the unknown value of μ (fixed in nature, but unknown to the scientists) lies within an interval (a, b) given the information provided by the data D, the assumptions A made on the behavior of the measurement mechanism, and whatever knowledge K might be available on the value of the constant μ. Again, those probabilities are conditional measures of uncertainty which describe the (necessarily probabilistic) conclusions of the scientists on the true value of μ, given available information and accepted assumptions. For example, after a classroom experiment to measure the gravitational field with a pendulum, a student may report (in m/sec^2) something like $\Pr(9.788 < g < 9.829 \,|\, D, A, K) = 0.95$, meaning that, under accepted knowledge K and assumptions A, the *observed* data D indicate that the true value of g lies within 9.788 and 9.829 with probability 0.95, a conditional uncertainty measure on a [0,1] scale. This is naturally compatible with the fact that the value of the gravitational field at the laboratory may well be known with high precision from available literature or from precise previous experiments, but the student may have been instructed *not* to use that information as part of the accepted knowledge K. Under some conditions, it is also true that if the same *procedure* were actually used by many other students with similarly obtained data sets, their reported intervals would actually cover the true value of g in approximately 95% of the cases, thus providing some form of *calibration* for the student's probability statement (see section "►Frequentist Properties").

Prediction. An experiment is made to count the number r of times that an event E takes place in each of n replications of a well defined situation; it is observed that E does take place r_i times in replication i, and it is desired to forecast the number of times r that E will take place in a future, similar situation. This is a *prediction* problem on the value of an *observable* (discrete) quantity r, given the information provided by data D, accepted assumptions A on the probability mechanism which generates the r_i's, and any relevant available knowledge K. Hence, simply the computation of the probabilities $\{\Pr(r \,|\, r_1, \ldots, r_n, A, K)\}$, for $r = 0, 1, \ldots$, is required. For example, the quality assurance engineer of a firm which produces automobile restraint systems may report something like $\Pr(r = 0 \,|\, r_1 = \ldots = r_{10} = 0, A, K) = 0.953$, after observing that the entire production of airbags in each of $n = 10$ consecutive months has yielded no complaints from their clients. This should be regarded as a measure, on a $[0, 1]$ scale, of the conditional uncertainty, given observed data, accepted assumptions and contextual knowledge, associated with the event that no airbag complaint will come from next month's production and, if conditions remain constant, this is also an estimate of the proportion of months expected to share this desirable property.

A similar problem may naturally be posed with continuous observables. For instance, after measuring some continuous magnitude in each of n randomly chosen elements within a population, it may be desired to forecast the proportion of items in the whole population whose magnitude satisfies some precise specifications. As an example, after measuring the breaking strengths $\{x_1, \ldots, x_{10}\}$ of ten randomly chosen safety belt webbings to verify whether or not they satisfy the requirement of remaining above 26 kN, the quality assurance engineer may report something like $\Pr(x > 26 \,|\, x_1, \ldots, x_{10}, A, K) = 0.9987$. This should be regarded as a measure, on a $[0, 1]$ scale, of the conditional uncertainty (given observed data, accepted assumptions and contextual knowledge) associated with the event that a randomly chosen safety belt webbing will support no less than 26 kN. If production conditions remain constant, it will also be an estimate of the proportion of safety belts which will conform to this particular specification.

Often, additional information of future observations is provided by related covariates. For instance, after observing the outputs $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ which correspond to a sequence $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of different production conditions, it may be desired to forecast the output $\boldsymbol{y}$ which would correspond to a particular set $\boldsymbol{x}$ of production conditions. For instance, the viscosity of commercial condensed milk is required to be within specified values a and b; after measuring the viscosities $\{y_1, \ldots, y_n\}$ which correspond to samples of condensed milk produced under different physical conditions $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, production engineers will require probabilities of the form $\Pr(a < y < b \,|\, \boldsymbol{x}, (y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n), A, K)$. This is a conditional measure of the uncertainty (always given observed data, accepted assumptions and contextual knowledge) associated with the event that condensed milk produced under conditions $\boldsymbol{x}$ will actually satisfy the required viscosity specifications.

Statistical Inference and Decision Theory

Decision theory not only provides a precise methodology to deal with decision problems under uncertainty, but its solid axiomatic basis also provides a powerful reinforcement to the logical power of the Bayesian approach. We now summarize the basic argument.

A decision problem exists whenever there are two or more possible courses of action; let $\mathcal{A}$ be the class of possible actions. Moreover, for each $a \in \mathcal{A}$, let $\boldsymbol{\Theta}_a$ be the set of *relevant events* which may affect the result of choosing a, and let $c(a, \boldsymbol{\theta}) \in \mathcal{C}_a$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_a$, be the *consequence* of having chosen action a when event $\boldsymbol{\theta}$ takes place. The class of

pairs $\{(\Theta_a, \mathcal{C}_a), a \in \mathcal{A}\}$ describes the *structure* of the decision problem. Without loss of generality, it may be assumed that the possible actions are mutually exclusive, for otherwise one would work with the appropriate Cartesian product.

Different sets of principles have been proposed to capture a minimum collection of logical rules that could sensibly be required for "rational" decision-making. These all consist of axioms with a strong intuitive appeal; examples include the *transitivity* of preferences (if $a_1 > a_2$ given C, and $a_2 > a_3$ given C, then $a_1 > a_3$ given C), and the *sure-thing principle* (if $a_1 > a_2$ given C and E, and $a_1 > a_2$ given C and $\bar{E}$, then $a_1 > a_2$ given C). Notice that these rules are *not* intended as a description of actual human decision-making, but as a *normative* set of principles to be followed by someone who aspires to achieve coherent decision-making.

There are naturally different options for the set of acceptable principles, but all of them lead basically to the same conclusions, namely:

1. Preferences among consequences should necessarily be measured with a real-valued bounded *utility* function $u(c) = u(a, \boldsymbol{\theta})$ which specifies, on some numerical scale, their desirability.
2. The uncertainty of relevant events should be measured with a set of *probability* distributions $\{(p(\boldsymbol{\theta} \mid C, a), \boldsymbol{\theta} \in \Theta_a), a \in \mathcal{A}\}$ describing their plausibility given the conditions C under which the decision must be taken.
3. The desirability of the available actions is measured by their corresponding *expected utility*

$$\overline{u}(a \mid C) = \int_{\Theta_a} u(a, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid C, a)\, d\boldsymbol{\theta}, \quad a \in \mathcal{A}.$$

It is often convenient to work in terms of the non-negative *loss* function defined by

$$\ell(a, \boldsymbol{\theta}) = \sup_{a \in \mathcal{A}} \{u(a, \boldsymbol{\theta})\} - u(a, \boldsymbol{\theta}),$$

which directly measures, as a function of $\boldsymbol{\theta}$, the "penalty" for choosing a wrong action. The relative undesirability of available actions $a \in \mathcal{A}$ is then measured by their *expected loss*

$$\overline{\ell}(a \mid C) = \int_{\Theta_a} \ell(a, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid C, a)\, d\boldsymbol{\theta}, \quad a \in \mathcal{A}.$$

Notice that, in particular, the argument described above establishes the need to quantify the uncertainty about all relevant unknown quantities (the actual values of the $\boldsymbol{\theta}$'s), and specifies that this quantification *must* have the mathematical structure of probability distributions. These probabilities are conditional on the circumstances C under which the decision is to be taken, which typically, but not necessarily, include the results D of some relevant experimental or observational data.

It has been argued that the development described above (which is not questioned when decisions have to be made) does not apply to problems of statistical inference, where no specific decision making is envisaged. However, there are two powerful counterarguments to this. Indeed, (1) a problem of statistical inference is typically considered worth analyzing because it *may* eventually help make sensible decisions (as Ramsey (1931) put it, a lump of arsenic is poisonous because it *may* kill someone, not because it has actually killed someone), and (2) it has been shown (Bernardo 1979a) that statistical inference on $\boldsymbol{\theta}$ actually *has* the mathematical structure of a decision problem, where the class of alternatives is the functional space

$$\mathcal{A} = \left\{ p(\boldsymbol{\theta} \mid D); \quad p(\boldsymbol{\theta} \mid D) > 0, \int_{\Theta} p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} = 1 \right\}$$

of the conditional probability distributions of $\boldsymbol{\theta}$ given the data, and the utility function is a measure of the amount of information about $\boldsymbol{\theta}$ which the data may be expected to provide.

Exchangeability and Representation Theorem

Available data often take the form of a set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of "homogeneous" (possibly multidimensional) observations, in the precise sense that only their *values* matter and not the *order* in which they appear. Formally, this is captured by the notion of *exchangeability*. The set of random vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is exchangeable if their joint distribution is invariant under permutations. An infinite sequence $\{\boldsymbol{x}_j\}$ of random vectors is exchangeable if all its finite subsequences are exchangeable. Notice that, in particular, any random sample from any model is exchangeable in this sense. The concept of exchangeability, introduced by de Finetti (1937) put it, is central to modern statistical thinking. Indeed, the general *representation theorem* implies that if a set of observations is assumed to be a subset of an exchangeable sequence, then it constitutes *a random sample* from some probability model $\{p(\boldsymbol{x} \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, $\boldsymbol{x} \in \mathcal{X}$, described in terms of (labeled by) some *parameter vector* $\boldsymbol{\omega}$; furthermore this parameter $\boldsymbol{\omega}$ is *defined* as the limit (as $n \to \infty$) of some function of the observations. Available information about the value of $\boldsymbol{\omega}$ in prevailing conditions C is *necessarily* described by *some* probability distribution $p(\boldsymbol{\omega} \mid C)$.

For example, in the case of a sequence $\{x_1, x_2, \ldots\}$ of dichotomous exchangeable random quantities $x_j \in \{0, 1\}$, de Finetti's representation theorem establishes that the

joint distribution of $(x_1, \ldots, x_n)$ has an *integral representation* of the form

$$p(x_1, \ldots, x_n \,|\, C) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} \, p(\theta \,|\, C) \, d\theta,$$

$$\theta = \lim_{n \to \infty} \frac{r}{n},$$

where $r = \sum x_j$ is the number of positive trials. This is nothing but the joint distribution of a set of (conditionally) independent Bernoulli trials with parameter θ, over which some probability distribution $p(\theta \,|\, C)$ is therefore proven to exist. More generally, for sequences of arbitrary random quantities $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$, exchangeability leads to integral representations of the form

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \,|\, C) = \int_{\boldsymbol{\Omega}} \prod_{i=1}^{n} p(\boldsymbol{x}_i \,|\, \boldsymbol{\omega}) \, p(\boldsymbol{\omega} \,|\, C) \, d\boldsymbol{\omega},$$

where $\{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ denotes some probability *model*, $\boldsymbol{\omega}$ is the limit as $n \to \infty$ of some function $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ of the observations, and $p(\boldsymbol{\omega} \,|\, C)$ is some probability distribution over $\boldsymbol{\Omega}$. This formulation includes "nonparametric" (distribution free) modeling, where $\boldsymbol{\omega}$ may index, for instance, all continuous probability distributions on $\mathcal{X}$. Notice that $p(\boldsymbol{\omega} \,|\, C)$ does *not* describe a possible variability of $\boldsymbol{\omega}$ (since $\boldsymbol{\omega}$ will typically be a fixed *unknown* vector), but a description on the uncertainty associated with its actual value.

Under appropriate conditioning, exchangeability is a very general assumption, a powerful extension of the traditional concept of a *random sample*. Indeed, many statistical analyses directly assume data (or subsets of the data) to be a random sample of conditionally independent observations from some probability model, so that $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \,|\, \boldsymbol{\omega}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i \,|\, \boldsymbol{\omega})$; but *any* random sample is exchangeable, since $\prod_{i=1}^{n} p(\boldsymbol{x}_i \,|\, \boldsymbol{\omega})$ is obviously invariant under permutations. Notice that the observations in a random sample are only independent *conditional* on the parameter value $\boldsymbol{\omega}$; as nicely put by Lindley, the mantra that the observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in a random sample are independent is ridiculous when they are used to infer $\boldsymbol{x}_{n+1}$. Notice also that, under exchangeability, the general representation theorem provides an *existence theorem* for a probability distribution $p(\boldsymbol{\omega} \,|\, C)$ on the parameter space Ω, and that this is an argument which only depends on mathematical probability theory.

Another important consequence of exchangeability is that it provides a formal *definition* of the parameter $\boldsymbol{\omega}$ which labels the model as the limit, as $n \to \infty$, of *some* function $f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ of the observations; the function f obviously depends both on the assumed model and the chosen parametrization. For instance, in the case of a sequence of Bernoulli trials, the parameter θ is *defined* as the limit, as $n \to \infty$, of the relative frequency r/n. It follows that, under exchangeability, the sentence "the true value of $\boldsymbol{\omega}$" has a well-defined meaning, if only asymptotically verifiable. Moreover, if two different models have parameters which are functionally related by their definition, then the corresponding posterior distributions may be meaningfully compared, for they refer to functionally related quantities. For instance, if a finite subset $\{x_1, \ldots, x_n\}$ of an exchangeable sequence of integer observations is assumed to be a random sample from a Poisson distribution $\mathrm{Po}(x \,|\, \lambda)$, so that $\mathrm{E}[x \,|\, \lambda] = \lambda$, then λ is *defined* as $\lim_{n \to \infty} \{\bar{x}_n\}$, where $\bar{x}_n = \sum_j x_j / n$; similarly, if for some fixed non-zero integer r, the same data are assumed to be a random sample for a negative binomial $\mathrm{NBi}(x \,|\, r, \theta)$, so that $\mathrm{E}[x \,|\, \theta, r] = r(1-\theta)/\theta$, then θ is *defined* as $\lim_{n \to \infty} \{r/(\bar{x}_n + r)\}$. It follows that $\theta \equiv r/(\lambda + r)$ and, hence, θ and $r/(\lambda + r)$ may be treated as the *same* (unknown) quantity whenever this might be needed as, for example, when comparing the relative merits of these alternative probability models.

The Bayesian Paradigm

The statistical analysis of some observed data D typically begins with some informal *descriptive* evaluation, which is used to suggest a tentative, formal *probability model* $\{p(D \,|\, \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ assumed to represent, for some (unknown) value of $\boldsymbol{\omega}$, the probabilistic mechanism which has generated the observed data D. The arguments outlined in section "▶Foundations" establish the logical need to assess a *prior* probability distribution $p(\boldsymbol{\omega} \,|\, K)$ over the parameter space $\boldsymbol{\Omega}$, describing the available knowledge K about the value of $\boldsymbol{\omega}$ prior to the data being observed. It then follows from standard probability theory that, if the probability model is correct, all available information about the value of $\boldsymbol{\omega}$ after the data D have been observed is contained in the corresponding *posterior* distribution whose probability density, $p(\boldsymbol{\omega} \,|\, D, A, K)$, is immediately obtained from Bayes' theorem,

$$p(\boldsymbol{\omega} \,|\, D, A, K) = \frac{p(D \,|\, \boldsymbol{\omega}) \, p(\boldsymbol{\omega} \,|\, K)}{\int_{\boldsymbol{\Omega}} p(D \,|\, \boldsymbol{\omega}) \, p(\boldsymbol{\omega} \,|\, K) \, d\boldsymbol{\omega}},$$

where A stands for the assumptions made on the probability model. It is this systematic use of Bayes' theorem to incorporate the information provided by the data that justifies the adjective *Bayesian* by which the paradigm is usually known. It is obvious from Bayes' theorem that any value of $\boldsymbol{\omega}$ with zero prior density will have zero posterior density. Thus, it is typically assumed (by appropriate restriction, if necessary, of the *parameter space* $\boldsymbol{\Omega}$) that prior distributions are *strictly positive* (as Savage put it,

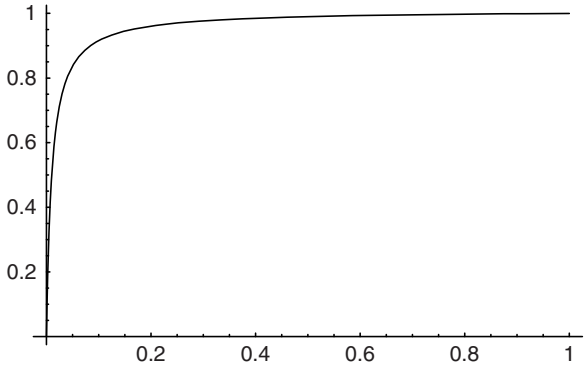
keep the mind open, or at least ajar). To simplify the presentation, the accepted assumptions A and the available knowledge K are often omitted from the notation, but the fact that *all* statements about $\boldsymbol{\omega}$ given D are *also* conditional to A and K should always be kept in mind.

Example 1 (Bayesian inference with a finite parameter space). Let $p(D\,|\,\theta)$, where $\theta \in \{\theta_1, \ldots, \theta_m\}$, be the probability mechanism which is assumed to have generated the observed data D, so that θ may only take a *finite* number of values. Using the finite form of Bayes' theorem, and omitting the prevailing conditions from the notation, the posterior probability of θ_i after data D have been observed is

$$\Pr(\theta_i\,|\,D) = \frac{p(D\,|\,\theta_i)\,\Pr(\theta_i)}{\sum_{j=1}^{m} p(D\,|\,\theta_j)\,\Pr(\theta_j)}, \quad i = 1, \ldots, m.$$

For any prior distribution $p(\theta) = \{\Pr(\theta_1), \ldots, \Pr(\theta_m)\}$ describing available knowledge about the value of θ, $\Pr(\theta_i\,|\,D)$ measures how likely should θ_i be judged, given both the initial knowledge described by the prior distribution, and the information provided by the data D.

An important, frequent application of this simple technique is provided by probabilistic diagnosis. For example, consider the simple situation where a particular test designed to detect a virus is known from laboratory research to give a positive result in 98% of infected people and in 1% of non-infected. Then, the posterior probability that a person who tested positive is infected is given by $\Pr(V\,|\,+) = (0.98\,p)/\{0.98\,p + 0.01\,(1-p)\}$ as a function of $p = \Pr(V)$, the prior probability of a person being infected (the *prevalence* of the infection in the population under study). Figure 1 shows $\Pr(V\,|\,+)$ as a function of $\Pr(V)$.



Bayesian Statistics. Fig. 1 Posterior probability of infection $\Pr(V\,|\,+)$ given a positive test, as a function of the prior probability of infection $\Pr(V)$

As one would expect, the posterior probability is only zero if the prior probability is zero (so that it is *known* that the population is free of infection) and it is only one if the prior probability is one (so that it is *known* that the population is universally infected). Notice that if the infection is rare, then the posterior probability of a randomly chosen person being infected will be relatively low even if the test is positive. Indeed, for say $\Pr(V) = 0.002$, one finds $\Pr(V\,|\,+) = 0.164$, so that in a population where only 0.2% of individuals are infected, only 16.4% of those testing positive within a random sample will actually prove to be infected: most positives would actually be *false* positives.

In the rest of this section, we describe in some detail the learning process described by Bayes' theorem, discuss its implementation in the presence of nuisance parameters, show how it can be used to forecast the value of future observations, and analyze its large sample behavior.

The Learning Process

In the Bayesian paradigm, the process of learning from the data is systematically implemented by making use of Bayes' theorem to combine the available prior information with the information provided by the data to produce the required posterior distribution. Computation of posterior densities is often facilitated by noting that Bayes' theorem may be simply expressed as

$$p(\boldsymbol{\omega}\,|\,D) \propto p(D\,|\,\boldsymbol{\omega})\,p(\boldsymbol{\omega}),$$

(where $\propto$ stands for 'proportional to' and where, for simplicity, the accepted assumptions A and the available knowledge K have been omitted from the notation), since the missing proportionality constant $[\int_{\Omega} p(D\,|\,\boldsymbol{\omega})\,p(\boldsymbol{\omega})\,d\boldsymbol{\omega}]^{-1}$ may always be deduced from the fact that $p(\boldsymbol{\omega}\,|\,D)$, a probability density, must integrate to one. Hence, to identify the form of a posterior distribution it suffices to identify a *kernel* of the corresponding probability density, that is a function $k(\boldsymbol{\omega})$ such that $p(\boldsymbol{\omega}\,|\,D) = c(D)\,k(\boldsymbol{\omega})$ for some $c(D)$ which does not involve $\boldsymbol{\omega}$. In the examples which follow, this technique will often be used.

An *improper prior function* is defined as a positive function $\pi(\boldsymbol{\omega})$ such that $\int_{\Omega} \pi(\boldsymbol{\omega})\,d\boldsymbol{\omega}$ is not finite. The formal expression of Bayes' theorem, remains technically valid if $p(\boldsymbol{\omega})$ is replaced by an improper prior function $\pi(\boldsymbol{\omega})$ provided the proportionality constant exists, thus leading to a well defined *proper* posterior density $\pi(\boldsymbol{\omega}\,|\,D) \propto p(D\,|\,\boldsymbol{\omega})\pi(\boldsymbol{\omega})$. It will later be established (section "▶Reference Analysis") that Bayes' theorem also remains philosophically valid if $p(\boldsymbol{\omega})$ is replaced by an appropriately chosen *reference* "noninformative" (typically improper) prior function $\pi(\boldsymbol{\omega})$.

Considered as a function of $\boldsymbol{\omega}$, $p(D\,|\,\boldsymbol{\omega})$ is often referred to as the *likelihood function*. Thus, Bayes' theorem is simply expressed in words by the statement that *the posterior is proportional to the likelihood times the prior*. It follows from Bayes' theorem that, provided the *same* prior $p(\boldsymbol{\omega})$ is used, two different data sets D_1 and D_2, with possibly different probability models $p_1(D_1\,|\,\boldsymbol{\omega})$ and $p_2(D_2\,|\,\boldsymbol{\omega})$ but yielding *proportional* likelihood functions, will produce identical posterior distributions for $\boldsymbol{\omega}$. This immediate consequence of Bayes theorem has been proposed as a principle on its own, the *likelihood principle*, and it is seen by many as an obvious requirement for reasonable statistical inference. In particular, for any given prior $p(\boldsymbol{\omega})$, the posterior distribution does not depend on the set of possible data values, or the *outcome space*. Notice, however, that the likelihood principle only applies to inferences about the parameter vector $\boldsymbol{\omega}$ *once the data have been obtained*. Consideration of the outcome space is essential, for instance, in model criticism, in the design of experiments, in the derivation of predictive distributions, or (see section "►Reference Analysis") in the construction of objective Bayesian procedures.

Naturally, the terms prior and posterior are only *relative* to a particular set of data. As one would expect from the coherence induced by probability theory, if data $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are sequentially presented, the final result will be the same whether data are globally or sequentially processed. Indeed, $p(\boldsymbol{\omega}\,|\,\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1}) \propto p(\boldsymbol{x}_{i+1}\,|\,\boldsymbol{\omega})\, p(\boldsymbol{\omega}\,|\,\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$, for $i = 1, \ldots, n-1$, so that the "posterior" at a given stage becomes the "prior" at the next.

In most situations, the posterior distribution is "sharper" than the prior so that, in most cases, the density $p(\boldsymbol{\omega}\,|\,\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i+1})$ will be more concentrated around the true value of $\boldsymbol{\omega}$ than $p(\boldsymbol{\omega}\,|\,\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i)$. However, this is not always the case: occasionally, a "surprising" observation will increase, rather than decrease, the uncertainty about the value of $\boldsymbol{\omega}$. For instance, in probabilistic diagnosis, a sharp posterior probability distribution (over the possible causes $\{\omega_1, \ldots, \omega_k\}$ of a syndrome) describing, a "clear" diagnosis of disease ω_i (that is, a posterior with a large probability for ω_i) would typically update to a less concentrated posterior probability distribution over $\{\omega_1, \ldots, \omega_k\}$ if a new clinical analysis yielded data which were unlikely under ω_i.

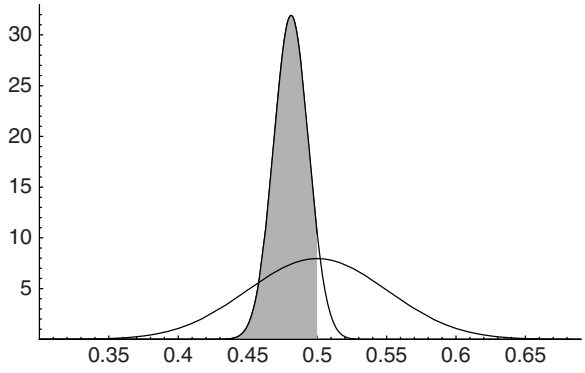
For a given probability model, one may find that some particular function of the data $\boldsymbol{t} = \boldsymbol{t}(D)$ is a *sufficient* statistic in the sense that, given the model, $\boldsymbol{t}(D)$ contains all information about $\boldsymbol{\omega}$ which is available in D. Formally, $\boldsymbol{t} = \boldsymbol{t}(D)$ is sufficient if (and only if) there exist nonnegative functions f and g such that the likelihood function may be factorized in the form $p(D\,|\,\boldsymbol{\omega}) = f(\boldsymbol{\omega}, \boldsymbol{t})g(D)$. A sufficient statistic always exists, for $\boldsymbol{t}(D) = D$ is obviously sufficient; however, a much simpler sufficient statistic, with a fixed dimensionality which is independent of the sample size, often exists. In fact this is known to be the case whenever the probability model belongs to the *generalized exponential family*, which includes many of the more frequently used probability models. It is easily established that if $\boldsymbol{t}$ is sufficient, the posterior distribution of $\boldsymbol{\omega}$ only depends on the data D through $\boldsymbol{t}(D)$, and may be directly computed in terms of $p(\boldsymbol{t}\,|\,\boldsymbol{\omega})$, so that, $p(\boldsymbol{\omega}\,|\,D) = p(\boldsymbol{\omega}\,|\,\boldsymbol{t}) \propto p(\boldsymbol{t}\,|\,\boldsymbol{\omega})\, p(\boldsymbol{\omega})$.

Naturally, for fixed data and model assumptions, different priors lead to different posteriors. Indeed, Bayes' theorem may be described as a data-driven probability transformation machine which maps prior distributions (describing prior knowledge) into posterior distributions (representing combined prior and data knowledge). It is important to analyze whether or not sensible changes in the prior would induce noticeable changes in the posterior. Posterior distributions based on reference "noninformative" priors play a central role in this ►*sensitivity analysis* context. Investigation of the sensitivity of the posterior to changes in the prior is an important ingredient of the comprehensive analysis of the sensitivity of the final results to *all* accepted assumptions which any responsible statistical study should contain.

Example 2 (Inference on a binomial parameter). If the data D consist of n Bernoulli observations with parameter θ which contain r positive trials, then $p(D\,|\,\theta, n) = \theta^r(1-\theta)^{n-r}$, so that $\boldsymbol{t}(D) = \{r, n\}$ is sufficient. Suppose that prior knowledge about θ may be approximately described by a ►Beta distribution $\mathrm{Be}(\theta\,|\,\alpha, \beta)$, so that $p(\theta\,|\,\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$. Using Bayes' theorem, the posterior density of θ is $p(\theta\,|\,r, n, \alpha, \beta) \propto \theta^r(1-\theta)^{n-r}\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1}$, the Beta distribution $\mathrm{Be}(\theta\,|\,r+\alpha, n-r+\beta)$.

Suppose, for example, that in the light of precedent surveys, available information on the proportion θ of citizens who would vote for a particular political measure in a referendum is described by a Beta distribution $\mathrm{Be}(\theta\,|\,50, 50)$, so that it is judged to be equally likely that the referendum would be won or lost, and it is judged that the probability that either side wins less than 60% of the vote is 0.95.

A random survey of size 1500 is then conducted, where only 720 citizens declare to be in favor of the proposed measure. Using the results above, the corresponding posterior distribution is then $\mathrm{Be}(\theta\,|\,770, 830)$. These prior and posterior densities are plotted in Fig. 2; it may be appreciated that, as one would expect, the effect of the data is to drastically reduce the initial uncertainty on the value of θ



Bayesian Statistics. Fig. 2 Prior and posterior densities of the proportion θ of citizens that would vote in favor of a referendum text

and, hence, on the referendum outcome. More precisely, $\Pr(\theta < 0.5 \,|\, 720, 1, 500, H, K) = 0.933$ (shaded region in Fig. 2) so that, after the information from the survey has been included, the probability that the referendum will be lost should be judged to be about 93%.

The general situation where the vector of interest is not the whole parameter vector $\boldsymbol{\omega}$, but some function $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ of possibly lower dimension than $\boldsymbol{\omega}$, will now be considered. Let D be some observed data, let $\{p(D \,|\, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$ be a probability model assumed to describe the probability mechanism which has generated D, let $p(\boldsymbol{\omega})$ be a probability distribution describing any available information on the value of $\boldsymbol{\omega}$, and let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \Theta$ be a function of the original parameters over whose value inferences based on the data D are required. Any valid conclusion on the value of the *vector of interest* $\boldsymbol{\theta}$ will then be contained in its posterior probability distribution $p(\boldsymbol{\theta} \,|\, D)$ which is conditional on the observed data D and will naturally also depend, although not explicitly shown in the notation, on the assumed model $\{p(D \,|\, \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, and on the available prior information encapsulated by $p(\boldsymbol{\omega})$. The required posterior distribution $p(\boldsymbol{\theta} \,|\, D)$ is found by standard use of probability calculus. Indeed, by Bayes' theorem, $p(\boldsymbol{\omega} \,|\, D) \propto p(D \,|\, \boldsymbol{\omega})\, p(\boldsymbol{\omega})$. Moreover, let $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\omega}) \in \boldsymbol{\Lambda}$ be some other function of the original parameters such that $\boldsymbol{\psi} = \{\boldsymbol{\theta}, \boldsymbol{\lambda}\}$ is a one-to-one transformation of $\boldsymbol{\omega}$, and let $J(\boldsymbol{\omega}) = (\partial \boldsymbol{\psi} / \partial \boldsymbol{\omega})$ be the corresponding Jacobian matrix. Naturally, the introduction of $\boldsymbol{\lambda}$ is not necessary if $\boldsymbol{\theta}(\boldsymbol{\omega})$ is a one-to-one transformation of $\boldsymbol{\omega}$. Using standard change-of-variable probability techniques, the posterior density of $\boldsymbol{\psi}$ is

$$p(\boldsymbol{\psi} \,|\, D) = p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, D) = \left[\frac{p(\boldsymbol{\omega} \,|\, D)}{|J(\boldsymbol{\omega})|} \right]_{\boldsymbol{\omega} = \boldsymbol{\omega}(\boldsymbol{\psi})}$$

and the required posterior of $\boldsymbol{\theta}$ is the appropriate *marginal* density, obtained by integration over the *nuisance parameter* $\boldsymbol{\lambda}$,

$$p(\boldsymbol{\theta} \,|\, D) = \int_{\boldsymbol{\Lambda}} p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, D)\, d\boldsymbol{\lambda}.$$

Notice that elimination of unwanted nuisance parameters, a simple integration within the Bayesian paradigm is, however, a difficult (often polemic) problem for conventional statistics.

Sometimes, the range of possible values of $\boldsymbol{\omega}$ is effectively restricted by contextual considerations. If $\boldsymbol{\omega}$ is known to belong to $\boldsymbol{\Omega}_c \subset \boldsymbol{\Omega}$, the prior distribution is only positive in $\boldsymbol{\Omega}_c$ and, using Bayes' theorem, it is immediately found that the restricted posterior is

$$p(\boldsymbol{\omega} \,|\, D, \boldsymbol{\omega} \in \boldsymbol{\Omega}_c) = \frac{p(\boldsymbol{\omega} \,|\, D)}{\int_{\boldsymbol{\Omega}_c} p(\boldsymbol{\omega} \,|\, D)}, \quad \boldsymbol{\omega} \in \boldsymbol{\Omega}_c,$$

and obviously vanishes if $\boldsymbol{\omega} \notin \boldsymbol{\Omega}_c$. Thus, to incorporate a restriction on the possible values of the parameters, it suffices to *renormalize* the unrestricted posterior distribution to the set $\boldsymbol{\Omega}_c \subset \boldsymbol{\Omega}$ of parameter values which satisfy the required condition. Incorporation of known constraints on the parameter values, a simple renormalization within the Bayesian pardigm, is another very difficult problem for conventional statistics.

Example 3 (Inference on normal parameters). Let $D = \{x_1, \ldots x_n\}$ be a random sample from a normal distribution $N(x \,|\, \mu, \sigma)$. The corresponding likelihood function is immediately found to be proportional to $\sigma^{-n} \exp[-n\{s^2 + (\bar{x} - \mu)^2\}/(2\sigma^2)]$, with $n\bar{x} = \sum_i x_i$, and $ns^2 = \sum_i (x_i - \bar{x})^2$. It may be shown (see section "►Reference Analysis") that absence of initial information on the value of both μ and σ may formally be described by a joint prior function which is uniform in both μ and $\log(\sigma)$, that is, by the (improper) prior function $p(\mu, \sigma) = \sigma^{-1}$. Using Bayes' theorem, the corresponding joint posterior is

$$p(\mu, \sigma \,|\, D) \propto \sigma^{-(n+1)} \exp[-n\{s^2 + (\bar{x} - \mu)^2\}/(2\sigma^2)].$$

Thus, using the Gamma integral in terms of $\lambda = \sigma^{-2}$ to integrate out σ,

$$p(\mu \,|\, D) \propto \int_0^{\infty} \sigma^{-(n+1)} \exp\left[-\frac{n}{2\sigma^2} [s^2 + (\bar{x} - \mu)^2] \right] d\sigma \propto [s^2 + (\bar{x} - \mu)^2]^{-n/2},$$

which is recognized as a kernel of the Student density $\mathrm{St}\left(\mu \,|\, \bar{x}, s/\sqrt{n-1}, n-1\right)$. Similarly, integrating out μ,

$$p(\sigma \,|\, D) \propto \int_{-\infty}^{\infty} \sigma^{-(n+1)} \exp\left[-\frac{n}{2\sigma^2}\left[s^2 + (\bar{x}-\mu)^2\right]\right] d\mu \propto \sigma^{-n} \exp\left[-\frac{ns^2}{2\sigma^2}\right].$$

Changing variables to the precision $\lambda = \sigma^{-2}$ results in $p(\lambda \,|\, D) \propto \lambda^{(n-3)/2} e^{ns^2\lambda/2}$, a kernel of the Gamma density $\mathrm{Ga}(\lambda \,|\, (n-1)/2, ns^2/2)$. In terms of the standard deviation σ this becomes $p(\sigma \,|\, D) = p(\lambda \,|\, D)|\partial\lambda/\partial\sigma| = 2\sigma^{-3}\mathrm{Ga}(\sigma^{-2} \,|\, (n-1)/2, ns^2/2)$, a square-root inverted gamma density.

A frequent example of this scenario is provided by laboratory measurements made in conditions where central limit conditions apply, so that (assuming no experimental bias) those measurements may be treated as a random sample from a normal distribution centered at the quantity μ which is being measured, and with some (unknown) standard deviation σ. Suppose, for example, that in an elementary physics classroom experiment to measure the gravitational field g with a pendulum, a student has obtained $n = 20$ measurements of g yielding (in m/s^2) a mean $\bar{x} = 9.8087$, and a standard deviation $s = 0.0428$. Using no other information, the corresponding posterior distribution is $p(g \,|\, D) = \mathrm{St}(g \,|\, 9.8087, 0.0098, 19)$ represented in the upper panel of Fig. 3. In particular, $\Pr(9.788 < g < 9.829 \,|\, D) = 0.95$, so that, with the information provided by this experiment, the value of g at the location of the laboratory may be expected to lie between 9.788 and 9.829 with probability 0.95.

Formally, the posterior distribution of g should be restricted to $g > 0$; however, as immediately obvious from Fig. 3, this would not have any appreciable effect, due to the fact that the likelihood function is actually concentrated on positive g values.

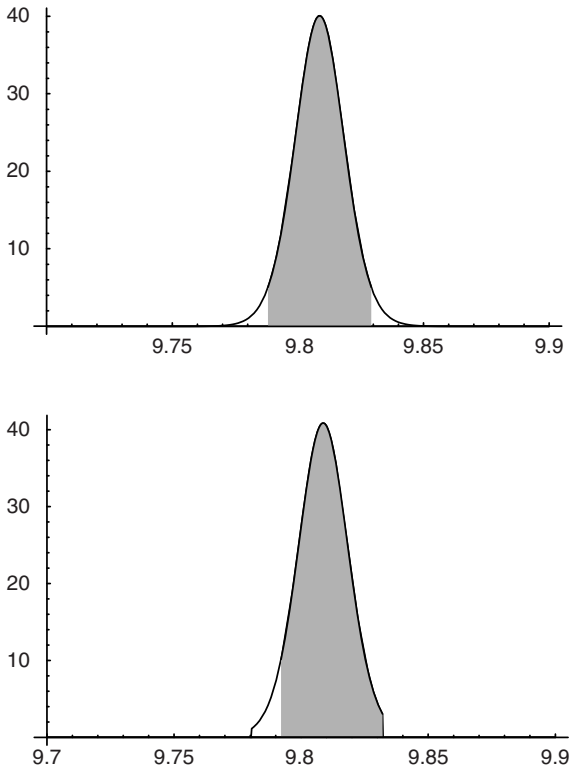
Suppose now that the student is further instructed to incorporate into the analysis the fact that the value of the gravitational field g at the laboratory is known to lie between 9.7803 m/s^2 (average value at the Equator) and 9.8322 m/s^2 (average value at the poles). The updated posterior distribution will then be

$$p(g \,|\, D, g \in G_c) = \frac{\mathrm{St}\left(g \,|\, m, s/\sqrt{n-1}, n\right)}{\int_{g \in G_c} \mathrm{St}\left(g \,|\, m, s/\sqrt{n-1}, n\right)}, \quad g \in G_c,$$

represented in lower panel of Fig. 3, where $G_c = \{g; 9.7803 < g < 9.8322\}$. Simple ▶numerical integration may be used to verify that $\Pr(g > 9.792 \,|\, D, g \in G_c) = 0.95$. Moreover, if inferences about the standard deviation σ of the measurement procedure are also requested, the corresponding posterior distribution is found to be $p(\sigma \,|\, D) = 2\sigma^{-3}\mathrm{Ga}(\sigma^{-2} \,|\, 9.5, 0.0183)$. This has a mean $\mathrm{E}[\sigma \,|\, D] = 0.0458$ and yields $\Pr(0.0334 < \sigma < 0.0642 \,|\, D) = 0.95$.

Predictive Distributions

Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathcal{X}$, be a set of exchangeable observations, and consider now a situation where it is desired to predict the value of a future observation $\boldsymbol{x} \in \mathcal{X}$ generated by the same random mechanism that has generated the data D. It follows from the foundations arguments discussed in section "▶Foundations" that the solution to this prediction problem is simply encapsulated by the *predictive* distribution $p(\boldsymbol{x} \,|\, D)$ describing the uncertainty on the value that $\boldsymbol{x}$ will take, given the information provided by D and any other available knowledge. Suppose that contextual information suggests the assumption that data D may be considered to be a random sample from a distribution in the family $\{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, and let $p(\boldsymbol{\omega})$ be a prior distribution describing available information on the value of $\boldsymbol{\omega}$. Since $p(\boldsymbol{x} \,|\, \boldsymbol{\omega}, D) = p(\boldsymbol{x} \,|\, \boldsymbol{\omega})$, it then



Bayesian Statistics. Fig. 3 Posterior densities $p(g \,|\, m, s, n)$ of the value g of the gravitational field

follows from standard probability theory that $p(\boldsymbol{x}\,|\,D) = \int_{\Omega} p(\boldsymbol{x}\,|\,\boldsymbol{\omega})\, p(\boldsymbol{\omega}\,|\,D)\, d\boldsymbol{\omega}$, which is an average of the probability distributions of $\boldsymbol{x}$ conditional on the (unknown) value of $\boldsymbol{\omega}$, weighted with the posterior distribution of $\boldsymbol{\omega}$ given D.

If the assumptions on the probability model are correct, the posterior predictive distribution $p(\boldsymbol{x}\,|\,D)$ will converge, as the sample size increases, to the distribution $p(\boldsymbol{x}\,|\,\boldsymbol{\omega})$ which has generated the data. Indeed, the best technique to assess the quality of the inferences about $\boldsymbol{\omega}$ encapsulated in $p(\boldsymbol{\omega}\,|\,D)$ is to check against the observed data the predictive distribution $p(\boldsymbol{x}\,|\,D)$ generated by $p(\boldsymbol{\omega}\,|\,D)$.

Example 4 (Prediction in a Poisson process). Let $D = \{r_1, \ldots, r_n\}$ be a random sample from a Poisson distribution $\text{Po}(r\,|\,\lambda)$ with parameter λ, so that $p(D\,|\,\lambda) \propto \lambda^t e^{-\lambda n}$, where $t = \sum r_i$. It may be shown (see section "►Reference Analysis") that absence of initial information on the value of λ may be formally described by the (improper) prior function $p(\lambda) = \lambda^{-1/2}$. Using Bayes' theorem, the corresponding posterior is

$$p(\lambda\,|\,D) \propto \lambda^t e^{-\lambda n}\, \lambda^{-1/2} \propto \lambda^{t-1/2} e^{-\lambda n},$$

the kernel of a Gamma density $\text{Ga}(\lambda\,|\,, t + 1/2, n)$, with mean $(t + 1/2)/n$. The corresponding predictive distribution is the Poisson-Gamma mixture

$$p(r\,|\,D) = \int_0^{\infty} \text{Po}(r\,|\,\lambda)\, \text{Ga}\left(\lambda\,|\,, t + \frac{1}{2}, n\right) d\lambda$$
$$= \frac{n^{t+1/2}}{\Gamma(t+1/2)}\, \frac{1}{r!}\, \frac{\Gamma(r+t+1/2)}{(1+n)^{r+t+1/2}}.$$

Suppose, for example, that in a firm producing automobile restraint systems, the entire production in each of 10 consecutive months has yielded no complaint from their clients. With no additional information on the average number λ of complaints per month, the quality assurance department of the firm may report that the probabilities that r complaints will be received in the next month of production are given by the last equation, with $t = 0$ and $n = 10$. In particular, $p(r = 0\,|\,D) = 0.953$, $p(r = 1\,|\,D) = 0.043$, and $p(r = 2\,|\,D) = 0.003$. Many other situations may be described with the same model. For instance, if meteorological conditions remain similar in a given area, $p(r = 0\,|\,D) = 0.953$ would describe the chances of no flash flood next year, given 10 years without flash floods in the area.

Example 5 (Prediction in a Normal process). Consider now prediction of a continuous variable. Let $D = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $N(x\,|\,\mu, \sigma)$. As mentioned in Example 3, absence of initial information on the values of both μ and σ is formally described by the *improper* prior function $p(\mu, \sigma) = \sigma^{-1}$, and this leads to the joint posterior density describe above. The corresponding (posterior) predictive distribution is

$$p(x\,|\,D) = \int_0^{\infty} \int_{-\infty}^{\infty} \text{N}(x\,|\,\mu, \sigma)\, p(\mu, \sigma\,|\,D)\, d\mu d\sigma$$
$$= \text{St}\left(x\,|\,\overline{x}, s\sqrt{\frac{n+1}{n-1}}, n-1\right).$$

If μ is known to be positive, the appropriate prior function will be the restricted function $p(\mu, \sigma) = \sigma^{-1}$ if $\mu > 0$ and $p(\mu, \sigma) = 0$ otherwise. However, the result will still hold, provided the likelihood function $p(D\,|\,\mu, \sigma)$ is concentrated on positive μ values. Suppose, for example, that in the firm producing automobile restraint systems, the observed breaking strengths of $n = 10$ randomly chosen safety belt webbings have mean $\overline{x} = 28.011$ kN and standard deviation $s = 0.443$ kN, and that the relevant engineering specification requires breaking strengths to be larger than 26 kN. If data may truly be assumed to be a random sample from a normal distribution, the likelihood function is only appreciable for positive μ values, and only the information provided by this small sample is to be used, then the quality engineer may claim that the probability that a safety belt randomly chosen from the same batch as the sample tested would satisfy the required specification is $\Pr(x > 26\,|\,D) = 0.9987$. Besides, if conditions remain constant, 99.87% of the safety belt webbings may be expected to have acceptable breaking strengths.

Asymptotic Behavior

The behavior of posterior distributions when the sample size is large is now considered. This is important for, at least, two different reasons: (1) asymptotic results provide useful first-order approximations when actual samples are relatively large, and (2) objective Bayesian methods typically depend on the asymptotic properties of the assumed model. Let $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x} \in \mathcal{X}$, be a random sample of size n from $\{p(\boldsymbol{x}\,|\,\boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$. It may be shown that, as $n \to \infty$, the posterior distribution $p(\boldsymbol{\omega}\,|\,D)$ of a *discrete* parameter $\boldsymbol{\omega}$ typically converges to a degenerate distribution which gives probability one to the true value of $\boldsymbol{\omega}$, and that the posterior distribution of a *continuous* parameter $\boldsymbol{\omega}$ typically converges to a normal distribution centered at its *maximum likelihood estimate* $\hat{\boldsymbol{\omega}}$ (MLE), with a variance matrix which decreases with n as $1/n$.

Consider first the situation where $\boldsymbol{\Omega} = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots\}$ consists of a *countable* (possibly infinite) set of values, such that the probability model which corresponds to the true parameter value $\boldsymbol{\omega}_t$ is *distinguishable* from

the others in the sense that the logarithmic divergence $\delta\{p(\boldsymbol{x}\,|\,\boldsymbol{\omega}_i)\,|\,p(\boldsymbol{x}\,|\,\boldsymbol{\omega}_t)\}$ of each of the $p(\boldsymbol{x}\,|\,\boldsymbol{\omega}_i)$ from $p(\boldsymbol{x}\,|\,\boldsymbol{\omega}_t)$ is strictly positive. Taking logarithms in Bayes' theorem, defining $z_j = \log[p(\boldsymbol{x}_j\,|\,\boldsymbol{\omega}_i)/p(\boldsymbol{x}_j\,|\,\boldsymbol{\omega}_t)]$, $j = 1,\ldots,n$, and using the strong law of large numbers on the n conditionally independent and identically distributed random quantities $z_1,\ldots,z_n$, it may be shown that

$$\lim_{n\to\infty} p(\boldsymbol{\omega}_t\,|\,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = 1,$$
$$\lim_{n\to\infty} p(\boldsymbol{\omega}_i\,|\,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = 0, \quad i \neq t.$$

Thus, under appropriate regularity conditions, the posterior probability of the true parameter value converges to one as the sample size grows.

Consider now the situation where $\boldsymbol{\omega}$ is a k-dimensional *continuous* parameter. Expressing Bayes' theorem as $p(\boldsymbol{\omega}\,|\,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \propto \exp\left\{\log[p(\boldsymbol{\omega})] + \sum_{j=1}^{n}\log[p(\boldsymbol{x}_j\,|\,\boldsymbol{\omega})]\right\}$, expanding $\sum_j \log[p(\boldsymbol{x}_j\,|\,\boldsymbol{\omega})]$ about its maximum (the MLE $\hat{\boldsymbol{\omega}}$), and assuming regularity conditions (to ensure that terms of order higher than quadratic may be ignored and that the sum of the terms from the likelihood will dominate the term from the prior) it is found that the posterior density of $\boldsymbol{\omega}$ is the approximate k-variate normal

$$p(\boldsymbol{\omega}\,|\,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \approx \mathrm{N}_k\{\hat{\boldsymbol{\omega}}, \boldsymbol{S}(D,\hat{\boldsymbol{\omega}})\},$$
$$\boldsymbol{S}^{-1}(D,\boldsymbol{\omega}) = \left(-\sum_{l=1}^{n}\frac{\partial^2\log[p(\boldsymbol{x}_l\,|\,\boldsymbol{\omega})]}{\partial\omega_i\partial\omega_j}\right).$$

A simpler, but somewhat poorer, approximation may be obtained by using the strong law of large numbers on the sums above to establish that $\boldsymbol{S}^{-1}(D,\hat{\boldsymbol{\omega}}) \approx n\boldsymbol{F}(\hat{\boldsymbol{\omega}})$, where $\boldsymbol{F}(\boldsymbol{\omega})$ is *Fisher's information matrix*, with general element

$$\boldsymbol{F}_{ij}(\boldsymbol{\omega}) = -\int_X p(\boldsymbol{x}\,|\,\boldsymbol{\omega})\,\frac{\partial^2\log[p(\boldsymbol{x}\,|\,\boldsymbol{\omega})]}{\partial\omega_i\partial\omega_j}\,d\boldsymbol{x},$$

so that

$$p(\boldsymbol{\omega}\,|\,\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \approx N_k(\boldsymbol{\omega}\,|\,\hat{\boldsymbol{\omega}}, n^{-1}\boldsymbol{F}^{-1}(\hat{\boldsymbol{\omega}})).$$

Thus, under appropriate regularity conditions, the posterior probability density of the parameter vector $\boldsymbol{\omega}$ approaches, as the sample size grows, a multivarite normal density centered at the MLE $\hat{\boldsymbol{\omega}}$, with a variance matrix which decreases with n as n^{-1}.

Example 2 (Inference on a binomial parameter, continued). Let $D = (x_1,\ldots,x_n)$ consist of n independent Bernoulli trials with parameter θ, so that $p(D\,|\,\theta,n) = \theta^r(1-\theta)^{n-r}$. This likelihood function is maximized at $\hat{\theta} = r/n$, and Fisher's information function is $F(\theta) = \theta^{-1}(1-\theta)^{-1}$. Thus, using the results above, the posterior distribution of θ will be the approximate normal,

$$p(\theta\,|\,r,n) \approx \mathrm{N}\left(\theta\,|\,\hat{\theta}, s(\hat{\theta})/\sqrt{n}\right), \quad s(\theta) = \{\theta(1-\theta)\}^{1/2}$$

with mean $\hat{\theta} = r/n$ and variance $\hat{\theta}(1-\hat{\theta})/n$. This will provide a reasonable approximation to the exact posterior if (1) the prior $p(\theta)$ is relatively "flat" in the region where the likelihood function matters, and (2) both r and n are moderately large. If, say, $n = 1{,}500$ and $r = 720$, this leads to $p(\theta\,|\,D) \approx \mathrm{N}(\theta\,|\,0.480, 0.013)$, and to $\Pr(\theta > 0.5\,|\,D) \approx 0.940$, which may be compared with the exact value $\Pr(\theta > 0.5\,|\,D) = 0.933$ obtained from the posterior distribution which corresponds to the prior $\mathrm{Be}(\theta\,|\,50,50)$.

It follows from the *joint* posterior asymptotic behavior of $\boldsymbol{\omega}$ and from the properties of the multivariate normal distribution (see ▶Multivariate Normal Distributions) that, if the parameter vector is decomposed into $\boldsymbol{\omega} = (\boldsymbol{\theta},\boldsymbol{\lambda})$, and Fisher's information matrix is correspondingly partitioned, so that

$$\boldsymbol{F}(\boldsymbol{\omega}) = \boldsymbol{F}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \begin{pmatrix} \boldsymbol{F}_{\theta\theta}(\boldsymbol{\theta},\boldsymbol{\lambda}) & \boldsymbol{F}_{\theta\lambda}(\boldsymbol{\theta},\boldsymbol{\lambda}) \\ \boldsymbol{F}_{\lambda\theta}(\boldsymbol{\theta},\boldsymbol{\lambda}) & \boldsymbol{F}_{\lambda\lambda}(\boldsymbol{\theta},\boldsymbol{\lambda}) \end{pmatrix}$$

and

$$\boldsymbol{S}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \boldsymbol{F}^{-1}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \begin{pmatrix} \boldsymbol{S}_{\theta\theta}(\boldsymbol{\theta},\boldsymbol{\lambda}) & \boldsymbol{S}_{\theta\lambda}(\boldsymbol{\theta},\boldsymbol{\lambda}) \\ \boldsymbol{S}_{\lambda\theta}(\boldsymbol{\theta},\boldsymbol{\lambda}) & \boldsymbol{S}_{\lambda\lambda}(\boldsymbol{\theta},\boldsymbol{\lambda}) \end{pmatrix},$$

then the *marginal* posterior distribution of $\boldsymbol{\theta}$ will be

$$p(\boldsymbol{\theta}\,|\,D) \approx \mathrm{N}\{\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}}, n^{-1}\boldsymbol{S}_{\theta\theta}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\lambda}})\},$$

while the *conditional* posterior distribution of $\boldsymbol{\lambda}$ given $\boldsymbol{\theta}$ will be

$$p(\boldsymbol{\lambda}\,|\,\boldsymbol{\theta},D) \approx \mathrm{N}\left\{\boldsymbol{\lambda}\,|\,\hat{\boldsymbol{\lambda}} - \boldsymbol{F}_{\lambda\lambda}^{-1}(\boldsymbol{\theta},\hat{\boldsymbol{\lambda}})\right.$$
$$\left.\boldsymbol{F}_{\lambda\theta}(\boldsymbol{\theta},\hat{\boldsymbol{\lambda}})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}),\, n^{-1}\boldsymbol{F}_{\lambda\lambda}^{-1}(\boldsymbol{\theta},\hat{\boldsymbol{\lambda}})\right\}.$$

Notice that $\boldsymbol{F}_{\lambda\lambda}^{-1} = \boldsymbol{S}_{\lambda\lambda}$ if (and only if) $\boldsymbol{F}$ is block diagonal, i.e., if (and only if) $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are asymptotically independent.

Example 3 (Inference on normal parameters, continued). Let $D = (x_1,\ldots,x_n)$ be a random sample from a normal distribution $\mathrm{N}(x\,|\,\mu,\sigma)$. The corresponding likelihood function $p(D\,|\,\mu,\sigma)$ is maximized at $(\hat{\mu},\hat{\sigma}) = (\bar{x},s)$, and Fisher's information matrix is diagonal, with $F_{\mu\mu} = \sigma^{-2}$. Hence, the posterior distribution of μ is *approximately* $\mathrm{N}\left(\mu\,|\,\bar{x}, s/\sqrt{n}\right)$; this may be compared with the *exact* result $p(\mu\,|\,D) = \mathrm{St}\left(\mu\,|\,\bar{x}, s/\sqrt{n-1}, n-1\right)$ obtained previously under the assumption of no prior knowledge.

Inference Summaries

From a Bayesian viewpoint, the final outcome of a problem of inference about *any* unknown quantity is nothing but the corresponding posterior distribution. Thus, given some data D and conditions C, *all* that can be said about any function $\boldsymbol{\omega}$ of the parameters which govern the model is contained in the posterior distribution $p(\boldsymbol{\omega}\,|\,D,C)$, and

all that can be said about some function $\boldsymbol{y}$ of future observations from the same model is contained in its posterior predictive distribution $p(\boldsymbol{y}\,|\,D, C)$. As mentioned before, Bayesian inference may technically be described as a decision problem where the space of available actions is the class of those posterior probability distributions of the quantity of interest which are compatible with accepted assumptions.

However, to make it easier for the user to assimilate the appropriate conclusions, it is often convenient to *summarize* the information contained in the posterior distribution by (1) providing values of the quantity of interest which, in the light of the data, are likely to be "close" to its true value and by (2) measuring the compatibility of the results with hypothetical values of the quantity of interest which might have been suggested in the context of the investigation. In this section, those Bayesian counterparts of traditional *estimation* and *hypothesis testing* problems are briefly considered.

Estimation

In one or two dimensions, a graph of the posterior probability density of the quantity of interest (or the probability mass function in the discrete case) immediately conveys an intuitive, "impressionist" summary of the main conclusions which may possibly be drawn on its value. Indeed, this is greatly appreciated by users, and may be quoted as an important asset of Bayesian methods. From a plot of its posterior density, the region where (given the data) a univariate quantity of interest is likely to lie is easily distinguished. For instance, all important conclusions about the value of the gravitational field in Example 3 are qualitatively available from Fig. 3. However, this does not easily extend to more than two dimensions and, besides, *quantitative* conclusions (in a simpler form than that provided by the mathematical expression of the posterior distribution) are often required.

Point Estimation. Let D be the available data, which are assumed to have been generated by a probability model $\{p(D\,|\,\boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$, and let $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ be the quantity of interest. A *point estimator* of $\boldsymbol{\theta}$ is some function of the data $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(D)$ which could be regarded as an appropriate proxy for the actual, unknown value of $\boldsymbol{\theta}$. Formally, to choose a point estimate for $\boldsymbol{\theta}$ is a *decision problem*, where the action space is the class $\boldsymbol{\Theta}$ of possible $\boldsymbol{\theta}$ values. From a decision-theoretic perspective, to choose a point estimate $\tilde{\boldsymbol{\theta}}$ of some quantity $\boldsymbol{\theta}$ is a *decision* to act as if the true value of $\boldsymbol{\theta}$ were $\tilde{\boldsymbol{\theta}}$, not to assert something about the value of $\boldsymbol{\theta}$ (although desire to assert something simple may well be the reason to obtain an estimate). As prescribed by the foundations of decision theory (section "▶Foundations"), to solve this decision problem it is necessary to specify a *loss function* $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ measuring the consequences of acting *as if* the true value of the quantity of interest were $\tilde{\boldsymbol{\theta}}$, when it is actually $\boldsymbol{\theta}$. The expected posterior loss if $\tilde{\boldsymbol{\theta}}$ were used is

$$\overline{\ell}[\tilde{\boldsymbol{\theta}}\,|\,D] = \int_{\Theta} \ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}\,|\,D)\, d\boldsymbol{\theta},$$

and the corresponding *Bayes estimator* $\boldsymbol{\theta}^*$ is that function of the data, $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(D)$, which minimizes this expectation.

Example 6 (Conventional Bayes estimators). For any given model and data, the Bayes estimator obviously depends on the chosen ▶loss function. The loss function is context specific, and should be chosen in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for those situations where no particular uses are envisaged. These loss functions produce estimates which may be regarded as simple descriptions of the *location* of the posterior distribution. For example, if the loss function is quadratic, so that $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$, then the Bayes estimator is the *posterior mean* $\boldsymbol{\theta}^* = E[\boldsymbol{\theta}\,|\,D]$, assuming that the mean exists. Similarly, if the loss function is a zero-one function, so that $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0$ if $\tilde{\boldsymbol{\theta}}$ belongs to a ball or radius ϵ centered in $\boldsymbol{\theta}$ and $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 1$ otherwise, then the Bayes estimator $\boldsymbol{\theta}^*$ tends to the *posterior mode* as the ball radius ϵ tends to zero, assuming that a unique mode exists. If θ is univariate and the loss function is linear, so that $\ell(\tilde{\theta}, \theta) = c_1(\tilde{\theta} - \theta)$ if $\tilde{\theta} \geq \theta$, and $\ell(\tilde{\theta}, \theta) = c_2(\theta - \tilde{\theta})$ otherwise, then the Bayes estimator is the *posterior quantile* of order $c_2/(c_1 + c_2)$, so that $\Pr[\theta < \theta^*] = c_2/(c_1 + c_2)$. In particular, if $c_1 = c_2$, the Bayes estimator is the *posterior median*. The results derived for linear loss functions clearly illustrate the fact that *any* possible parameter value may turn out be the Bayes estimator: it all depends on the loss function describing the consequences of the anticipated uses of the estimate.

Example 7 (Intrinsic estimation). Conventional loss functions are typically non-invariant under reparametrization, so that the Bayes estimator $\boldsymbol{\phi}^*$ of a one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ of the original parameter $\boldsymbol{\theta}$ is not necessarily $\boldsymbol{\phi}(\boldsymbol{\theta}^*)$ (the *univariate* posterior median, which *is* invariant, is an interesting exception). Moreover, conventional loss functions focus on the "distance" between the estimate $\tilde{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}$, rather than on the "distance" between the probability models they label. Intrinsic losses directly focus on how different the probability *model* $p(D\,|\,\boldsymbol{\theta}, \boldsymbol{\lambda})$ is from its closest approximation within the family $\{p(D\,|\,\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}_i), \boldsymbol{\lambda}_i \in \Lambda\}$, and typically produce invariant solutions. An attractive example is the *intrinsic discrepancy*, $d(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$ defined as the minimum logarithmic divergence between a probability model labeled by $\boldsymbol{\theta}$

and a probability model labeled by $\tilde{\boldsymbol{\theta}}$. When there are no nuisance parameters, this is given by

$$d(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \min\{\delta(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta}), \delta(\boldsymbol{\theta} \,|\, \tilde{\boldsymbol{\theta}})\},$$
$$\delta(\boldsymbol{\theta}_i \,|\, \boldsymbol{\theta}) = \int_{\mathcal{T}} p(\boldsymbol{t} \,|\, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{t} \,|\, \boldsymbol{\theta})}{p(\boldsymbol{t} \,|\, \boldsymbol{\theta}_i)} \, d\boldsymbol{t},$$

where $\boldsymbol{t} = \boldsymbol{t}(D) \in \mathcal{T}$ is *any* sufficient statistic (which may well be the whole data set D). The definition is easily extended to problems with nuisance parameters; in this case,

$$d(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \min_{\boldsymbol{\lambda}_i \in \boldsymbol{\Lambda}} d(\tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}_i, \boldsymbol{\theta}, \boldsymbol{\lambda})$$

measures the logarithmic divergence from $p(\boldsymbol{t} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ of its closest approximation with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, and the loss function now depends on the complete parameter vector $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Although not explicitly shown in the notation, the intrinsic discrepancy function typically depends on the sample size n; indeed, when the data consist of a random sample $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from some model $p(\boldsymbol{x} \,|\, \boldsymbol{\theta})$ then

$$\delta(\boldsymbol{\theta}_i \,|\, \boldsymbol{\theta}) = n \int_{\mathcal{X}} p(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x} \,|\, \boldsymbol{\theta})}{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}_i)} \, d\boldsymbol{x},$$

so that the discrepancy associated with the full model is simply n times the discrepancy which corresponds to a single observation. The intrinsic discrepancy is a symmetric, non-negative loss function with a direct interpretation in information-theoretic terms as the minimum amount of information which is expected to be necessary to distinguish between the model $p(D \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and its closest approximation within the class $\{p(D \,|\, \tilde{\boldsymbol{\theta}}, \boldsymbol{\lambda}_i), \boldsymbol{\lambda}_i \in \boldsymbol{\Lambda}\}$. Moreover, it is invariant under one-to-one reparametrization of the parameter of interest $\boldsymbol{\theta}$, and does not depend on the choice of the nuisance parameter $\boldsymbol{\lambda}$. The *intrinsic estimator* is naturally obtained by minimizing the posterior expected intrinsic discrepancy

$$\overline{d}(\tilde{\boldsymbol{\theta}} \,|\, D) = \int_{\boldsymbol{\Lambda}} \int_{\Theta} d(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}, \boldsymbol{\lambda}) p(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, D) \, d\boldsymbol{\theta} d\boldsymbol{\lambda}.$$

Since the intrinsic discrepancy is invariant under reparametrization, minimizing its posterior expectation produces invariant estimators.

Example 2 (Inference on a binomial parameter, continued). In the estimation of a binomial proportion θ, given data $D = (n, r)$ and a Beta prior $\text{Be}(\theta \,|\, \alpha, \beta)$, the Bayes estimator associated with the quadratic loss (the corresponding posterior mean) is $E[\theta \,|\, D] = (r+\alpha)/(n+\alpha+\beta)$, while the quadratic loss based estimator of, say, the log-odds $\phi(\theta) = \log[\theta/(1-\theta)]$, is $E[\phi \,|\, D] = \psi(r+\alpha) - \psi(n-r+\beta)$ (where $\psi(x) = d\log[\Gamma(x)]/dx$ is the *digamma* function), which is *not* equal to $\phi(E[\theta \,|\, D])$. The intrinsic loss function in this problem is

$$d(\tilde{\theta}, \theta) = n \min\{\delta(\tilde{\theta} \,|\, \theta), \delta(\theta \,|\, \tilde{\theta})\},$$
$$\delta(\theta_i \,|\, \theta) = \theta \log \frac{\theta}{\theta_i} + (1-\theta) \log \frac{1-\theta}{1-\theta_i},$$

and the corresponding intrinsic estimator θ^* is obtained by minimizing the expected posterior loss $\overline{d}(\tilde{\theta} \,|\, D) = \int d(\tilde{\theta}, \theta) \, p(\theta \,|\, D) \, d\theta$. The exact value of θ^* may be obtained by numerical minimization, but a very good approximation is given by

$$\theta^* \approx \frac{1}{2} \frac{r+\alpha}{n+\alpha+\beta} + \frac{1}{2} \frac{e^{\psi(r+\alpha)}}{e^{\psi(r+\alpha)} + e^{\psi(n-r+\beta)}}.$$

Since intrinsic estimation is an invariant procedure, the intrinsic estimator of the log-odds will simply be the log-odds of the intrinsic estimator of θ. As one would expect, when $r+\alpha$ and $n-r+\beta$ are both large, all Bayes estimators of any well-behaved function $\phi(\theta)$ will cluster around $\phi(E[\theta \,|\, D])$.

Interval Estimation. To describe the inferential content of the posterior distribution of the quantity of interest $p(\boldsymbol{\theta} \,|\, D)$ it is often convenient to quote regions $R \subset \boldsymbol{\Theta}$ of given probability under $p(\boldsymbol{\theta} \,|\, D)$. For example, the identification of regions containing 50%, 90%, 95%, or 99% of the probability under the posterior may be sufficient to convey the general quantitative messages implicit in $p(\boldsymbol{\theta} \,|\, D)$; indeed, this is the intuitive basis of graphical representations of univariate distributions like those provided by boxplots. Any region $R \subset \boldsymbol{\Theta}$ such that $\int_R p(\boldsymbol{\theta} \,|\, D) d\boldsymbol{\theta} = q$ (so that, given data D, the true value of $\boldsymbol{\theta}$ belongs to R with probability q), is said to be a posterior *q-credible region* of $\boldsymbol{\theta}$. Notice that this provides immediately a direct intuitive statement about the unknown quantity of interest $\boldsymbol{\theta}$ in probability terms, in marked contrast to the circumlocutory statements provided by frequentist confidence intervals. Clearly, for any given q there are generally infinitely many credible regions. A credible region is invariant under reparametrization; thus, for any q-credible region R of $\boldsymbol{\theta}$, $\boldsymbol{\phi}(R)$ is a q-credible region of $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$. Sometimes, credible regions are selected to have minimum size (length, area, volume), resulting in *highest probability density* (HPD) regions, where all points in the region have larger probability density than all points outside. However, HPD regions are *not* invariant under reparametrization: the image $\boldsymbol{\phi}(R)$ of an HPD region R will be a credible region for $\boldsymbol{\phi}$, but will not generally be HPD; indeed, there is no compelling reason to restrict attention to HPD credible regions. Posterior quantiles are often used to derive credible regions. Thus, if $\boldsymbol{\theta}_q = \boldsymbol{\theta}_q(D)$ is the $100q\%$ posterior quantile of $\boldsymbol{\theta}$, then $R = \{\boldsymbol{\theta}; \boldsymbol{\theta} \leq \boldsymbol{\theta}_q\}$ is a one-sided, typically unique

q-credible region, and it is invariant under reparametrization. Indeed, *probability centered* q-credible regions of the form $R = \{\boldsymbol{\theta}; \boldsymbol{\theta}_{(1-q)/2} \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_{(1+q)/2}\}$ are easier to compute, and are often quoted in preference to HPD regions.

Example 3 (Inference on normal parameters, continued). In the numerical example about the value of the gravitational field described in the top panel of Fig. 3, the interval $[9.788, 9.829]$ in the unrestricted posterior density of g is a HPD, 95%-credible region for g. Similarly, the interval $[9.7803, 9.8322]$ in the bottom panel of Fig. 3 is also a 95%-credible region for g, but it is not HPD.

The concept of a credible region for a function $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$ of the parameter vector is trivially extended to prediction problems. Thus, a posterior q-credible region for $\boldsymbol{x} \in \boldsymbol{\chi}$ is a subset R of the outcome space $\mathcal{X}$ with posterior predictive probability q, so that $\int_R p(\boldsymbol{x} \mid D) d\boldsymbol{x} = q$.

For a description of the choice of credible regions using the intrinsic loss function, see Bernardo (2005b).

Hypothesis Testing

The posterior distribution $p(\boldsymbol{\theta} \mid D)$ of the quantity of interest $\boldsymbol{\theta}$ conveys immediate intuitive information on those values of $\boldsymbol{\theta}$ which, given the assumed model, may be taken to be *compatible* with the observed data D, namely, those with a relatively high probability density. Sometimes, a *restriction* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ of the possible values of the quantity of interest (where $\boldsymbol{\Theta}_0$ may possibly consists of a single value $\boldsymbol{\theta}_0$) is suggested in the course of the investigation as deserving special consideration, either because restricting $\boldsymbol{\theta}$ to $\boldsymbol{\Theta}_0$ would greatly simplify the model, or because there are additional, context specific arguments suggesting that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Intuitively, the *hypothesis* $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ should be judged to be *compatible* with the observed data D if there are elements in $\boldsymbol{\Theta}_0$ with a relatively high posterior density. However, a more precise conclusion is often required and, once again, this is made possible by adopting a decision-oriented approach. Formally, testing the hypothesis $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$ is a *decision problem* where the action space has only two elements, namely to accept (a_0) or to reject (a_1) the proposed restriction. To solve this decision problem, it is necessary to specify an appropriate loss function, $\ell(a_i, \boldsymbol{\theta})$, measuring the consequences of accepting or rejecting H_0 as a function of the actual value $\boldsymbol{\theta}$ of the vector of interest. Notice that this requires the statement of an *alternative* a_1 to accepting H_0; this is only to be expected, for an action is taken not because it is good, but because it is better than anything else that has been imagined.

Given data D, the optimal action will be to reject H_0 if (and only if) the expected posterior loss of accepting, $\int_\Theta \ell(a_0, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$, is larger than the expected posterior loss of rejecting, $\int_\Theta \ell(a_1, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$, that is, if (and only if)

$$\int_\Theta [\ell(a_0, \boldsymbol{\theta}) - \ell(a_1, \boldsymbol{\theta})]\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} = \int_\Theta \Delta\ell(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta} > 0.$$

Therefore, only the loss difference $\Delta\ell(\boldsymbol{\theta}) = \ell(a_0, \boldsymbol{\theta}) - \ell(a_1, \boldsymbol{\theta})$, which measures the *advantage* of rejecting H_0 as a function of $\boldsymbol{\theta}$, has to be specified. Thus, as common sense dictates, the hypothesis H_0 should be rejected whenever the expected advantage of rejecting H_0 is positive.

A crucial element in the specification of the loss function is a description of what is actually meant by rejecting H_0. By assumption a_0 means to act *as if* H_0 were true, i.e.,as if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, but there are at least two obvious options for the alternative action a_1. This may either mean (1) the *negation* of H_0, that is to act as if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$ or, alternatively, it may rather mean (2) to reject the simplification implied by H_0 and to keep the unrestricted model, $\boldsymbol{\theta} \in \Theta$, which is true by assumption. Both options have been analyzed in the literature, although it may be argued that the problems of scientific data analysis where hypothesis testing procedures are typically used are better described by the second alternative. Indeed, an established model, identified by $H_0 \equiv \{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0\}$, is often embedded into a more general model, $\{\boldsymbol{\theta} \in \Theta, \boldsymbol{\Theta}_0 \subset \Theta\}$, constructed to include possibly promising departures from H_0, and it is required to verify whether presently available data D are still compatible with $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, or whether the extension to $\boldsymbol{\theta} \in \Theta$ is really required.

Example 8 (Conventional hypothesis testing). Let $p(\boldsymbol{\theta} \mid D)$, $\boldsymbol{\theta} \in \Theta$, be the posterior distribution of the quantity of interest, let a_0 be the decision to work under the restriction $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ and let a_1 be the decision to work under the complementary restriction $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$. Suppose, moreover, that the loss structure has the simple, zero-one form given by $\{\ell(a_0, \boldsymbol{\theta}) = 0, \ell(a_1, \boldsymbol{\theta}) = 1\}$ if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ and, similarly, $\{\ell(a_0, \boldsymbol{\theta}) = 1, \ell(a_1, \boldsymbol{\theta}) = 0\}$ if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$, so that the *advantage* $\Delta\ell(\boldsymbol{\theta})$ of rejecting H_0 is 1 if $\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0$ and it is -1 otherwise. With this loss function it is immediately found that the optimal action is to reject H_0 if (and only if) $\Pr(\boldsymbol{\theta} \notin \boldsymbol{\Theta}_0 \mid D)$ is larger than $\Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \mid D)$. Notice that this formulation requires that $\Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta}_0) > 0$, that is, that the hypothesis H_0 has a strictly positive prior probability. If $\boldsymbol{\theta}$ is a continuous parameter and $\boldsymbol{\Theta}_0$ has zero measure (for instance if H_0 consists of a single point $\boldsymbol{\theta}_0$), this requires the use of a non-regular "sharp" prior concentrating a positive probability mass on $\boldsymbol{\Theta}_0$.

Example 9 (Intrinsic hypothesis testing). Again, let $p(\boldsymbol{\theta}\,|\,D)$, $\boldsymbol{\theta} \in \Theta$, be the posterior distribution of the quantity of interest, and let a_0 be the decision to work under the restriction $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, but let a_1 now be the decision to keep the general, unrestricted model $\boldsymbol{\theta} \in \Theta$. In this case, the advantage $\Delta\ell(\boldsymbol{\theta})$ of rejecting H_0 as a function of $\boldsymbol{\theta}$ may safely be assumed to have the form $\Delta\ell(\boldsymbol{\theta}) = d(\boldsymbol{\Theta}_0, \boldsymbol{\theta}) - d^*$, for some $d^* > 0$, where (1) $d(\boldsymbol{\Theta}_0, \boldsymbol{\theta})$ is some measure of the discrepancy between the assumed model $p(D\,|\,\boldsymbol{\theta})$ and its closest approximation within the class $\{p(D\,|\,\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0\}$, such that $d(\boldsymbol{\Theta}_0, \boldsymbol{\theta}) = 0$ whenever $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and (2) d^* is a context dependent *utility constant* which measures the (necessarily positive) advantage of being able to work with the simpler model when it is true. Choices of both $d(\boldsymbol{\Theta}_0, \boldsymbol{\theta})$ and d^* which may be appropriate for general use will now be described.

For reasons similar to those supporting its use in point estimation, an attractive choice for the function $d(\boldsymbol{\Theta}_0, \boldsymbol{\theta})$ is an appropriate extension of the intrinsic discrepancy; when there are no nuisance parameters, this is given by

$$d(\boldsymbol{\Theta}_0, \boldsymbol{\theta}) = \inf_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0} \min\{\delta(\boldsymbol{\theta}_0\,|\,\boldsymbol{\theta}),\, \delta(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}_0)\}$$

where $\delta(\boldsymbol{\theta}_0\,|\,\boldsymbol{\theta}) = \int_T p(\boldsymbol{t}\,|\,\boldsymbol{\theta}) \log\{p(\boldsymbol{t}\,|\,\boldsymbol{\theta})/p(\boldsymbol{t}\,|\,\boldsymbol{\theta}_0)\}\, d\boldsymbol{t}$, and $\boldsymbol{t} = \boldsymbol{t}(D) \in T$ is *any* sufficient statistic, which may well be the whole dataset D. As before, if the data $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ consist of a random sample from $p(\boldsymbol{x}\,|\,\boldsymbol{\theta})$, then

$$\delta(\boldsymbol{\theta}_0\,|\,\boldsymbol{\theta}) = n \int_{\mathcal{X}} p(\boldsymbol{x}\,|\,\boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}\,|\,\boldsymbol{\theta})}{p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_0)}\, d\boldsymbol{x}.$$

Naturally, the loss function $d(\boldsymbol{\Theta}_0, \boldsymbol{\theta})$ reduces to the intrinsic discrepancy $d(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ of Example 6 when $\boldsymbol{\Theta}_0$ contains a single element $\boldsymbol{\theta}_0$. Besides, as in the case of estimation, the definition is easily extended to problems with nuisance parameters, with

$$d(\boldsymbol{\Theta}_0, \boldsymbol{\theta}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0, \boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}} d(\boldsymbol{\theta}_0, \boldsymbol{\lambda}_o, \boldsymbol{\theta}, \boldsymbol{\lambda}).$$

The hypothesis H_0 should be rejected if the posterior expected advantage of rejecting is

$$\overline{d}(\boldsymbol{\Theta}_0\,|\,D) = \int_{\boldsymbol{\Lambda}} \int_{\Theta} d(\boldsymbol{\Theta}_0, \boldsymbol{\theta}, \boldsymbol{\lambda})\, p(\boldsymbol{\theta}, \boldsymbol{\lambda}\,|\,D)\, d\boldsymbol{\theta} d\boldsymbol{\lambda} > d^*,$$

for some $d^* > 0$. It is easily verified that the function $\overline{d}(\boldsymbol{\Theta}_0, D)$ is nonnegative. Morovever, if $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ is a one-to-one transformation of $\boldsymbol{\theta}$, then $\overline{d}(\boldsymbol{\phi}(\boldsymbol{\Theta}_0), D) = \overline{d}(\boldsymbol{\Theta}_0, D)$, so that the expected intrinsic loss of rejecting H_0 is invariant under reparametrization.

It may be shown that, as the sample size increases, the expected value of $\overline{d}(\boldsymbol{\Theta}_0, D)$ under sampling tends to one when H_0 is true, and tends to infinity otherwise; thus $\overline{d}(\boldsymbol{\Theta}_0, D)$ may be regarded as a continuous, positive measure of how inappropriate (in loss of information units) it would be to simplify the model by accepting H_0. In traditional language, $\overline{d}(\boldsymbol{\Theta}_0, D)$ is a *test statistic* for H_0 and the hypothesis should be rejected if the value of $\overline{d}(\boldsymbol{\Theta}_0, D)$ exceeds some *critical value* d^*. In sharp contrast to conventional hypothesis testing, this critical value d^* is found to be a context specific, positive utility constant d^*, which may precisely be described as the number of *information units* which the decision maker is prepared to lose in order to be able to work with the simpler model H_0, and does not depend on the sampling properties of the probability model. The procedure may be used with standard, continuous regular priors even in *sharp* hypothesis testing, when $\boldsymbol{\Theta}_0$ is a zero-measure set (as would be the case if $\boldsymbol{\theta}$ is continuous and $\boldsymbol{\Theta}_0$ contains a single point $\boldsymbol{\theta}_0$). Naturally, to implement the test, the utility constant d^* which defines the rejection region must be chosen.

All measurements are based on a comparison with a standard; comparison with the "canonical" problem of testing a value $\mu = \mu_0$ for the mean of a normal distribution with known variance (see below) makes it possible to *calibrate* this *information scale*. Values of $\overline{d}(\boldsymbol{\Theta}_0, D)$ of about one should be regarded as an indication of no evidence against H_0, since the expected value of $\overline{d}(\boldsymbol{\Theta}_0, D)$ under H_0 is exactly equal to one. Values of $\overline{d}(\boldsymbol{\Theta}_0, D)$ of about 2.5, and 5 should be respectively regarded as an indication of mild evidence against H_0, and significant evidence against H_0 since, in the canonical normal problem, these values correspond to the observed sample mean $\overline{x}$ respectively lying two or three posterior standard deviations from the null value μ_0. Notice that, in sharp contrast to frequentist hypothesis testing, where it is hazily recommended to adjust the significance level for dimensionality and sample size, this provides an absolute scale (in information units) which remains valid for any sample size and any dimensionality.

Example 10 (Testing the value of a normal mean). Let the data $D = \{x_1, \ldots, x_n\}$ be a random sample from a normal distribution $\mathrm{N}(x\,|\,\mu, \sigma)$, where σ is assumed to be known, and consider the "canonical" problem of testing whether these data are or are not compatible with some specific sharp hypothesis $H_0 \equiv \{\mu = \mu_0\}$ on the value of the mean.

The conventional approach to this problem requires a non-regular prior which places a probability mass, say p_0, on the value μ_0 to be tested, with the remaining $1 - p_0$ probability continuously distributed over $\Re$. If this prior is chosen to be $p(\mu\,|\,\mu \neq \mu_0) = N(\mu\,|\,\mu_0, \sigma_0)$, Bayes theorem

may be used to obtain the corresponding posterior probability,

$$\Pr[\mu_0 \mid D, \lambda] = \frac{B_{01}(D,\lambda)\, p_0}{(1-p_0) + p_0\, B_{01}(D,\lambda)},$$

$$B_{01}(D,\lambda) = \left(1+\frac{n}{\lambda}\right)^{1/2} \exp\left[-\frac{1}{2}\frac{n}{n+\lambda}\, z^2\right],$$

where $z = (\bar{x}-\mu_0)/\left(\sigma/\sqrt{n}\right)$ measures, in standard deviations, the distance between $\bar{x}$ and μ_0 and $\lambda = \sigma^2/\sigma_0^2$ is the ratio of model to prior variance. The function $B_{01}(D,\lambda)$, a ratio of (integrated) likelihood functions, is called the *Bayes factor* in favor of H_0. With a conventional zero-one loss function, H_0 should be rejected if $\Pr[\mu_0 \mid D, \lambda] < 1/2$. The choices $p_0 = 1/2$ and $\lambda = 1$ or $\lambda = 1/2$, describing particular forms of *sharp* prior knowledge, have been suggested in the literature for routine use. The conventional approach to sharp hypothesis testing deals with situations of *concentrated* prior probability; it *assumes* important prior knowledge about the value of μ and, hence, should *not* be used unless this is an appropriate assumption. Moreover, as pointed out by Bartlett (1957), the resulting posterior probability is extremely sensitive to the specific prior specification. In most applications, H_0 is really a hazily defined small region rather than a point. For moderate sample sizes, the posterior probability $\Pr[\mu_0 \mid D, \lambda]$ is an *approximation* to the posterior probability $\Pr[\mu_0 - \epsilon < \mu < \mu_0 - \epsilon \mid D, \lambda]$ for some small interval around μ_0 which would have been obtained from a regular, continuous prior heavily concentrated around μ_0; however, this approximation *always* breaks down for sufficiently large sample sizes. One consequence (which is immediately apparent from the last two equations) is that for any *fixed* value of the pertinent statistic z, the posterior probability of the null, $\Pr[\mu_0 \mid D, \lambda]$, tends to one as $n \to \infty$. Far from being specific to this example, this unappealing behavior of posterior probabilities based on sharp, non-regular priors (discovered by Lindley 1957, and generally known as *Lindley's paradox*) is *always* present in the conventional Bayesian approach to *sharp* hypothesis testing.

The intrinsic approach may be used without assuming any sharp prior knowledge. The intrinsic discrepancy is $d(\mu_0,\mu) = n(\mu-\mu_0)^2/(2\sigma^2)$, a simple transformation of the standardized distance between μ and μ_0. As later explained (section "▶Reference Analysis"), absence of initial information about the value of μ may formally be described in this problem by the (improper) uniform prior function $p(\mu) = 1$; Bayes' theorem may then be used to obtain the corresponding (proper) posterior distribution, $p(\mu \mid D) = N\left(\mu \mid \bar{x}, \sigma/\sqrt{n}\right)$. The expected value of $d(\mu_0,\mu)$ with respect to this posterior is $\overline{d}(\mu_0, D) = (1+z^2)/2$, where $z = (\bar{x}-\mu_0)/\left(\sigma/\sqrt{n}\right)$ is the standardized distance between $\bar{x}$ and μ_0. As foretold by the general theory, the expected value of $\overline{d}(\mu_0, D)$ under repeated sampling is one if $\mu = \mu_0$, and increases linearly with n if $\mu = \mu_0$. Moreover, in this canonical example, to reject H_0 whenever $|z| > 2$ or $|z| > 3$, that is whenever μ_0 is two or three posterior standard deviations away from $\bar{x}$, respectively corresponds to rejecting H_0 whenever $\overline{d}(\mu_0, D)$ is larger than 2.5, or larger than 5. But the information scale is independent of the problem, so that rejecting the null whenever its expected discrepancy from the true model is larger than $d^* = 5$ units of information is a *general* rule (and one which corresponds to the conventional "3σ" rule in the canonical normal case).

If σ is unknown, the intrinsic discrepancy becomes

$$d(\mu_0, \mu, \sigma) = \frac{n}{2} \log\left[1 + \left(\frac{\mu-\mu_0}{\sigma}\right)^2\right].$$

Moreover, as mentioned before, absence of initial information about both μ and σ may be described by the (improper) prior function $p(\mu,\sigma) = \sigma^{-1}$. The intrinsic test statistic $\overline{d}(\mu_0, D)$ is found as the expected value of $d(\mu_0, \mu, \sigma)$ under the corresponding joint posterior distribution; this may be exactly expressed in terms of hypergeometric functions, and is approximated by

$$\overline{d}(\mu_0, D) \approx \frac{1}{2} + \frac{n}{2} \log\left(1 + \frac{t^2}{n}\right),$$

where t is the traditional statistic $t = \sqrt{n-1}(\bar{x}-\mu_0)/s$, $ns^2 = \sum_j (x_j - \bar{x})^2$. For instance, for samples sizes 5, 30 and 1,000, and using the utility constant $d^* = 5$, the hypothesis H_0 would be rejected whenever $|t|$ is respectively larger than 5.025, 3.240, and 3.007.

Reference Analysis

Under the Bayesian paradigm, the outcome of any inference problem (the posterior distribution of the quantity of interest) combines the information provided by the data with relevant available prior information. In many situations, however, either the available prior information on the quantity of interest is too vague to warrant the effort required to have it formalized in the form of a probability distribution, or it is too subjective to be useful in scientific communication or public decision making. It is therefore important to be able to identify the mathematical form of a "noninformative" prior, a prior that would have a minimal effect, relative to the data, on the posterior inference. More formally, suppose that the probability mechanism which has generated the available data D is assumed to be $p(D \mid \boldsymbol{\omega})$, for some $\boldsymbol{\omega} \in \Omega$,

and that the quantity of interest is some real-valued function $\theta = \theta(\boldsymbol{\omega})$ of the model parameter $\boldsymbol{\omega}$. Without loss of generality, it may be assumed that the probability model is of the form $p(D \mid \theta, \boldsymbol{\lambda})$, $\theta \in \Theta$, $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, where $\boldsymbol{\lambda}$ is some appropriately chosen nuisance parameter vector. As described in section "▶The Bayesian Paradigm", to obtain the required posterior distribution of the quantity of interest $p(\theta \mid D)$ it is necessary to specify a *joint* prior $p(\theta, \boldsymbol{\lambda})$. It is now required to identify the form of that joint prior $\pi_\theta(\theta, \boldsymbol{\lambda})$, the θ-*reference prior*, which would have a *minimal effect* on the corresponding posterior distribution of θ,

$$\pi(\theta \mid D) \propto \int_{\boldsymbol{\Lambda}} p(D \mid \theta, \boldsymbol{\lambda})\, \pi_\theta(\theta, \boldsymbol{\lambda})\, d\boldsymbol{\lambda},$$

a prior which, to use a conventional expression, "would let the data speak for themselves" about the likely value of θ. Properly defined, reference *posterior* distributions have an important role to play in scientific communication, for they provide the answer to a central question in the sciences: conditional on the assumed model $p(D \mid \theta, \boldsymbol{\lambda})$, and on any further assumptions of the value of θ on which there might be universal agreement, the reference posterior $\pi(\theta \mid D)$ should specify what *could* be said about θ *if* the only available information about θ were some well-documented data D.

Much work has been done to formulate "reference" priors which would make the idea described above mathematically precise. This section concentrates on an approach that is based on information theory to derive reference distributions which may be argued to provide the most advanced general procedure available. In the formulation described below, far from ignoring prior knowledge, the reference posterior exploits certain well-defined features of a *possible* prior, namely those describing a situation where relevant knowledge about the quantity of interest (beyond that universally accepted) may be held to be negligible compared to the information about that quantity which repeated experimentation (from a particular data generating mechanism) might possibly provide. Reference analysis is appropriate in contexts where the set of inferences which could be drawn in this *possible* situation is considered to be pertinent.

Any statistical analysis contains a fair number of subjective elements; these include (among others) the data selected, the model assumptions, and the choice of the quantities of interest. Reference analysis may be argued to provide an "objective" Bayesian solution to statistical inference problems in just the same sense that conventional statistical methods claim to be "objective": in that the solutions only depend on model assumptions and observed data. The whole topic of objective Bayesian methods is, however, subject to polemic; interested readers will find in Bernardo (2005a) and references therein some pointers to the relevant literature.

Reference Distributions

One parameter. Consider the experiment which consists of the observation of data D, generated by a random mechanism $p(D \mid \theta)$ which only depends on a real-valued parameter $\theta \in \Theta$, and let $\boldsymbol{t} = \boldsymbol{t}(D) \in \mathcal{T}$ be *any* sufficient statistic (which may well be the complete data set D). In Shannon's general information theory, the *amount of information* $I^\theta\{\mathcal{T}, p(\theta)\}$ which may be expected to be provided by D, or (equivalently) by $\boldsymbol{t}(D)$, about the value of θ is defined by

$$I^\theta\{\mathcal{T}, p(\theta)\} = \int_{\mathcal{T}} \int_{\Theta} p(\boldsymbol{t}, \theta) \log \frac{p(\boldsymbol{t}, \theta)}{p(\boldsymbol{t})p(\theta)}\, d\theta d\boldsymbol{t}$$

$$= \mathrm{E}_{\boldsymbol{t}}\left[\int_{\Theta} p(\theta \mid \boldsymbol{t}) \log \frac{p(\theta \mid \boldsymbol{t})}{p(\theta)}\, d\theta \right]$$

the expected logarithmic divergence of the prior from the posterior. This is naturally a *functional* of the prior $p(\theta)$: the larger the prior information, the smaller the information which the data may be expected to provide. The functional $I^\theta\{\mathcal{T}, p(\theta)\}$ is concave, non-negative, and invariant under one-to-one transformations of θ. Consider now the amount of information $I^\theta\{\mathcal{T}^k, p(\theta)\}$ about θ which may be expected from the experiment which consists of k conditionally independent replications $\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_k\}$ of the original experiment. As $k \to \infty$, such an experiment would provide any *missing information* about θ which could possibly be obtained within this framework; thus, as $k \to \infty$, the functional $I^\theta\{\mathcal{T}^k, p(\theta)\}$ will approach the missing information about θ associated with the prior $p(\theta)$. Intuitively, a θ-"noninformative" prior is one which *maximizes the missing information* about θ. Formally, if $\pi_k(\theta)$ denotes the prior density which maximizes $I^\theta\{\mathcal{T}^k, p(\theta)\}$ in the class $\mathcal{P}$ of strictly positive prior distributions which are compatible with accepted assumptions on the value of θ (which may well be the class of *all* strictly positive proper priors) then the θ-reference prior $\pi(\theta)$ is the limit as $k \to \infty$ (in a sense to be made precise) of the sequence of priors $\{\pi_k(\theta), k = 1, 2, \ldots\}$.

Notice that this limiting procedure is *not* some kind of asymptotic approximation, but an essential element of the *definition* of a reference prior. In particular, this definition implies that reference distributions only depend on the *asymptotic* behavior of the assumed probability model, a feature which greatly simplifies their actual derivation.

Example 11 (Maximum entropy). If θ may only take a *finite* number of values, so that the parameter space is $\Theta = \{\theta_1, \ldots, \theta_m\}$ and $p(\theta) = \{p_1, \ldots, p_m\}$, with $p_i = \Pr(\theta = \theta_i)$, then the missing information associated to $\{p_1, \ldots, p_m\}$ may be shown to be

$$\lim_{k \to \infty} I^{\theta}\{T^k, p(\theta)\} = H(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log(p_i),$$

that is, the *entropy* of the prior distribution $\{p_1, \ldots, p_m\}$. Thus, in the finite case, if there is no further structure in the problem (which should then be taken into account), the reference prior is that with *maximum entropy* in the class $\mathcal{P}$ of priors compatible with accepted assumptions. Consequently, the reference prior algorithm contains "maximum entropy" priors as the particular case which obtains when the parameter space is *finite*, the *only* case where the original concept of ►entropy (in statistical mechanics, as a measure of uncertainty) is unambiguous and well-behaved. If, in particular, $\mathcal{P}$ contains *all* priors over $\{\theta_1, \ldots, \theta_m\}$, then the reference prior is the uniform prior, $\pi(\theta) = \{1/m, \ldots, 1/m\}$.

Formally, the *reference prior function* $\pi(\theta)$ of a univariate parameter θ is defined to be the limit of the sequence of the proper priors $\pi_k(\theta)$ which maximize $I^{\theta}\{T^k, p(\theta)\}$ in the precise sense that, for any value of the sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(D)$, the *reference posterior*, the pointwise limit $\pi(\theta \mid \boldsymbol{t})$ of the corresponding sequence of posteriors $\{\pi_k(\theta \mid \boldsymbol{t})\}$, may be obtained from $\pi(\theta)$ by formal use of Bayes theorem, so that $\pi(\theta \mid \boldsymbol{t}) \propto p(\boldsymbol{t} \mid \theta)\, \pi(\theta)$.

Reference prior *functions* are often simply called reference priors, even though they are usually *not* probability distributions. They should *not* be considered as expressions of belief, but technical devices to obtain (proper) posterior distributions which are a limiting form of the posteriors which could have been obtained from possible prior beliefs which were relatively uninformative with respect to the quantity of interest when compared with the information which data could provide.

If (1) the sufficient statistic $\boldsymbol{t} = \boldsymbol{t}(D)$ is a consistent estimator $\tilde{\theta}$ of a continuous parameter θ, and (2) the class $\mathcal{P}$ contains *all* strictly positive priors, then the reference prior may be shown to have a simple form in terms of any *asymptotic* approximation to the posterior distribution of θ. Notice that, by construction, an *asymptotic* approximation to the posterior does *not* depend on the prior. Specifically, if the posterior density $p(\theta \mid D)$ has an asymptotic approximation of the form $p(\theta \mid \tilde{\theta}, n)$, the reference prior is simply

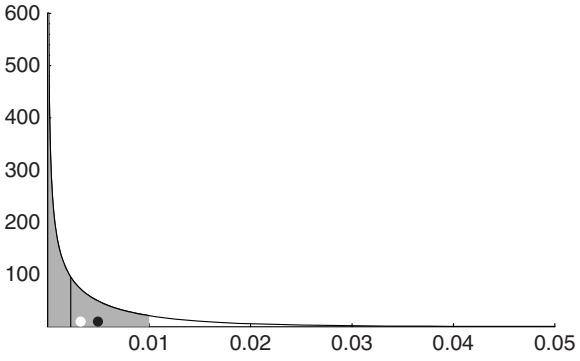
$$\pi(\theta) \propto p(\theta \mid \tilde{\theta}, n)\Big|_{\tilde{\theta}=\theta}$$

One-parameter reference priors are shown to be *invariant* under reparametrization; thus, if $\psi = \psi(\theta)$ is a piecewise one-to-one function of θ, then the ψ-reference prior is simply the appropriate probability transformation of the θ-reference prior.

Example 12 (Jeffreys' prior). If θ is univariate and continuous, and the posterior distribution of θ given $\{x_1 \ldots, x_n\}$ is asymptotically normal with standard deviation $s(\tilde{\theta})/\sqrt{n}$, then, using the last displayed equation, the reference prior function is $\pi(\theta) \propto s(\theta)^{-1}$. Under regularity conditions (often satisfied in practice, see section "►Asymptopic Behavior"), the posterior distribution of θ is asymptotically normal with variance $n^{-1} F^{-1}(\hat{\theta})$, where $F(\theta)$ is Fisher's information function and $\hat{\theta}$ is the MLE of θ. Hence, the reference prior function in these conditions is $\pi(\theta) \propto F(\theta)^{1/2}$, which is known as Jeffreys' prior. It follows that the reference prior algorithm contains Jeffreys' priors as the particular case which obtains when the probability model only depends on a single continuous univariate parameter, there are regularity conditions to guarantee ►asymptotic normality, and there is no additional information, so that the class of possible priors $\mathcal{P}$ contains all strictly positive priors over Θ. These are precisely the conditions under which there is general agreement on the use of Jeffreys' prior as a "noninformative" prior.

Example 2 (Inference on a binomial parameter, continued). Let data consist of a sequence $D = \{x_1, \ldots, x_n\}$ of n conditionally independent Bernoulli trials, so that $p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$, $x \in \{0,1\}$; this is a regular, one-parameter continuous model, whose Fisher's information function is $F(\theta) = \theta^{-1}(1-\theta)^{-1}$. Thus, the reference prior $\pi(\theta)$ is proportional to $\theta^{-1/2}(1-\theta)^{-1/2}$, so that the reference prior is the (proper) Beta distribution $\mathrm{Be}(\theta \mid 1/2, 1/2)$. Since the reference algorithm is invariant under reparametrization, the reference prior of $\phi(\theta) = 2\,arc\sin\sqrt{\theta}$ is $\pi(\phi) = \pi(\theta)/|\partial\phi/\partial\theta| = 1$; thus, the reference prior is *uniform on the variance-stabilizing transformation* $\phi(\theta) = 2arc\sin\sqrt{\theta}$, a feature generally true under regularity conditions. In terms of the original parameter θ, the corresponding reference posterior is $\mathrm{Be}(\theta \mid r + 1/2, n - r + 1/2)$, where $r = \sum x_j$ is the number of positive trials.

Suppose, for example, that $n = 100$ randomly selected people have been tested for an infection and that all tested negative, so that $r = 0$. The reference posterior distribution of the proportion θ of people infected is then the Beta distribution $\mathrm{Be}(\theta \mid 0.5, 100.5)$, represented in Fig. 4. It may well be known that the infection was rare, leading to the assumption that $\theta < \theta_0$, for some upper bound



Bayesian Statistics. Fig. 4 Posterior distribution of the proportion of infected people in the population, given the results of $n = 100$ tests, none of which were positive

θ_0; the (restricted) reference prior would then be of the form $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ if $\theta < \theta_0$, and zero otherwise. However, provided the likelihood is concentrated in the region $\theta < \theta_0$, the corresponding posterior would virtually be identical to $\text{Be}(\theta \,|\, 0.5, 100.5)$. Thus, just on the basis of the observed experimental results, one may claim that the proportion of infected people is surely smaller than 5% (for the reference posterior probability of the event $\theta > 0.05$ is 0.001), that θ is smaller than 0.01 with probability 0.844 (area of the shaded region in Fig. 4), that it is equally likely to be over or below 0.23% (for the median, represented by a vertical line, is 0.0023), and that the probability that a person randomly chosen from the population is infected is 0.005 (the posterior mean, represented in the figure by a black circle), since $\Pr(x = 1 \,|\, r, n) = \text{E}[\theta \,|\, r, n] = 0.005$. If a particular point estimate of θ is required (say a number to be quoted in the summary headline) the *intrinsic* estimator suggests itself; this is found to be $\theta^* = 0.0032$ (represented in the figure with a white circle). Notice that the traditional solution to this problem, based on the asymptotic behavior of the MLE, here $\hat{\theta} = r/n = 0$ for any n, makes absolutely no sense in this scenario.

One nuisance parameter. The extension of the reference prior algorithm to the case of two parameters follows the usual mathematical procedure of reducing the problem to a sequential application of the established procedure for the single parameter case. Thus, if the probability model is $p(\boldsymbol{t} \,|\, \theta, \lambda)$, $\theta \in \Theta$, $\lambda \in \Lambda$ and a θ-reference prior $\pi_\theta(\theta, \lambda)$ is required, the reference algorithm proceeds in two steps:

1. Conditional on θ, $p(\boldsymbol{t} \,|\, \theta, \lambda)$ only depends on the nuisance parameter λ and, hence, the one-parameter algorithm may be used to obtain the *conditional* reference prior $\pi(\lambda \,|\, \theta)$.
2. If $\pi(\lambda \,|\, \theta)$ is proper, this may be used to integrate out the nuisance parameter thus obtaining the one-parameter integrated model $p(\boldsymbol{t} \,|\, \theta) = \int_\Lambda p(\boldsymbol{t} \,|\, \theta, \lambda)\, \pi(\lambda \,|\, \theta)\, d\lambda$, to which the one-parameter algorithm may be applied again to obtain $\pi(\theta)$. The θ-reference prior is then $\pi_\theta(\theta, \lambda) = \pi(\lambda \,|\, \theta)\, \pi(\theta)$, and the required reference posterior is $\pi(\theta \,|\, \boldsymbol{t}) \propto p(\boldsymbol{t} \,|\, \theta) \pi(\theta)$.

If the conditional reference prior is *not* proper, then the procedure is performed within an increasing sequence $\{\Lambda_i\}$ of subsets converging to Λ over which $\pi(\lambda \,|\, \theta)$ is integrable. This makes it possible to obtain a corresponding sequence of θ-reference posteriors $\{\pi_i(\theta \,|\, \boldsymbol{t}\}$ for the quantity of interest θ, and the required reference posterior is the corresponding pointwise limit $\pi(\theta \,|\, \boldsymbol{t}) = \lim_i \pi_i(\theta \,|\, \boldsymbol{t})$. A θ-reference prior is then defined as a positive function $\pi_\theta(\theta, \lambda)$ which may be formally used in Bayes' theorem as a prior to obtain the reference posterior, i.e., such that, for any $\boldsymbol{t} \in T$, $\pi(\theta \,|\, \boldsymbol{t}) \propto \int_\Lambda p(\boldsymbol{t} \,|\, \theta, \lambda)\, \pi_\theta(\theta, \lambda)\, d\lambda$. The approximating sequences should be *consistently* chosen within a given model. Thus, given a probability model $\{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ an appropriate approximating sequence $\{\boldsymbol{\Omega}_i\}$ should be chosen for the whole parameter space $\boldsymbol{\Omega}$. Thus, if the analysis is done in terms of, say, $\boldsymbol{\psi} = \{\psi_1, \psi_2\} \in \Psi(\boldsymbol{\Omega})$, the approximating sequence should be chosen such that $\Psi_i = \psi(\boldsymbol{\Omega}_i)$. A natural approximating sequence in location-scale problems is $\{\mu, \log \sigma\} \in [-i, i]^2$.

The θ-reference prior does *not* depend on the choice of the nuisance parameter λ; thus, for any $\psi = \psi(\theta, \lambda)$ such that (θ, ψ) is a one-to-one function of (θ, λ), the θ-reference prior in terms of (θ, ψ) is simply $\pi_\theta(\theta, \psi) = \pi_\theta(\theta, \lambda)/|\partial(\theta, \psi)/\partial(\theta, \lambda)|$, the appropriate probability transformation of the θ-reference prior in terms of (θ, λ). Notice, however, that the reference prior *may* depend on the parameter of interest; thus, the θ-reference prior may differ from the ϕ-reference prior unless either ϕ is a piecewise one-to-one transformation of θ, or ϕ is asymptotically independent of θ. This is an expected consequence of the fact that the conditions under which the missing information about θ is maximized are not generally the same as the conditions which maximize the missing information about some function $\phi = \phi(\theta, \lambda)$.

The *non-existence* of a unique "noninformative prior" which would be appropriate for any inference problem within a given model was established by Dawid et al. (1973), when they showed that this is incompatible with *consistent marginalization*. Indeed, if given the model $p(D \,|\, \theta, \lambda)$, the reference posterior of the quantity of interest θ, $\pi(\theta \,|\, D) = \pi(\theta \,|\, \boldsymbol{t})$, only depends on the data through a statistic $\boldsymbol{t}$ whose sampling distribution, $p(\boldsymbol{t} \,|\, \theta, \lambda) = p(\boldsymbol{t} \,|\, \theta)$, only depends on θ, one would expect the reference posterior to be of the form $\pi(\theta \,|\, \boldsymbol{t}) \propto \pi(\theta)\, p(\boldsymbol{t} \,|\, \theta)$

for some prior $\pi(\theta)$. However, examples were found where this cannot be the case if a *unique* joint "noninformative" prior were to be used for all possible quantities of interest.

Example 13 (Regular two dimensional continuous reference prior functions). If the joint posterior distribution of (θ, λ) is asymptotically normal, then the θ-reference prior may be derived in terms of the corresponding Fisher's information matrix, $\boldsymbol{F}(\theta, \lambda)$. Indeed, if

$$\boldsymbol{F}(\theta,\lambda) = \begin{pmatrix} F_{\theta\theta}(\theta,\lambda) & F_{\theta\lambda}(\theta,\lambda) \\ F_{\theta\lambda}(\theta,\lambda) & F_{\lambda\lambda}(\theta,\lambda) \end{pmatrix}, \quad \text{and}$$
$$\boldsymbol{S}(\theta,\lambda) = \boldsymbol{F}^{-1}(\theta,\lambda),$$

then the θ-reference prior is $\pi_\theta(\theta,\lambda) = \pi(\lambda \,|\, \theta)\,\pi(\theta)$, where

$$\pi(\lambda \,|\, \theta) \propto F_{\lambda\lambda}^{1/2}(\theta,\lambda), \quad \lambda \in \Lambda.$$

If $\pi(\lambda \,|\, \theta)$ is proper,

$$\pi(\theta) \propto \exp\Big\{ \int_\Lambda \pi(\lambda \,|\, \theta)\, \log\Big[S_{\theta\theta}^{-1/2}(\theta,\lambda)\Big]\, d\lambda \Big\}, \quad \theta \in \Theta.$$

If $\pi(\lambda \,|\, \theta)$ is not proper, integrations are performed on an approximating sequence $\{\Lambda_i\}$ to obtain a sequence $\{\pi_i(\lambda \,|\, \theta)\, \pi_i(\theta)\}$, (where $\pi_i(\lambda \,|\, \theta)$ is the proper renormalization of $\pi(\lambda \,|\, \theta)$ to Λ_i) and the θ-reference prior $\pi_\theta(\theta,\lambda)$ is defined as its appropriate limit. Moreover, if (1) both $F_{\lambda\lambda}^{1/2}(\theta,\lambda)$ and $S_{\theta\theta}^{-1/2}(\theta,\lambda)$ *factorize*, so that

$$S_{\theta\theta}^{-1/2}(\theta,\lambda) \propto f_\theta(\theta)\, g_\theta(\lambda), \quad F_{\lambda\lambda}^{1/2}(\theta,\lambda) \propto f_\lambda(\theta)\, g_\lambda(\lambda),$$

and (2) the parameters θ and λ are *variation independent*, so that Λ does not depend on θ, then the θ-reference prior is simply $\pi_\theta(\theta,\lambda) = f_\theta(\theta)\, g_\lambda(\lambda)$, even if the conditional reference prior $\pi(\lambda \,|\, \theta) = \pi(\lambda) \propto g_\lambda(\lambda)$ (which will not depend on θ) is actually improper.

Example 3 (Inference on normal parameters, continued). The information matrix which corresponds to a normal model $\mathrm{N}(x \,|\, \mu, \sigma)$ is

$$\boldsymbol{F}(\mu,\sigma) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix},$$
$$\boldsymbol{S}(\mu,\sigma) = \boldsymbol{F}^{-1}(\mu,\sigma) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{pmatrix};$$

hence $F_{\sigma\sigma}^{1/2}(\mu,\sigma) = \sqrt{2}\,\sigma^{-1} = f_\sigma(\mu)\, g_\sigma(\sigma)$, with $g_\sigma(\sigma) = \sigma^{-1}$, and $\pi(\sigma \,|\, \mu) = \sigma^{-1}$. Similarly, $S_{\mu\mu}^{-1/2}(\mu,\sigma) = \sigma^{-1} = f_\mu(\mu)\, g_\mu(\sigma)$, with $f_\mu(\mu) = 1$, and $\pi(\mu) = 1$. Therefore, the μ-reference prior is $\pi_\mu(\mu,\sigma) = \pi(\sigma \,|\, \mu)\,\pi(\mu) = \sigma^{-1}$, as already anticipated. Moreover, as one would expect from the fact that $\boldsymbol{F}(\mu,\sigma)$ is diagonal and also anticipated, it is similarly found that the σ-reference prior is $\pi_\sigma(\mu,\sigma) = \sigma^{-1}$, the same as $\pi_\mu(\mu,\sigma)$.

Suppose, however, that the quantity of interest is *not* the mean μ or the standard deviation σ, but the *standardized* mean $\phi = \mu/\sigma$. Fisher's information matrix in terms of the parameters ϕ and σ is $\boldsymbol{F}(\phi,\sigma) = J^t\, \boldsymbol{F}(\mu,\sigma)\, J$, where $J = (\partial(\mu,\sigma)/\partial(\phi,\sigma))$ is the Jacobian of the inverse transformation; this yields

$$\boldsymbol{F}(\phi,\sigma) = \begin{pmatrix} 1 & \phi\sigma^{-1} \\ \phi\sigma^{-1} & \sigma^{-2}(2+\phi^2) \end{pmatrix},$$
$$\boldsymbol{S}(\phi,\sigma) = \begin{pmatrix} 1+\frac{1}{2}\phi^2 & -\frac{1}{2}\phi\sigma \\ -\frac{1}{2}\phi\sigma & \frac{1}{2}\sigma^2 \end{pmatrix}.$$

Thus, $S_{\phi\phi}^{-1/2}(\phi,\sigma) \propto \left(1+\frac{1}{2}\phi^2\right)^{-1/2}$ and $F_{\sigma\sigma}^{1/2}(\phi,\sigma) \propto \sigma^{-1}(2+\phi^2)^{1/2}$. Hence, using again the results in Example 13, $\pi_\phi(\phi,\sigma) = \left(1+\frac{1}{2}\phi^2\right)^{-1/2}\sigma^{-1}$. In the original parametrization, this is $\pi_\phi(\mu,\sigma) = \left(1+\frac{1}{2}(\mu/\sigma)^2\right)^{-1/2}\sigma^{-2}$, which is *very* different from $\pi_\mu(\mu,\sigma) = \pi_\sigma(\mu,\sigma) = \sigma^{-1}$. The corresponding reference posterior of ϕ is $\pi(\phi \,|\, x_1, \ldots, x_n) \propto \left(1+\frac{1}{2}\phi^2\right)^{-1/2} p(t \,|\, \phi)$ where $t = (\sum x_j)/\left(\sum x_j^2\right)^{1/2}$, a one-dimensional (marginally sufficient) statistic whose sampling distribution, $p(t \,|\, \mu,\sigma) = p(t \,|\, \phi)$, only depends on ϕ. Thus, the reference prior algorithm is seen to be consistent under marginalization.

Many parameters. The reference algorithm is easily generalized to an arbitrary number of parameters. If the model is $p(\boldsymbol{t} \,|\, \omega_1, \ldots, \omega_m)$, a joint reference prior

$$\pi(\theta_m \,|\, \theta_{m-1}, \ldots, \theta_1) \times \ldots \times \pi(\theta_2 \,|\, \theta_1) \times \pi(\theta_1)$$

may sequentially be obtained for each *ordered* parametrization $\{\theta_1(\boldsymbol{\omega}), \ldots, \theta_m(\boldsymbol{\omega})\}$ of interest, and these are invariant under reparametrization of any of the $\theta_i(\boldsymbol{\omega})$'s. The choice of the ordered parametrization $\{\theta_1, \ldots, \theta_m\}$ precisely describes the particular prior required, namely that which *sequentially* maximizes the missing information about each of the θ_i's, conditional on $\{\theta_1, \ldots, \theta_{i-1}\}$, for $i = m, m-1, \ldots, 1$.

Example 14 (Stein's paradox). Let D be a random sample from a m-variate normal distribution with mean $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_m\}$ and unitary variance matrix. The reference prior which corresponds to any permutation of the μ_i's is uniform, and this prior leads indeed to appropriate reference posterior distributions for any of the μ_i's, namely $\pi(\mu_i \,|\, D) = N\left(\mu_i \,|\, \bar{x}_i, 1/\sqrt{n}\right)$. Suppose, however, that the quantity of interest is $\theta = \sum_i \mu_i^2$, the distance of $\boldsymbol{\mu}$ to the origin. As showed by Stein in the 1950s, the

posterior distribution of θ based on that uniform prior (or in any "flat" *proper* approximation) has very undesirable properties; this is due to the fact that a uniform (or nearly uniform) prior, although "noninformative" with respect to each of the individual μ_i's, is actually highly informative on the sum of their squares, introducing a severe positive bias (Stein's paradox). However, the reference prior which corresponds to a parametrization of the form $\{\theta, \lambda_1, \ldots, \lambda_{m-1}\}$ produces, for any choice of the nuisance parameters $\lambda_i = \lambda_i(\boldsymbol{\mu})$, the reference posterior $\pi(\theta \,|\, D) = \pi(\theta \,|\, t) \propto \theta^{-1/2} \chi^2(nt \,|\, m, n\theta)$, where $t = \sum_i \bar{x}_i^2$, and this posterior is shown to have the appropriate consistency properties.

Far from being specific to Stein's example, the inappropriate behavior in problems with many parameters of specific marginal posterior distributions derived from multivariate "flat" priors (proper or improper) is indeed very frequent. Hence, sloppy, uncontrolled use of "flat" priors (rather than the relevant reference priors), is strongly discouraged.

Limited information. Although often used in contexts where no universally agreed prior knowledge about the quantity of interest is available, the reference algorithm may be used to specify a prior which incorporates any acceptable prior knowledge; it suffices to maximize the missing information within the class $\mathcal{P}$ of priors which is compatible with such accepted knowledge. Indeed, by progressive incorporation of further restrictions into $\mathcal{P}$, the reference prior algorithm becomes a method of (prior) *probability assessment*. As described below, the problem has a fairly simple analytical solution when those restrictions take the form of known expected values. The incorporation of other type of restrictions usually involves numerical computations.

Example 15 (Univariate restricted reference priors). If the probability mechanism which is assumed to have generated the available data only depends on a univarite continuous parameter $\theta \in \Theta \subset \Re$, and the class $\mathcal{P}$ of acceptable priors is a class of proper priors which satisfies some expected value restrictions, so that

$$\mathcal{P} = \left\{ p(\theta); \quad p(\theta) > 0, \int_\Theta p(\theta)\, d\theta = 1, \right.$$
$$\left. \int_\Theta g_i(\theta)\, p(\theta)\, d\theta = \beta_i,\ i = 1, \ldots, m \right\}$$

then the (restricted) reference prior is

$$\pi(\theta \,|\, \mathcal{P}) \propto \pi(\theta)\, \exp\left[\textstyle\sum_{j=1}^{m} \gamma_i\, g_i(\theta)\right]$$

where $\pi(\theta)$ is the unrestricted reference prior and the γ_i's are constants (the corresponding Lagrange multipliers), to be determined by the restrictions which define $\mathcal{P}$. Suppose, for instance, that data are considered to be a random sample from a location model centered at θ, and that it is further assumed that $\mathrm{E}[\theta] = \mu_0$ and that $\mathrm{Var}[\theta] = \sigma_0^2$. The unrestricted reference prior for any regular location problem may be shown to be uniform. Thus, the restricted reference prior must be of the form $\pi(\theta \,|\, \mathcal{P}) \propto \exp\{\gamma_1 \theta + \gamma_2 (\theta - \mu_0)^2\}$, with $\int_\Theta \theta\, \pi(\theta \,|\, \mathcal{P})\, d\theta = \mu_0$ and $\int_\Theta (\theta - \mu_0)^2\, \pi(\theta \,|\, \mathcal{P})\, d\theta = \sigma_0^2$. Hence, $\pi(\theta \,|\, \mathcal{P})$ is a *normal* distribution with the specified mean and variance.

Frequentist Properties

Bayesian methods provide a *direct* solution to the problems typically posed in statistical inference; indeed, posterior distributions precisely state what can be said about unknown quantities of interest *given* available data and prior knowledge. In particular, unrestricted reference posterior distributions state what could be said if no prior knowledge about the quantities of interest were available.

A frequentist analysis of the behavior of Bayesian procedures under repeated sampling may, however, be illuminating, for this provides some interesting connections between frequentist and Bayesian inference. It is found that the frequentist properties of Bayesian reference procedures are typically excellent, and may be used to provide a form of calibration for reference posterior probabilities.

Point Estimation. It is generally accepted that, as the sample size increases, a "good" estimator $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ ought to get the correct value of $\boldsymbol{\theta}$ eventually, that is to be *consistent*. Under appropriate regularity conditions, any Bayes estimator $\boldsymbol{\phi}^*$ of any function $\boldsymbol{\phi}(\boldsymbol{\theta})$ converges in probability to $\boldsymbol{\phi}(\boldsymbol{\theta})$, so that sequences of Bayes estimators are typically *consistent*. Indeed, it is known that if there is a consistent sequence of estimators, then Bayes estimators are consistent. The rate of convergence is often best for reference Bayes estimators.

It is also generally accepted that a "good" estimator should be *admissible*, that is, *not dominated* by any other estimator in the sense that its expected loss under sampling (conditional to $\boldsymbol{\theta}$) cannot be larger for all $\boldsymbol{\theta}$ values than that corresponding to another estimator. Any *proper* Bayes estimator is admissible; moreover, as established by Wald (1950), a procedure *must* be Bayesian (proper or improper) to be admissible. Most published admissibility results refer to quadratic loss functions, but they often extend to more general loss functions. Reference Bayes estimators are typically admissible with respect to intrinsic loss functions.

Notice, however, that many other apparently intuitive frequentist ideas on estimation have been proved to be potentially misleading. For example, given a sequence of n

Bernoulli observations with parameter θ resulting in r positive trials, the *best unbiased* estimate of θ^2 is found to be $r(r-1)/\{n(n-1)\}$, which yields $\tilde{\theta}^2 = 0$ when $r = 1$; but to estimate the probability of two positive trials as zero, when one positive trial has been observed, is not at all sensible. In marked contrast, any Bayes reference estimator provides a reasonable answer. For example, the intrinsic estimator of θ^2 is simply $(\theta^*)^2$, where θ^* is the intrinsic estimator of θ described in section "►Estimation". In particular, if $r = 1$ and $n = 2$ the intrinsic estimator of θ^2 is (as one would naturally expect) $(\theta^*)^2 = 1/4$.

Interval Estimation. As the sample size increases, the frequentist coverage probability of a posterior q-credible region typically converges to q so that, for *large samples*, Bayesian credible intervals may (under regularity conditions) be interpreted as *approximate* frequentist confidence regions: under repeated sampling, a Bayesian q-credible region of $\boldsymbol{\theta}$ based on a large sample will cover the true value of $\boldsymbol{\theta}$ approximately $100q\%$ of times. Detailed results are readily available for univariate problems. For instance, consider the probability model $\{p(D\,|\,\boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, let $\theta = \theta(\boldsymbol{\omega})$ be any univariate quantity of interest, and let $\boldsymbol{t} = \boldsymbol{t}(D) \in T$ be any sufficient statistic. If $\theta_q(\boldsymbol{t})$ denotes the $100q\%$ quantile of the posterior distribution of θ which corresponds to some unspecified prior, so that

$$\Pr[\theta \le \theta_q(\boldsymbol{t})\,|\,\boldsymbol{t}] = \int_{\theta \le \theta_q(\boldsymbol{t})} p(\theta\,|\,\boldsymbol{t})\, d\theta = q,$$

then the coverage probability of the q-credible interval $\{\theta; \theta \le \theta_q(\boldsymbol{t})\}$,

$$\Pr[\theta_q(\boldsymbol{t}) \ge \theta\,|\,\boldsymbol{\omega}] = \int_{\theta_q(\boldsymbol{t}) \ge \theta} p(\boldsymbol{t}\,|\,\boldsymbol{\omega})\, d\boldsymbol{t},$$

is such that $\Pr[\theta_q(\boldsymbol{t}) \ge \theta\,|\,\boldsymbol{\omega}] = \Pr[\theta \le \theta_q(\boldsymbol{t})\,|\,\boldsymbol{t}] + O(n^{-1/2})$. This *asymptotic* approximation is true for *all* (sufficiently regular) positive priors. However, the approximation is better, actually $O(n^{-1})$, for a particular class of priors known as (first-order) *probability matching* priors. Reference priors are typically found to be probability matching priors, so that they provide this improved asymptotic agreement. As a matter of fact, the agreement (in regular problems) is typically quite good even for relatively small samples.

Example 16 (Product of normal means). Consider the case where independent random samples $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$ have respectively been taken from the normal densities $N(x\,|\,\omega_1, 1)$ and $N(y\,|\,\omega_2, 1)$, and suppose that the quantity of interest is the product of their means, $\phi = \omega_1\omega_2$ (for instance, one may be interested in inferences about the area ϕ of a rectangular piece of land, given measurements $\{x_i\}$ and $\{y_j\}$ of its sides). Notice that this is a simplified version of a problem that it is often encountered in the sciences, where one is interested in the product of several magnitudes, all of which have been measured with error. Using the procedure described in Example 13, with the natural approximating sequence induced by $(\omega_1, \omega_2) \in [-i, i]^2$, the ϕ-reference prior is found to be

$$\pi_\phi(\omega_1, \omega_2) \propto \left(n\,\omega_1^2 + m\,\omega_2^2\right)^{-1/2},$$

very different from the uniform prior $\pi_{\omega_1}(\omega_1, \omega_2) = \pi_{\omega_2}(\omega_1, \omega_2) = 1$ which should be used to make objective inferences about either ω_1 or ω_2. The prior $\pi_\phi(\omega_1, \omega_2)$ may be shown to provide approximate agreement between Bayesian credible regions and frequentist confidence intervals for ϕ; indeed, this prior (with $m = n$) was originally suggested by Stein in the 1980s to obtain such approximate agreement. The same example was later used by Efron (1986) to stress the fact that, even within a fixed probability model $\{p(D\,|\,\boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$, the prior required to make objective inferences about some function of the parameters $\phi = \phi(\boldsymbol{\omega})$ must generally depend on the function ϕ.

The numerical agreement between reference Bayesian credible regions and frequentist confidence intervals is actually perfect in special circumstances. Indeed, as Lindley (1958) pointed out, this is the case in those problems of inference which may be transformed to location-scale problems.

Example 3 (Inference on normal parameters, continued). Let $D = \{x_1, \ldots x_n\}$ be a random sample from a normal distribution $N(x\,|\,\mu, \sigma)$. As mentioned before, the reference posterior of the quantity of interest μ is the Student distribution $\mathrm{St}\left(\mu\,|\,\bar{x}, s/\sqrt{n-1}, n-1\right)$. Thus, normalizing μ, the *posterior* distribution of $t(\mu) = \sqrt{n-1}(\bar{x} - \mu)/s$, as a function of μ given D, is the standard Student $\mathrm{St}(t\,|\,0, 1, n-1)$ with $n-1$ degrees of freedom. On the other hand, this function t is recognized to be precisely the conventional t statistic, whose *sampling distribution* is well known to *also* be standard Student with $n-1$ degrees of freedom. It follows that, *for all sample sizes*, posterior *reference* credible intervals for μ given the data will be *numerically identical* to frequentist confidence intervals based on the sampling distribution of t.

A similar result is obtained in inferences about the variance. Thus, the reference *posterior* distribution of $\lambda = \sigma^{-2}$ is the ►Gamma distribution $\mathrm{Ga}(\lambda\,|\,(n-1)/2, ns^2/2)$ and, hence, the *posterior* distribution of $r = ns^2/\sigma^2$, as a function of σ^2 given D, is a (central) χ^2 with $n-1$ degrees of freedom. But the function r is recognized to be a conventional statistic for this problem, whose *sampling distribution* is well known to *also* be χ^2 with $n-1$ degrees of freedom. It follows that, *for all sample sizes*, posterior *reference* credible intervals for σ^2 (or any one-to-one

function of σ^2) given the data will be *numerically identical* to frequentist confidence intervals based on the sampling distribution of r.

For a recent review or modern objective Bayesian inference, see Bernardo (2010).

Discussion

In writing a broad article it is always hard to decide what to leave out. This article focused on the basic concepts of the Bayesian paradigm; methodological topics which have unwillingly been omitted include design of experiments, sample surveys, linear models and sequential methods. The interested reader is referred to the bibliography for further information. This final section briefly reviews the main arguments for the Bayesian approach, and includes pointers to further issues which have not been discussed in more detail due to space limitations.

Coherence

By using probability distributions to characterize *all* uncertainties in the problem, the Bayesian paradigm reduces statistical inference to applied probability, thereby ensuring the coherence of the proposed solutions. There is no need to investigate, on a case by case basis, whether or not the solution to a particular problem is logically correct: a Bayesian result is only a *mathematical consequence of explicitly stated assumptions* and hence, unless a logical mistake has been committed in its derivation, it cannot be formally wrong. In marked contrast, conventional statistical methods are plagued with counterexamples. These include, among many others, negative estimators of positive quantities, q-confidence regions ($q < 1$) which consist of the whole parameter space, empty sets of "appropriate" solutions, and incompatible answers from alternative methodologies simultaneously supported by the theory.

The Bayesian approach does require, however, the specification of a (prior) probability distribution over the parameter space. The sentence "a prior distribution does not exist for this problem" is often stated to justify the use of non-Bayesian methods. However, the general representation theorem *proves the existence* of such a distribution whenever the observations are assumed to be exchangeable (and, if they are assumed to be a random sample then, *a fortiori*, they are assumed to be exchangeable). To ignore this fact, and to proceed as if a prior distribution did not exist, just because it is not easy to specify, is mathematically untenable.

Objectivity

It is generally accepted that any statistical analysis is subjective, in the sense that it is always conditional on accepted assumptions (on the structure of the data, on the probability model, and on the outcome space) and those assumptions, although possibly well founded, are definitely *subjective* choices. It is, therefore, mandatory to make all assumptions very explicit.

Users of conventional statistical methods rarely dispute the mathematical foundations of the Bayesian approach, but claim to be able to produce "objective" answers in contrast to the possibly subjective elements involved in the choice of the prior distribution.

Bayesian methods do indeed require the choice of a prior distribution, and critics of the Bayesian approach systematically point out that in many important situations, including scientific reporting and public decision making, the results must exclusively depend on documented data which might be subject to independent scrutiny. This is of course true, but those critics choose to ignore the fact that this particular case is covered within the Bayesian approach by the use of *reference* prior distributions which (1) are mathematically derived from the accepted probability model (and, hence, they are "objective" insofar as the choice of that model might be objective) and, (2) by construction, they produce posterior probability distributions which, given the accepted probability model, *only* contain the information about their values which data may provide and, *optionally*, any further contextual information over which there might be universal agreement.

An issue related to objectivity is that of the operational meaning of reference posterior probabilities; it is found that the analysis of their behavior under repeated sampling provides a suggestive form of calibration. Indeed, $\Pr[\boldsymbol{\theta} \in R \,|\, D] = \int_R \pi(\boldsymbol{\theta} \,|\, D)\, d\boldsymbol{\theta}$, the reference posterior probability that $\boldsymbol{\theta} \in R$, is *both* a measure of the conditional uncertainty (given the assumed model and the observed data D) about the event that the unknown value of $\boldsymbol{\theta}$ belongs to $R \subset \Theta$, and the limiting proportion of the regions which would cover $\boldsymbol{\theta}$ under repeated sampling conditional on data "sufficiently similar" to D. Under broad conditions (to guarantee regular asymptotic behavior), all large data sets from the same model are "sufficiently similar" among themselves in this sense and hence, given those conditions, reference posterior credible regions are *approximate* unconditional frequentist confidence regions.

The conditions for this approximate *unconditional* equivalence to hold exclude, however, important special cases, like those involving "extreme" or "relevant" observations. In very special situations, when probability models may be transformed to location-scale models, there is an exact unconditional equivalence; in those cases reference posterior credible intervals are, for any sample size, exact unconditional frequentist confidence intervals.

Applicability

In sharp contrast to most conventional statistical methods, which may only be exactly applied to a handful of relatively simple stylized situations, Bayesian methods are (in theory) totally general. Indeed, for a given probability model and prior distribution over its parameters, the derivation of posterior distributions is a well-defined mathematical exercise. In particular, Bayesian methods do not require any particular regularity conditions on the probability model, do not depend on the existence of sufficient statistics of finite dimension, do not rely on asymptotic relations, and do not require the derivation of any sampling distribution, nor (a fortiori) the existence of a "pivotal" statistic whose sampling distribution is independent of the parameters.

However, when used in complex models with many parameters, Bayesian methods often require the computation of multidimensional definite integrals and, for a long time in the past, this requirement effectively placed practical limits on the complexity of the problems which could be handled. This has dramatically changed in recent years with the general availability of large computing power, and the parallel development of simulation-based numerical integration strategies like *importance sampling* or ►*Markov chain Monte Carlo* (MCMC). These methods provide a structure within which many complex models may be analyzed using generic software. MCMC is numerical integration using Markov chains. Monte Carlo integration proceeds by drawing samples from the required distributions, and computing sample averages to approximate expectations. MCMC methods draw the required samples by running appropriately defined ►Markov chains for a long time; specific methods to construct those chains include the Gibbs sampler and the Metropolis algorithm, originated in the 1950s in the literature of statistical physics. The development of improved algorithms and appropriate diagnostic tools to establish their convergence, remains a very active research area.

Actual scientific research often requires the use of models that are far too complex for conventional statistical methods. This article concludes with a glimpse at some of them.

Hierarchical structures. Consider a situation where a possibly variable number n_i of observations, $\{\boldsymbol{x}_{ij}, j = 1, \ldots, n_i\}$, $i = 1, \ldots, m$, are made on each of m internally homogeneous subsets of some population. For instance, a firm might have chosen m production lines for inspection, and n_i items might have been randomly selected among those made by production line i, so that $\boldsymbol{x}_{ij}$ is the result of the measurements made on item j of production line i. As another example, animals of some species are captured to study their metabolism, and a blood sample taken before releasing them again; the procedure is repeated in the same habitat for some time, so that some of the animals are recaptured several times, and $\boldsymbol{x}_{ij}$ is the result of the analysis of the j-th blood sample taken from animal i. It those situations, it is often appropriate to assume that the n_i observations on subpopulation i are exchangeable, so that they may be treated as a random sample from some model $p(\boldsymbol{x} \mid \boldsymbol{\theta}_i)$ indexed by a parameter $\boldsymbol{\theta}_i$ which depends on the subpopulation observed, and that the parameters which label the subpopulations may also be assumed to be exchangeable, so that $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ may be treated as a random sample from some distribution $p(\boldsymbol{\theta} \mid \boldsymbol{\omega})$. Thus, the complete *hierarchical* model which is assumed to have generated the observed data $D = \{\boldsymbol{x}_{11}, \ldots, \boldsymbol{x}_{mn_m}\}$ is of the form

$$p(D \mid \boldsymbol{\omega}) = \int_{\Theta^m} \left[\prod_{j=1}^{n_i} p(\boldsymbol{x}_{ij} \mid \boldsymbol{\theta}_i) \right] \left[\prod_{i=1}^{m} p(\boldsymbol{\theta}_i \mid \boldsymbol{\omega}) \right] \left[\prod_{i=1}^{m} d\boldsymbol{\theta}_i \right].$$

Hence, under the Bayesian paradigm, a family of conventional probability models, say $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, and an appropriate "structural" prior $p(\boldsymbol{\theta} \mid \boldsymbol{\omega})$, may be naturally combined to produce a versatile, complex model $\{p(D \mid \boldsymbol{\omega}), \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$ whose analysis is often well beyond the scope of conventional statistics. The Bayesian solution only requires the specifiction a prior distribution $p(\boldsymbol{\omega})$, the use Bayes' theorem to obtain the corresponding posterior $p(\boldsymbol{\omega} \mid D) \propto p(D \mid \boldsymbol{\omega})\, p(\boldsymbol{\omega})$, and the performance of the appropriate probability transformations to derive the posterior distributions of the quantities of interest (which may well be functions of $\boldsymbol{\omega}$, functions of the $\boldsymbol{\theta}_i$'s, or functions of future observations). As in any other Bayesian analysis, the prior distribution $p(\boldsymbol{\omega})$ has to describe available knowledge about $\boldsymbol{\omega}$; if none is available, or if an objective analysis is required, an appropriate reference prior function $\pi(\boldsymbol{\omega})$ may be used.

Contextual information. In many problems of statistical inference, objective and universally agreed contextual information is available on the parameter values. This information is usually very difficult to handle within the framework of conventional statistics, but it is easily incorporated into a Bayesian analysis by simply restricting the prior distribution to the class $\mathcal{P}$ of priors which are compatible with such information. As an example, consider the frequent problem in archaeology of trying to establish the occupation period $[\alpha, \beta]$ of a site by some past culture on the basis of the radiocarbon dating of organic samples taken from the excavation. Radiocarbon dating is not precise, so that each dating x_i is typically taken to be a normal observation from a distribution $N(x \mid \mu(\theta_i), \sigma_i)$, where θ_i is the actual, unknown calendar date of the sample, $\mu(\theta)$

is an internationally agreed calibration curve, and σ_i is a known standard error quoted by the laboratory. The actual calendar dates $\{\theta_1, \ldots, \theta_m\}$ of the samples are typically assumed to be uniformly distributed within the occupation period $[\alpha, \beta]$; however, stratigraphic evidence indicates some partial orderings for, if sample i was found on top of sample j in undisturbed layers, then $\theta_i > \theta_j$. Thus, if $\mathcal{C}$ denotes the class of values of $\{\theta_1, \ldots, \theta_m\}$ which satisfy those known restrictions, data may be assumed to have been generated by the hierarchical model

$$p(x_1, \ldots, x_m \mid \alpha, \beta) = \int_{\mathcal{C}} \left[\prod_{i=1}^{m} N\left(x_i \mid \mu(\theta_i), \sigma_i^2\right) \right] (\beta - \alpha)^{-m} \, d\theta_1 \ldots d\theta_m.$$

Often, contextual information further indicates an absolute lower bound α_0 and an absolute upper bound β_0 for the period investigated, so that $\alpha_0 < \alpha < \beta < \beta_0$. If no further documented information is available, the corresponding restricted reference prior for the quantities of interest, $\{\alpha, \beta\}$ should be used; this is found to be $\pi(\alpha, \beta) \propto (\beta - \alpha)^{-1}$ whenever $\alpha_0 < \alpha < \beta < \beta_0$ and zero otherwise. The corresponding reference posterior $\pi(\alpha, \beta \mid x_1, \ldots, x_m) \propto p(x_1, \ldots, x_m \mid \alpha, \beta)\, \pi(\alpha, \beta)$ summarizes all available information on the occupation period.

Covariate information. Over the last 30 years, both linear and non-linear regression models have been analyzed from a Bayesian point of view at increasing levels of sophistication. These studies range from the elementary objective Bayesian analysis of simple linear regression structures (which parallel their frequentist counterparts) to the sophisticated analysis of time series involved in dynamic forecasting which often make use of complex hierarchical structures. The field is far too large to be reviewed in this article, but the bibliography contains some relevant pointers.

Model Criticism. It has been stressed that *any* statistical analysis is conditional on the accepted assumptions of the probability model which is presumed to have generated the data. Recent years have shown a huge effort into the development of Bayesian procedures for *model criticism* and *model choice*. Most of these are sophisticated elaborations of the procedures described in section "►Hypothesis Testing" under the heading of hypothesis testing. Again, this is too large a topic to be reviewed here, but some key references are included in the bibliography.

Acknowledgments

Work has been partially funded with Grant MTM2006-07801 of the MEC, Spain.

About the Author

Professor José Bernardo is founder co-President of the International Society for Bayesian Analysis (1992–1994). He is the personal organizer and Programme Committee Member of the Valencia International Meetings on Bayesian Statistics, established world forums on Bayesian Methods, held every 4 years in Spain since 1979. Spanish Royal Academy of Science invited Professor Bernardo to become the Editor of the special issue of its journal *the Rev. Acad. Cien. Madrid*, devoted entirely to Bayesian Methods in the Sciences (1999). He has co-edited (as the first author) eight proceedings of the Valencia meetings on Bayesian Statistics. Professor Bernardo is a Fellow of the American Statistical Association, Royal Statistical Society and Elected Member of the International Statistical Institute. He was Founding Editor of *Test* (1992–1997), associate editor for the *Journal of the Royal Statistical Society* (Series B) (1989–1993), *The Statistician* (1987–1997), *Questio* (1983–2002); Contributing Editor of the *Current Index of Statistics* and *Statistics Theory and Methods Abstracts* (1996–2003). Currently he is an Associate editor for the *Journal of the Iranian Statistical Society*. He was delivering lectures at more than 60 universities worldwide. He is a co-author (with Adrian Smith) of the well known text *Bayesian Theory* (Wiley, 1994).

Cross References

►Bayes' Theorem
►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Nonparametric Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Foundations of Probability
►Likelihood
►Markov Chain Monte Carlo
►Model Selection
►Prior Bayes: Rubin's View of Statistics
►Statistical Inference
►Statistical Inference: An Overview
►Statistics: An Overview

References and Further Reading

Bartlett M (1957) A comment on D.V. Lindley's statistical paradox. Biometrika 44:533–534
Berger JO (1985) Statistical decision theory and Bayesian analysis. Springer, Berlin
Berger JO (2000) Bayesian analysis: a look at today and thoughts of tomorrow. J Am Stat Assoc 95:1269–1276
Berger JO, Bernardo JM (1989) Estimating a product of means: Bayesian analysis with reference priors. J Am Stat Assoc 84:200–207

Berger JO, Bernardo JM (1992) On the development of reference priors. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian statistics, vol 4. Oxford University Press, Oxford, pp 35–60 (with discussion)
Berger J, Bernardo JM, Sun D (2009a) The formal definition of reference priors. Ann Stat 37:905–938
Berger JO, Bernardo JM, Sun D (2009b) Natural induction: an objective Bayesian approach. Rev Acad Sci Madrid A 103:125–159 (with discussion)
Bernardo JM (1979a) Expected information as expected utility. Ann Stat 7:686–690
Bernardo JM (1979b) Reference posterior distributions for Bayesian inference. J R Stat Soc B 41: 113–147 (with discussion). In: Tiao GC, Polson GC (eds) Reprinted in Bayesian Inference 1. Edward Elgar, Oxford, pp 229–263
Bernardo JM (1997) Noninformative priors do not exist. J Stat Plann Infer 65:159–189 (with discussion)
Bernardo JM (2005a) Reference analysis. In: Dey DK, Rao CR (eds) Handbook of Statistics, vol 25. Elsevier, Amsterdam, pp 17–90
Bernardo JM (2005b) Intrinsic credible regions: An objective Bayesian approach to interval estimation. Test 14:317–384 (with discussion)
Bernardo JM (2010) Integrated objective Bayesian estimation and hypothesis testing. In: Bernardo JM et al. (eds) Bayesian Statistics 9. Oxford: Oxford University Press, (to appear, with discussion)
Bernardo JM, Ramón JM (1998) An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. The Statistician 47:1–35
Bernardo JM, Rueda R (2002) Bayesian hypothesis testing: a reference approach. Int Stat Rev 70:351–372
Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, Chichester
Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) (2003) Bayesian statistics 7. Oxford University Press, Oxford
Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) (2007) Bayesian statistics 8. Oxford University Press, Oxford
Berry DA (1996) Statistics, a Bayesian perspective. Wadsworth, Belmont
Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Addison-Wesley, Reading
Dawid AP, Stone M, Zidek JV (1973) Marginalization paradoxes in Bayesian and structural inference. J R Stat Soc B 35:189–233 (with discussion)
de Finetti B (1970) Teoria delle Probabilità. Einaudi, Turin. English translation: Theory of Probability (1975) Wiley, Chichester
DeGroot MH (1970) Optimal statistical decisions. McGraw-Hill, New York
de Finetti B (1937) La prévision, ses lois logiques, ser sources sunjecives. Ann Inst Henri Poincaré 7:1–68
Efron B (1986) Why isn't everyone a Bayesian? Am Stat 40:1–11 (with discussion)
Geisser S (1993) Predictive inference: an introduction. Chapman and Hall, London
Gelfand AE, Smith AFM (1990) Sampling based approaches to calculating marginal densities. J Am Stat Assoc 85: 398–409
Gelman A, Carlin JB, Stern H, Rubin DB (1995) Bayesian data analysis. Chapman and Hall, London
Gilks WR, Richardson S, Spiegelhalter DJ (1996) Markov chain Monte Carlo in practice. Chapman and Hall, London
Jaynes ET (1976) Confidence intervals vs Bayesian intervals. In: Harper WL, Hooker CA (eds) Foundations of probability theory, statistical inference and statistical theories of science, vol 2. Reidel, Dordrecht, pp 175–257 (with discussion)
Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, Oxford
Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90: 773–795
Laplace PS (1812) Théorie Analitique des Probabilités. Paris: Gauthier-Villars
Lindley DV (1957) A statistical paradox. Biometrika 44:187–192
Lindley DV (1958) Fiducial distribution and Bayes theorem. J R Stat Soc B 20:102–107
Lindley DV (1965) Introduction to probability and statistics from a Bayesian viewpoint. Cambridge University Press, Cambridge
Lindley DV (1972) Bayesian Statistics, a review. SIAM, Philadelphia
Lindley DV (1985) Making Decisions, 2nd edn. Wiley, Chichester
Lindley DV (2000) The philosophy of statistics. The Statistician 49:293–337 (with discussion)
O'Hagan A (1994) Bayesian Inference. Edward Arnold, London
Press SJ (1972) Applied multivariate analysis: using Bayesian and frequentist methods of inference. Krieger, Melbourne
Ramsey FP (1931) Truth and probability. In: Braithwaite RB (ed) The foundations of mathematics and other logical essays. London: Kegan Paul, pp 156–198
Wald A (1950) Statistical decision functions. Wiley, Chichester
West M, Harrison PJ (1989) Bayesian forecasting and dynamic models. Springer, Berlin
Zellner A (1971) An introduction to Bayesian inference in econometrics. Wiley, New York. Reprinted in 1987, Krieger, Melbourne

Bayesian Versus Frequentist Statistical Reasoning

Jordi Vallverdú
Universitat Autònoma de Barcelona, Catalonia, Spain

We can consider the existence of two main statistical schools: Bayesian and frequentist. Both provide ways to deal with probability, although their methods and theories are mutually exclusive (Vallverdú 2008).

Bayesian Statistics

From a historical perspective, Bayesian appeared first, in 1763, when Richard Price published posthumously the paper of late Rev. Thomas Bayes "An Essay towards solving a Problem in the Doctrine of Chances" (Dale 2003). In this paper, Bayes presented his ideas about the best way

of dealing with probability (and trying to solve the problem of *inverse probability*), which can be exemplified today with the classic formula called "Bayes' Rule" or "Bayes' Theorem":

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

We must look at the notation and terminology involved:

- $P(A|B)$ is the *conditional probability* of A, given B. It is also called the *posterior probability* because it is derived from or depends upon the specified value of B.
- $P(B|A)$ is the conditional probability of B given A.
- $P(A)$ is the *prior probability* or *marginal probability* of A. It is "prior" in the sense that it does not take into account any information about B.
- $P(B)$ is the prior or marginal probability of B, and acts as a *normalizing constant*.

We can see, then, that our *posterior* belief $P(A|B)$ is calculated by multiplying our *prior* belief $P(A)$ by the *likelihood* $P(B|A)$ that B will occur if A is true. Although Bayes' method was enthusiastically taken up by Laplace and other leading probabilists of the day, it fell into disrepute in the nineteenth century because they did not yet know how to handle *prior probabilities* properly. The *prior probability* of A represents our best estimate of the probability of the fact we are considering, prior to consideration of the new piece of evidence. Therefore, in the Bayesian paradigm, current knowledge about the model parameters is expressed by placing a probability distribution on the parameters, called the "prior distribution." When new data become available, the information they contain regarding the model parameters is expressed in the "likelihood," which is proportional to the distribution of the observed data given the model parameters. This information is then combined with the prior to produce an updated probability distribution called the "posterior distribution," on which all Bayesian inference is based. ►Bayes' Theorem, an elementary identity in probability theory, states how the update is done mathematically: the posterior is proportional to the prior times the likelihood.

There are a large number of types of Bayesians (speaking ironically, Good (1971) spoke of "46,656 kinds of Bayesians"), depending on their attitude toward subjectivity in postulating priors. Some recent Bayesian books are Earman (1992), Howson and Urbach (1991), Bernardo and Smith (1996).

Frequentist Statistics

On the other hand, we have the frequentist paradigm. Its followers understand probability as a long-run frequency of a "repeatable" event and developed a notion of confidence intervals. Probability would be, then, a measurable frequency of events determined from repeated experiments. We can express it as:

$$P(A) = n/N,$$

where n is the number of times event A occurs in N opportunities.

From the frequentist viewpoint two closely related methods have been developed. One is the Neyman–Pearson theory of significance tests and the other is based on Fisher's notion of ►*p*-values. The researchers who follow this approach, consider frequentism as the only allowed statistical method for achieving sound scientific inferences (Mayo and Cox 2006).

The Debate: Degrees of Belief Versus Relative Frequencies

As early as in 1949, Maurice George Kendall (1949) wrote a paper, "On the Reconciliation of Theories of Probability," in which he coined the word "frequentist" and stated: "Few branches of scientific method have been subject to so much difference of opinion as the theory of probability." He tried to attempt mediation between the contestants, but failed.

Clearly, one of the recurrent arguments against/in favor of one of the two positions (frequentist or Bayesian) consists in saying that a true scientist is always/never frequentist/Bayesian (you can *choose* between the two possibilities). As an example of this confrontation see the ideas of Giere (1988): "Are Scientists Bayesian Agents? (. . .) The overwhelming conclusion is that humans are not bayesian agents," and of Efron (1986) or Cousins (1995). The last two do not need to be quoted. It seems to be an epistemological law about statistical practices: "A true scientist never belongs to the opposite statistical school" (Vallverdú 2008).

It could seem that frequentists are realists, when they consider relative frequencies and that Bayesian are subjective, when they defend degrees of belief of prior probabilities but the truth is that in cluster investigations, for example, the frequentist approach is just as subjective as the Bayesian approach, although the Bayesian approach is less ambitious in that it treats the analysis as a synthesis of data and personal judgments (possibly poor ones), rather than objective reality (Coory et al. 2009).

Why to Become Frequentist/Bayesian?

Bland and Altman (1998, p. 1160) have their own answer: "Most statisticians have become Bayesians or Frequentists as a result of their choice of university." And as the epidemiologist Berger (2003) says: "practicing epidemiologists are given little guidance in choosing between these approaches apart from the ideological adherence of mentors, colleagues and editors."

So, the arguments go beyond the ethereal philosophical arena and become more practical ones. Better opportunities to find a good job is an important argument, and the value of a Bayesian academic training is now accepted: "where once graduate students doing Bayesian dissertations were advised to try not to look too Bayesian when they went on the job market, now great numbers of graduate students try to include some Bayesian flavor in their dissertations to increase their marketability" (Wilson 2003, p. 372). Therefore, and following Hacking (1972, p. 133): "Euler at once retorted that this advice is metaphysical, not mathematical. Quite so! The choice of primitive concepts for inference *is* a matter of 'metaphysics'. The orthodox statistician has made one metaphysical choice and the Bayesian another." To be honest, there is not a fatally flawed position, but different context-based applications of both main approaches. As Gigerenzer et al. (1990) express "we need statistical thinking, not statistical rituals." Lilford and Braunholtz (1996, p. 604) go further: "when the situation is less clear cut (...) conventional statistics may drive decision makers into a corner and produce sudden, large changes in prescribing. The problem does not lie with any of the individual decision makers, but with the very philosophical basis of scientific inference. We propose that conventional statistics should not be used in such cases and that the Bayesian approach is both epistemologically and practically superior." There is also a structural aspect: computational facilities; due to recent innovations in scientific computing (faster computer processors) and drastic drops in the cost of computers, the number of statisticians trained in Bayesian methodology has increased (Tan 2001). Trying to offer a midpoint perspective, Berger (2003) proposes using both models and studying case by case their possibilities: "based on the philosophical foundations of the approaches, Bayesian models are best suited to addressing hypotheses, conjectures, or public-policy goals, while the frequentist approach is best suited to those epidemiological studies which can be considered 'experiments', i.e., testing constructed sets of data." Usually, we find no such equitable position.

Considering all these facts, we can conclude that both frequentist and Bayesian statisticians use sound science in their researches and that, in the end, this debate is a deep philosophical one, not a matter of rational argument.

Acknowledgment

This research was supported by the Project "El diseño del espacio en entornos de cognición distribuida: plantillas y affordances," MCI [FFI2008-01559/FISO].

About the Author

For biography *see* the entry ►Probability, History of.

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Foundations of Probability
►Frequentist Hypothesis Testing: A Defense
►Likelihood
►Model Selection
►Philosophical Foundations of Statistics
►Prior Bayes: Rubin's View of Statistics
►Probability Theory: An Outline
►Probability, History of
►Significance Tests: A Critique
►Statistical Inference
►Statistical Inference: An Overview
►Statistics: An Overview
►Statistics: Nelder's view

References and Further Reading

Berger ZD (2003) Bayesian and frequentist models: legitimate choices for different purposes. AEP 13(8):583
Berger JA et al (1997) Unified frequentist and Bayesian testing of a precise hypothesis. Stat Sci 12(3):133–160
Bernardo JM, Smith AFM (1996) Bayesian theory. Wiley, USA
Bland MJ, Altman DG (1998) Bayesian and frequentists. Br Med J 317:1151–1160
Coory MD, Wills RA, Barnett AG (2009) Bayesian versus frequentist statistical inference for investigating a one-off cancer cluster reported to a health department. BMC Med Res Methodol 9:3
Cousins RD (1995) Why isn't every physicist a Bayesian? Am J Phys 63:398
Dale AI (2003) Most honourable remembrance. The life and work of Thomas Bayes. Springer, New York
Earman J (1992) Bayes or bust? MIT Press, Cambridge, MA
Efron B (1986) Why isn't everyone a Bayesian? Am Stat 40:1–5
Giere R (1988) Understanding scientific reasoning. University of Chicago, Chicago, p 189
Gigerenzer G et al (1990) The Empire of Chance. How probability changed science and everyday life. Cambridge University Press, Cambridge
Good IJ (1971) 46,656 kinds of Bayesians. Am Stat 25:62–63
Hacking I (1972) Likelihood. Brit J Philos Sci 23:132–137
Howson C, Urbach P (1991) Bayesian reasoning in science. Nature 350:371–374
Kendall MG (1949) On the Reconciliation of Theories of Probability. Biometrika 36:101–116
Lilford RJ, Braunholtz D (1996) For debate: the statistical basis of public policy: a paradigm shift is overdue. Br Med J 313:603–607
Mayo DG, Cox DR (2006) Frequentist statistics as a theory of inductive inference. 2nd Lehmann symposium - optimality IMS lecture notes - monographs series, 1–28
Tan SB (2001) Bayesian methods for medical research. Ann Acad Med 30(4):444–446
Vallverdú J (2008) The false dilemma: Bayesian versus Frequentist. E – L O G O S Electronic Journal for Philosophy 1–17
Wilson G (2003) Tides of change: is Bayesianism the new paradigm in statistics? J Stat Plan Infer 113:3171–374

Bayesian vs. Classical Point Estimation: A Comparative Overview

Francisco J. Samaniego
Professor
University of California-Davis, Davis, CA, USA

The foundations of the classical theory of point estimation are embedded in the work of Frederick Gauss, Karl Pearson and Ronald Fisher, though there have been many other contributors, as documented in Stigler's (1986) historical masterpiece or, in more technical terms, in Lehmann and Casella (1998). In the framework of independent, identically distributed (i.i.d.) observations, the theory seeks to obtain good estimators (or "best guesses") of an unknown scalar or vector-valued parameter θ based on a "random sample" of observations drawn from a distribution indexed by this parameter. The adjective "frequentist" is often used in referring to classical methods, largely because the theory of their performance is based on the premise that the experiment from which data is drawn can be replicated repeatedly and that estimators (and other statistical procedures) may be evaluated and compared on the basis of their expected performance over the intended number of replications or on a hypothesized infinite number of i.i.d. trials. Finite sample methods leading, for example, to ►least-squares, best-unbiased or best-invariant estimators, and estimators based on asymptotic theory, the premier example of which is the maximum likelihood estimators proposed by Fisher, are generally studied separately and are, together, the mainstays of the theory and practice of frequentist estimation. These are discussed in more detail below, as well as elsewhere in the Encyclopedia.

Bayesian estimation theory tends to start at the same place outlined above. It begins with a model for the observable data, and assumes the existence of data upon which inference about a target parameter will be based. The important point of departure from classical inference is the position that uncertainty should be treated stochastically. From this, it follows that since the target parameter in a point estimation problem is unknown, one's uncertainty about its value is appropriately represented by a "prior probability distribution" on that parameter. The Bayesian position is not simply a whim or a matter of convenience; it is in fact motivated and defended through a system of axioms about the comparison of possible uncertain events and the fact that the axiom system leads to this position as a derived result. See De Groot (1970) for further details. The Bayesian paradigm for estimation can be by described as involving three steps: the specification of a prior distribution (through introspection or the elicitation of expert opinion), the updating of that prior on the basis of available data, leading to the "posterior distribution" on the parameter, and the estimation of the parameter based on characteristics of the posterior distribution. The mean of the posterior distribution is probably the most commonly used Bayesian point estimator.

Comparisons between frequentist and Bayesian estimators raise some challenging issues. There are important philosophical differences between the approaches that make it difficult to compare them side by side. For example, the Likelihood Principle (see, for example, Berger and Wolpert (1984)) stipulates that inference about an unknown parameter should depend of the experiment only through the observed data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ or, equivalently, through the likelihood function, which for the present purposes, may be thought of as a constant multiple of the joint density

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta),$$

with $\mathbf{x}$ fixed and known. As a consequence, the process of averaging over the sample space, as the frequentist does, for example, in minimizing the variance among unbiased estimators of a parameter, is inappropriate from the Bayesian viewpoint. Only the observed data, rather than what might have been observed in the experiment but wasn't, is relevant in Bayesian inference about θ. Maximum likelihood estimation, which obeys the likelihood principle in the sense that its calculation relies solely on the maximization of $L(\theta|\mathbf{x})$, might thus seem consistent with the Bayesian approach, but it runs afoul of that paradigm on the basis of the facts that it fails to quantify the uncertainty about θ stochastically and its entire theoretical justification is based on the familiar averaging process over the sample space, albeit in a limiting sense.

There are many texts, monographs and research papers which treat Bayesian and frequentist estimators and offer some discussion on how one might compare them. In this brief review, the discussion of this literature is clearly out of the question. We will limit ourselves to arguments and theory that may be found in the books by Lehmann and Casella (1998), Robert (2001), Cox (2006) and in the paper by Samaniego and Reneau (1994). A more comprehensive treatment of the subject of the comparative analysis of Bayesian and frequentist point estimators can be found in the monograph by Samaniego (2010). The references cited in the sources above are wide in scope and will provide the interested reader with an enormous collection of collateral reading that is highly relevant to the subject of interest here.

When one seeks to compare the Bayesian and frequentist approaches to point estimation, one might begin with a

foundational issue, namely, which approach has defensible logical underpinnings. As it happens, this is a rather easy matter to resolve. The classical theory of estimation has no logical underpinnings to speak of. The theories of unbiased estimators, invariant estimators and maximum likelihood estimators (MLEs) are justified by frequentists on intuitive grounds. For example, "unbiasedness" and "invariance" are intuitively appealing ways of restricting the class of all possible estimators, a class which is "too large" and contains no "best estimator" in any nontrivial problem. Maximum likelihood estimators "work well" based on a sample of infinite size, so perhaps they will work well the given finite-sample problem at hand. The frequentist approach to estimation is unabashedly *ad hoc*, and there is widespread recognition that the particular frequentist tool one might wish to use in different problems might vary, with least squares (or its well-known variants) used in regression and ANOVA problems, UMVUEs used in finite-sample problems in which their derivation is feasible and MLEs used in estimation problems in which the sample size is thought to be suitably large. In contrast with these circumstances, the Bayesian approach to point estimation is based on a system of "plausible" axioms about how one should deal with uncertainty. The fact that the Bayesian approach to statistical inference is built upon a logical basis which one can formally study and scrutinize places it on a higher "logical" plane than frequentist inference. Most readers of the axioms of Bayesian inference will find them "reasonable"; for example, the transitivity of one's assessments about which events are more likely to occur than others is typical of the assumptions one is expected to make. On logical grounds, the Bayesian approach appears to have a strong advantage. Is that enough to justify the claim that Bayesian inference is the preferred approach? The approach does provide a logically consistent process of prior assessment, updating and posterior assessment, leading to inference that can clearly be defended on logical grounds. A disturbing counterargument is that logic, by itself, is not enough. After all, it is possible to be perfectly logical and yet woefully wrong. Poor prior specification can lead to logically consistent estimation in which the Bayesian estimator is way off the mark. Other considerations must be brought to bear on these comparisons.

Bayesian estimation (as well as other modes of Bayesian inference) has a fairly apparent Achilles heal, the fact that the Bayesian interjects subjective prior information into the inference process, potentially infusing errors or biases that would not otherwise be present. This is a legitimate concern, and one that the "orthodox" Bayesian (that is, one who uses a proper probability distribution for his prior) must always be mindful of and careful about. While the usual counterargument does diffuse the criticism somewhat, a general fix does not exist. The argument that helps mitigate the criticism is that it is often the case that the Bayesian approach is essential in making sensible inferences. For example, a frequentist who observes ten heads in ten tosses of a newly minted coin has no choice but to estimate the probability of heads as $p = 1$, while a typical Bayesian, using a quite reasonable beta prior, might instead estimate p to be about 0.52. In situations such as this, Bayesian methods may be seen as a reasonable way of averting disaster.

There are many other grounds on which comparisons can be made. Regarding asymptotic performance for example, it is known that proper Bayes estimators based on priors that place positive mass on all open subsets of the parameter space enjoy the same asymptotic properties as the best asymptotically normal (BAN) estimators of frequentist estimation theory (most notably, MLEs). Multivariate analysis poses greater challenges to the Bayesian than to the frequentist, largely due to the difficulty of obtaining and quantifying useful prior information on high-dimensional parameters. The frequentists would seem to have a bit of an advantage in this area. Robert (2001) employs decision theoretic arguments quite prominently in his defense of Bayesian methods. One of the centerpieces of that defense is the so-called complete class theorem which, in essence, says that any admissible decision rule (here, estimator) is Bayes (or nearly Bayes) with respect to some prior. This statement begs the question: why ever use anything else? To some, this argument seems to have obvious weaknesses, and the question is rather easy to answer. The fact that an estimator belongs to a class that contains all the "good" estimators hardly justifies its use. After all, the estimator that always estimates the scalar parameter θ to be 5 is admissible but would (essentially) never be recommended for use. In addition to the inconclusive arguments above, there are examples on both sides that show that the other approach gives silly answers to particular statistical questions. Bayesians often use the term "incoherent" when referring to frequentist procedures of this sort. To the frequentist school, the most pressing concern is the very real possibility of obtaining poor answers from a Bayesian analysis due to poor prior input.

Samaniego and Reneau (1994) offer a quite different, performance-based comparison of Bayesian and frequentist estimators – the Bayes risk of an estimator relative to the "truth", with the latter modeled as a (possibly degenerate) "true prior distribution" on the unknown parameter. The threshold problem (i.e., finding the boundary separating Bayes estimators that outperform the frequentist estimator of choice from Bayes estimators that don't) is introduced. Explicit solutions to the threshold problem are

obtained in the context of sampling distributions belonging to one-parameter exponential families, conjugate prior families and squared error loss. In Samaniego (2010), subsequent extensions of this work to the estimation of high-dimensional parameters and to estimation under asymmetric loss are treated. While the "true state" of the target parameter remains unknown throughout these analyses, it is seen that useful practical guidance can nevertheless be gleaned from them. In many problems, the class of Bayes estimators that are superior to frequentist alternatives is surprisingly broad. Bayesians who are both misguided (with a poorly centered prior distribution) and stubborn (with a tightly concentrated prior distribution) will generally not do well. Interestingly, it is shown that one flaw or the other need not be fatal by itself. But perhaps the most important conclusion of these studies is the simple fact that that neither the Bayesian nor the frequentist approach to point estimation will be uniformly dominant in any well-defined point estimation problem. In all such problems, there will be a threshold separating "good" priors from "bad" ones, and the remaining challenge, one that is by no means trivial, is trying to make a sensible judgment about which side of the threshold one is on, given the prior information one has in hand. This examination may lead one to a Bayes estimator or to a frequentist estimator in the particular problem of interest.

About the Author

Francisco Samaniego holds the title of Distinguished Professor of Statistics at the University of California, Davis. He has authored two recent Springer monographs, the first on "system signatures," a notion that he introduced to the engineering literature in a 1985 IEEE paper, and the second on "comparative statistical inference" focusing on the relative performance of Bayesian and frequentist point estimators. He is a Fellow of the ASA, IMS, RSS and a member of the ISI. He is a former Editor of *The Journal of the American Statistical Association*. In 2008, he received the U.S. Army Samuel S. Wilks Award for career contributions to reliability.

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Estimation
►Estimation: An Overview
►Foundations of Probability
►Frequentist Hypothesis Testing: A Defense
►Likelihood
►Model Selection
►Philosophical Foundations of Statistics
►Prior Bayes: Rubin's View of Statistics
►Probability Theory: An Outline
►Probability, History of
►Significance Tests: A Critique
►Statistical Inference
►Statistical Inference: An Overview
►Statistics: An Overview

References and Further Reading

Berger JO, Wolpert R (1984) The likelihood principle. IMS Monograph Series, Hayward
Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge
De Groot MH (1970) Optimal statistical decisions. McGraw-Hill, New York
Lehmann E, Casella G (1998) Theory of point estimation, 2nd edn. Springer Verlag, New York
Robert C (2001) The Bayesian choice: a decision theoretic motivation, 2nd edn. Chapman and Hall, London
Samaniego FJ, Reneau DM (1994) Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. J Am Stat Assoc 89:947–957
Samaniego FJ (2010) A comparison of the Bayesian and frequentist approaches to estimation. Springer, New York
Stigler S (1986) The history of statistics: the measurement of uncertainty before 1900. Harvard University Press, Cambridge

Behrens–Fisher Problem

Allan S. Cohen, Seock-Ho Kim
Professors
University of Georgia, Athens, GA, USA

Introduction

The Behrens–Fisher problem is the problem in statistics of hypothesis testing and interval estimation regarding the difference between the means of two independent normal populations without assuming the variances are equal. The solution of this problem was first offered by Behrens (1929) and reformulated later by Fisher (1939) using

$$t' = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = t_1 \sin\theta - t_2 \cos\theta,$$

where the sample mean $\bar{x}_1$ and sample variance s_1^2 are obtained from the random sample of size n_1 from the normal distribution with mean μ_1 and variance σ_1^2, $t_1 = (\bar{x}_1 - \mu_1)/\sqrt{s_1^2/n_1}$ has a t distribution with $\nu_1 = n_1 - 1$ degrees of freedom, the respective quantities with subscript 2 are defined similarly, and $\tan\theta = (s_1/\sqrt{n_1})/(s_2/\sqrt{n_2})$ or

$\theta = \tan^{-1}[(s_1/\sqrt{n_1})/(s_2/\sqrt{n_2})]$. The distribution of t' is the Behrens–Fisher distribution. It is, hence, a mixture of the two t distributions.

Under the usual null hypothesis of H_0: $\mu_1 = \mu_2$, the test statistic t', can be obtained and compared with the percentage points of the Behrens–Fisher distribution. Tables for the Behrens–Fisher distribution are available from Fisher and Yates (1957) and Lindley and Scott (1995). The table entries are based on the four numbers: ν_1, ν_2, θ, and the Type I error rate α. For example, Fisher and Yates (1957) presented significance points of the Behrens–Fisher distribution in two tables, one for $\nu_1 = \nu_2 = 8, 12, 24, \infty$, $\theta = 0°, 15°, 30°, 45°, 60°, 75°, 90°$, and $\alpha = 0.05, 0.01$, and the other for ν_1 that is greater than $\nu_2 = 1, 2, 3, 4, 5, 6, 7$, $\theta = 0°, 15°, 30°, 45°, 60°, 75°, 90°$, and $\alpha = 0.10, 0.05, 0.02, 0.01$.

Using the Behrens–Fisher distribution, the $100(1-\alpha)\%$ interval that contains $\mu_1 - \mu_2$ can be constructed with

$$\bar{x}_1 - \bar{x}_2 \pm t'_{\alpha/2}(\nu_1, \nu_2, \theta)\sqrt{s_1^2/n_1 + s_2^2/n_2},$$

where the probability that $t' > t'_{\alpha/2}(\nu_1, \nu_2, \theta)$ is $\alpha/2$ or, equivalently, $\Pr[t' > t'_{\alpha/2}(\nu_1, \nu_2, \theta)] = \alpha/2$.

The Behrens–Fisher t' statistic and the Behrens–Fisher distribution are based on Fisher's (1935) fiducial approach. The approach is to find a fiducial probability distribution that is a probability distribution of a parameter from observed data. Consequently, the interval that involves $t'_{\alpha/2}(\nu_1, \nu_2, \theta)$ is referred to as the $100(1 - \alpha)\%$ fiducial interval.

Example

Driving times from a person's house to work were measured for two different routes with $n_1 = 5$ and $n_2 = 11$ (see Lehmann 1975, p. 83). The ordered data from the first route are 6.5, 6.8, 7.1, 7.3, 10.2 yielding $\bar{x}_1 = 7.580$ and $s_1^2 = 2.237$, and the data from the second route are 5.8, 5.8, 5.9, 6.0, 6.0, 6.0, 6.3, 6.3, 6.4, 6.5, 6.5 yielding $\bar{x}_2 = 6.136$ and $s_2^2 = 0.073$. It is assumed that the two independent samples were drawn from two normal distributions having means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. A researcher wants to know whether the average driving times differed for the two routes.

The test statistic under the null hypothesis of equal population means is $t' = 2.143$ with $\nu_1 = 4$, $\nu_2 = 10$, and $\theta = 83.078$. The result, $\Pr(t' > 2.143) = 0.049$, indicates the null hypothesis cannot be rejected at $\alpha = 0.05$ when the alternative hypothesis is non-directional, H_a: $\mu_1 \neq \mu_2$, because $p = 0.098$. The corresponding 95% interval for the population mean difference is $[-0.431, 3.308]$.

Other Solutions

The Student's t test (see ▶Student's t-Tests) for independent means can be used when the two population variances are assumed to be equal and $\sigma_1^2 = \sigma_2^2 = \sigma^2$,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2/n_1 + s_p^2/n_2}},$$

where the pooled variance that provides the estimate of the common population variance σ^2 is defined as $s_p^2 = \left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right]/(n_1 + n_2 - 2)$. It has a t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. The example data yield the Student's $t = 3.220$, $\nu = 14$, the two-tailed $p = 0.006$, and the 95% confidence interval of $[0.482, 2.405]$. The null hypothesis of equal population means is rejected at the nominal $\alpha = 0.05$, and the confidence interval does not contain 0.

When the two variances cannot be assumed to be the same, there are several alternative solutions in addition to use the Behrens–Fisher t' statistic. One simple way to solve this two means problem, called the smaller degrees of freedom t test, is to use the same t' statistic that has a t distribution with different degrees of freedom (e.g., Moore 2007, p. 465):

$$t' \sim t[\min(\nu_1, \nu_2)],$$

where the degrees of freedom is the smaller value of ν_1 or ν_2. Note that this method should be used only if no statistical software is available because it yields a conservative test result and a wider confidence interval. The example data yield $t' = 2.143$, $\nu = 4$, the two-tailed $p = 0.099$, and the 95% confidence interval of $[-0.427, 3.314]$. The null hypothesis of equal population means is not rejected at $\alpha = 0.05$, and the confidence interval contains 0.

Welch (1938)'s approximate t test also uses the same t' statistic that has a t distribution with the approximate degrees of freedom ν' (see Moore 2007, p. 474):

$$t' \sim t(\nu'),$$

where $\nu' = 1/[c^2/\nu_1 + (1 - c)^2/\nu_2]$ with $c = \left(s_1^2/n_1\right)/\left[\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)\right]$. It can be noted that the equivalent of this Welch's approximate t test was proposed by Smith (1936). Moore (2007) indicated that the approximation is accurate when both sample sizes are 5 or larger. Although there are other solutions, Welch's approximate t test currently seems to be the best practical solution to the Behrens–Fisher problem because of its availability from popular statistical software including SPSS and SAS. The example data yield $t' = 2.143$, $\nu' = 4.118$, the two-tailed $p = 0.097$, and the 95% confidence interval of $[-0.406, 3.293]$. The null hypothesis of equal population means is

not rejected at $\alpha = 0.05$, and the confidence interval contains 0.

In addition to the previous method, Welch (1938, 1947) and Aspin (1948) presented an approximation of the distribution of t' by the method of moments (i.e., Welch-Aspin t test; see Kim and Cohen 1998, for the detailed expansion terms for the approximation). The example data yield $t' = 2.143$ and the critical value under the Welch-Aspin t test for the two-tailed test is 2.715 at $\alpha = 0.05$. The corresponding 95% confidence interval is [−0.386, 3.273]. Again, the null hypothesis of equal population means is not rejected at $\alpha = 0.05$, and the confidence interval contains 0.

The Bayesian solution to the Behrens–Fisher problem was offered by Jeffreys (1940). When uninformative uniform priors are used for the population parameters, the Bayesian solution to the Behrens–Fisher problem is identical to that of Fisher's (1939). The Bayesian highest posterior density interval that contains the population mean difference with the probability of $1 - \alpha$ is identical to the $100(1 - \alpha)\%$ fiducial interval.

There are many solutions to the Behrens–Fisher problem based on the frequentist approach of the Neyman and Pearson (1928) sampling theory. Among the methods, Kim and Cohen (1998) indicated that Welch (1938, 1947), Aspin (1948), and Tsui and Weerahandi (1989) are the most important ones from the frequentist perspective. The critical values and the confidence intervals from various methods under the frequentist approach are in general different from either fiducial or Bayesian approach. For the one-sided alternative hypothesis, however, it is interesting to note that the generalized extreme region to obtain the generalized p by Tsui and Weerahandi (1989) is identical to the extreme area from the Behrens–Fisher t' statistic (see also Weerahandi 1995, pp. 174–181).

For the example data, the smaller degrees of freedom t test yielded the most conservative result with the largest critical value and the widest confidence interval. The Student's t test yielded the smallest critical value and the shortest confidence interval. All other intervals lie between these two intervals. Robinson (1982) pointed out that the differences between many solutions to the Behrens–Fisher problem might be less than their differences from the Student's t test when sample sizes are greater than 10.

The popular statistical software programs SPSS and SAS produce results from the Welch's approximate t test and the Student's t test as well as the respective confidence intervals. It is essential to have a table that contains the percentage points of the Behrens–Fisher distribution or computer programs that can calculate the tail areas and percentage values in order to use the Behrens–Fisher t test or to obtain the fiducial interval. Note that the Welch's approximate t test may not be as effective as the Welch-Aspin t test (Wang 1971). Note also that the sequential testing of the population means based on the result from either Levene's test of the equal population variances from SPSS or the folded F test from SAS is not recommended in general due to the complicated nature of control of the Type I error in the sequential testing.

About the Authors

Dr. Allan S. Cohen is Professor of Research, Evaluation, Measurement, and Statistics Program, Department of Educational Psychology and Instructional Technology, and Director, the Georgia Center for Assessment, both at the University of Georgia. He has authored or co-authored over 70 journal articles, 250 conference papers and technical reports as well as several tests, test manuals, and computer programs. He is an editorial board member of Applied Psychological Measurement and the Journal of Educational Measurement and is a Member of the American Educational Research Association and the National Council on Measurement in Education.

Dr. Seock-Ho Kim is Professor of Research, Evaluation, Measurement, and Statistics Program, Department of Educational Psychology and Instructional Technology, University of Georgia. He has co-authored over 40 papers and several computer programs. He has been the Editorial board member of *Educational and Psychological Measurement*, *Journal of Educational Measurement*, *School Psychology Quarterly*, and *Measurement in Physical Education and Exercise Science*. He has a co-authored book with Frank B. Baker entitled *Item Response Theory: Parameter Estimation Techniques* (2nd edn, Dekker, 2004). He is a Member of the American Statistical Association, Institute of Mathematical Statistics, American Educational Research Association, American Psychological Association, National Council on Measurement in Education, and Psychometric Society. Professor Seock-Ho Kim has served as Vice-President and President of the Korean-American Educational Researchers Association.

Cross References

►Fiducial Inference
►Heteroscedasticity
►Permutation Tests
►Statistical Software: An Overview
►Student's t-Tests

References and Further Reading

Aspin AA (1948) An examination and further development of a formula arising in the problem of comparing two mean values. Biometrika 35:88–96

Behrens WU (1929) Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen (A contribution to error estimation with few observations). Landwirtschaftliche Jahrbücher 68:807–837
Fisher RA (1935) The fiducial argument in statistical inference. Ann Eugenic 6:391–398
Fisher RA (1939) The comparison of samples with possibly unequal variances. Ann Eugenic 9:174–180
Fisher RA, Yates F (1957) Statistical tables for biological, agricultural and medical research, 4th edn. Oliver and Boyd, Edinburgh, England
Jeffreys H (1940) Note on the Behrens–Fisher formula. Ann Eugenic 10:48–51
Kim S-H, Cohen AS (1998) On the Behrens–Fisher problem: a review. J Educ Behav Stat 23:356–377
Lehmann EL (1975) Nonparametrics: statistical methods based on ranks. Holden-Day, San Francisco
Lindley DV, Scott WF (1995) New Cambridge elementary statistical tables, 2nd edn. Cambridge University Press, Cambridge, England
Moore DS (2007) The basic practice of statistics, 4th edn. W.H. Freeman, New York
Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20A(175–240):263–294
Robinson GK (1982) Behrens–Fisher problem. In: Kotz S, Johnson NL, Read CB (eds) Encyclopedia of statistical sciences, vol 1. Wiley, New York pp 205–208
Smith HF (1936) The problem of comparing the results of two experiments with unequal errors. J Coun Sci Ind Res 9:211–212
Tsui K-H, Weerahandi S (1989) Generalized *p*-values in significance testing of hypotheses in the presence of nuisance parameters. J Am Stat Assoc 84:602–607; Correction 86:256
Wang YY (1971) Probability of the type I error of the Welch tests for the Behrens-Fisher problem. J Am Stat Assoc 66:605–608
Weerahandi S (1995) Exact statistical methods for data analysis. Springer, New York
Welch BL (1938) The significance of the difference between two means when the population variances are unequal. Biometrika 29:350–362
Welch BL (1947) The generalization of 'Student's' problem when several different population variances are involved. Biometrika 34:28–35

Best Linear Unbiased Estimation in Linear Models

Simo Puntanen[1], George P. H. Styan[2]
[1]University of Tampere, Tampere, Finland
[2]McGill University, Montréal, QC, Canada

Introduction

In this article we consider the general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{or in short } \mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V}\},$$

where $\mathbf{X}$ is a known $n \times p$ model matrix, the vector $\mathbf{y}$ is an observable n-dimensional random vector, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{\varepsilon}$ is an unobservable vector of random errors with expectation $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, and covariance matrix $cov(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, where $\sigma^2 > 0$ is an unknown constant. The nonnegative definite (possibly singular) matrix $\mathbf{V}$ is known. In our considerations σ^2 has no role and hence we may put $\sigma^2 = 1$.

As regards the notation, we will use the symbols $\mathbf{A}'$, $\mathbf{A}^-$, $\mathbf{A}^+$, $\mathscr{C}(\mathbf{A})$, $\mathscr{C}(\mathbf{A})^\perp$, and $\mathscr{N}(\mathbf{A})$ to denote, respectively, the transpose, a generalized inverse, the Moore–Penrose inverse, the column space, the orthogonal complement of the column space, and the null space, of the matrix $\mathbf{A}$. By $(\mathbf{A} : \mathbf{B})$ we denote the partitioned matrix with $\mathbf{A}$ and $\mathbf{B}$ as submatrices. By $\mathbf{A}^\perp$ we denote any matrix satisfying $\mathscr{C}(\mathbf{A}^\perp) = \mathscr{N}(\mathbf{A}') = \mathscr{C}(\mathbf{A})^\perp$. Furthermore, we will write $\mathbf{P}_\mathbf{A} = \mathbf{A}\mathbf{A}^+ = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ to denote the orthogonal projector (with respect to the standard inner product) onto $\mathscr{C}(\mathbf{A})$. In particular, we denote $\mathbf{H} = \mathbf{P}_\mathbf{X}$ and $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$. One choice for $\mathbf{X}^\perp$ is of course the projector $\mathbf{M}$.

Let $\mathbf{K}'\boldsymbol{\beta}$ be a given vector of parametric functions specified by $\mathbf{K}' \in \mathbb{R}^{q\times p}$. Our object is to find a (homogeneous) linear estimator $\mathbf{A}\mathbf{y}$ which would provide an unbiased and in some sense "best" estimator for $\mathbf{K}'\boldsymbol{\beta}$ under the model $\mathscr{M}$. However, not all parametric functions have linear unbiased estimators; those which have are called estimable parametric functions, and then there exists a matrix $\mathbf{A}$ such that

$$\mathrm{E}(\mathbf{A}\mathbf{y}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta} = \mathbf{K}'\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p.$$

Hence $\mathbf{K}'\boldsymbol{\beta}$ is estimable if and only if there exists a matrix $\mathbf{A}$ such that $\mathbf{K}' = \mathbf{A}\mathbf{X}$, i.e., $\mathscr{C}(\mathbf{K}) \subset \mathscr{C}(\mathbf{X}')$.

The ordinary ▶least squares estimator of $\mathbf{K}'\boldsymbol{\beta}$ is defined as $\mathrm{OLSE}(\mathbf{K}'\boldsymbol{\beta}) = \mathbf{K}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is any solution to the normal equation $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$; hence $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ minimizes $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and it can be expressed as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$, while $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$. Now the condition $\mathscr{C}(\mathbf{K}) \subset \mathscr{C}(\mathbf{X}')$ guarantees that $\mathbf{K}'\hat{\boldsymbol{\beta}}$ is unique, even though $\hat{\boldsymbol{\beta}}$ may not be unique.

The Best Linear Unbiased Estimator (BLUE) of $\mathbf{X}\boldsymbol{\beta}$

The expectation $\mathbf{X}\boldsymbol{\beta}$ is trivially estimable and $\mathbf{G}\mathbf{y}$ is unbiased for $\mathbf{X}\boldsymbol{\beta}$ whenever $\mathbf{G}\mathbf{X} = \mathbf{X}$. An unbiased linear estimator $\mathbf{G}\mathbf{y}$ for $\mathbf{X}\boldsymbol{\beta}$ is defined to be the *best* linear unbiased estimator, BLUE, for $\mathbf{X}\boldsymbol{\beta}$ under $\mathscr{M}$ if

$$cov(\mathbf{G}\mathbf{y}) \leq_\mathrm{L} cov(\mathbf{L}\mathbf{y}) \quad \text{for all } \mathbf{L}: \mathbf{L}\mathbf{X} = \mathbf{X},$$

where "$\leq_\mathrm{L}$" refers to the Löwner partial ordering. In other words, $\mathbf{G}\mathbf{y}$ has the smallest covariance matrix (in the Löwner sense) among all linear unbiased estimators. We

denote the BLUE of $\mathbf{X}\boldsymbol{\beta}$ as $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\tilde{\boldsymbol{\beta}}$. If $\mathbf{X}$ has full column rank, then $\boldsymbol{\beta}$ is estimable and an unbiased estimator $\mathbf{Ay}$ is the BLUE for $\boldsymbol{\beta}$ if $\mathbf{AVA}' \leq_{\mathrm{L}} \mathbf{BVB}'$ for all $\mathbf{B}$ such that $\mathbf{BX} = \mathbf{I}_p$. The Löwner ordering is a very strong ordering implying for example

$$var(\tilde{\beta}_i) \leq var\left(\beta_i^*\right), \quad i = 1, \ldots, p,$$
$$tracecov(\tilde{\boldsymbol{\beta}}) \leq tracecov(\boldsymbol{\beta}^*), \quad \det cov(\tilde{\boldsymbol{\beta}}) \leq \det cov(\boldsymbol{\beta}^*),$$

for any linear unbiased estimator $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$; here "det" denotes determinant.

The following theorem gives the "Fundamental BLUE equation"; see, e.g., Rao (1967), Zyskind (1967) and Puntanen et al. (2000).

Theorem 1 *Consider the general linear model* $\mathscr{M} = \{\boldsymbol{y}, \boldsymbol{X\beta}, \boldsymbol{V}\}$. *Then the estimator* $\boldsymbol{Gy}$ *is the BLUE for* $\boldsymbol{X\beta}$ *if and only if* $\boldsymbol{G}$ *satisfies the equation*

$$\boldsymbol{G}(\boldsymbol{X} : \boldsymbol{VX}^{\perp}) = (\boldsymbol{X} : \boldsymbol{0}). \tag{1}$$

The corresponding condition for $\boldsymbol{Ay}$ *to be the BLUE of an estimable parametric function* $\boldsymbol{K}'\boldsymbol{\beta}$ *is* $\boldsymbol{A}(\boldsymbol{X} : \boldsymbol{VX}^{\perp}) = (\boldsymbol{K}' : \boldsymbol{0})$.

It is sometimes convenient to express (1) in the following form, see Rao (1971).

Theorem 2 (Pandora's Box) *Consider the general linear model* $\mathscr{M} = \{\boldsymbol{y}, \boldsymbol{X\beta}, \boldsymbol{V}\}$. *Then the estimator* $\boldsymbol{Gy}$ *is the BLUE for* $\boldsymbol{X\beta}$ *if and only if there exists a matrix* $\boldsymbol{L} \in \mathbb{R}^{p\times n}$ *so that* $\boldsymbol{G}$ *is a solution to*

$$\begin{pmatrix} \boldsymbol{V} & \boldsymbol{X} \\ \boldsymbol{X}' & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{G}' \\ \boldsymbol{L} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{X}' \end{pmatrix}.$$

The equation (1) has a unique solution for $\mathbf{G}$ if and only if $\mathscr{C}(\mathbf{X} : \mathbf{V}) = \mathbb{R}^n$. Notice that under $\mathscr{M}$ we assume that the observed value of $\mathbf{y}$ belongs to the subspace $\mathscr{C}(\mathbf{X} : \mathbf{V})$ with probability 1; this is the consistency condition of the linear model, see, e.g., Baksalary et al. (1992). The consistency condition means, for example, that whenever we have some statements which involve the random vector $\mathbf{y}$, these statements need hold only for those values of $\mathbf{y}$ that belong to $\mathscr{C}(\mathbf{X} : \mathbf{V})$. The general solution for $\mathbf{G}$ can be expressed, for example, in the following ways:

$$\mathbf{G}_1 = \mathbf{X}(\mathbf{X}'\mathbf{W}^{-}\mathbf{X})^{-}\mathbf{X}'\mathbf{W}^{-} + \mathbf{F}_1(\mathbf{I}_n - \mathbf{WW}^{-}),$$

$$\mathbf{G}_2 = \mathbf{H} - \mathbf{HVM}(\mathbf{MVM})^{-}\mathbf{M} + \mathbf{F}_2[\mathbf{I}_n - \mathbf{MVM}(\mathbf{MVM})^{-}]\mathbf{M},$$

where $\mathbf{F}_1$ and $\mathbf{F}_2$ are arbitrary matrices, $\mathbf{W} = \mathbf{V} + \mathbf{XUX}'$ and $\mathbf{U}$ is any arbitrary conformable matrix such that $\mathscr{C}(\mathbf{W}) = \mathscr{C}(\mathbf{X} : \mathbf{V})$. Notice that even though $\mathbf{G}$ may not be unique, the numerical value of $\mathbf{Gy}$ is unique because $\mathbf{y} \in \mathscr{C}(\mathbf{X} : \mathbf{V})$. If $\mathbf{V}$ is positive definite, then $\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. Clearly $\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{Hy}$ is the BLUE under $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$. It is also worth noting that the matrix $\mathbf{G}$ satisfying (1) can be interpreted as a projector: it is a projector onto $\mathscr{C}(\mathbf{X})$ along $\mathscr{C}(\mathbf{VX}^{\perp})$, see Rao (1974).

OLSE vs. BLUE

Characterizing the equality of the Ordinary Least Squares Estimator (OLSE) and the BLUE has received a lot of attention in the literature, but the major breakthroughs were made by Rao (1967) and Zyskind (1967); for a detailed review, see Puntanen and Styan (1989). We present below six characterizations for the OLSE and the BLUE to be equal (with probability 1).

Theorem 3 (OLSE vs. BLUE) *Consider the general linear model* $\mathscr{M} = \{\boldsymbol{y}, \boldsymbol{X\beta}, \boldsymbol{V}\}$. *Then* $\mathrm{OLSE}(\boldsymbol{X\beta}) = \mathrm{BLUE}(\boldsymbol{X\beta})$ *if and only if any one of the following six equivalent conditions holds.* (*Note:* $\boldsymbol{V}$ *may be replaced by its Moore–Penrose inverse* $\boldsymbol{V}^{+}$ *and* $\boldsymbol{H}$ *and* $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{H}$ *may be interchanged.*)

1. $\boldsymbol{HV} = \boldsymbol{VH}$,
2. $\boldsymbol{HVM} = \boldsymbol{0}$,
3. $\mathscr{C}(\boldsymbol{VH}) \subset \mathscr{C}(\boldsymbol{H})$,
4. $\mathscr{C}(\boldsymbol{X})$ *has a basis comprising orthonormal eigenvectors of* $\boldsymbol{V}$,
5. $\boldsymbol{V} = \boldsymbol{HAH} + \boldsymbol{MBM}$ *for some* $\boldsymbol{A}$ *and* $\boldsymbol{B}$,
6. $\boldsymbol{V} = \alpha\boldsymbol{I}_n + \boldsymbol{HKH} + \boldsymbol{MLM}$ *for some* $\alpha \in \mathbb{R}$, *and* $\boldsymbol{K}$ *and* $\boldsymbol{L}$.

Two Linear Models

Consider now two linear models $\mathscr{M}_1 = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_1\}$ and $\mathscr{M}_2 = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}_2\}$, which differ only in their covariance matrices. For the proof of the following proposition and related discussion, see, e.g., Rao (1971, Theorem 5.2, Theorem 5.5), and Mitra and Moore (1973, Theorem 3.3, Theorems 4.1–4.2).

Theorem 4 *Consider the linear models* $\mathscr{M}_1 = \{\boldsymbol{y}, \boldsymbol{X\beta}, \boldsymbol{V}_1\}$ *and* $\mathscr{M}_2 = \{\boldsymbol{y}, \boldsymbol{X\beta}, \boldsymbol{V}_2\}$, *and let the notation* $\{\, \mathrm{BLUE}(\boldsymbol{X\beta} \mid \mathscr{M}_1) \,\} \subset \{\, \mathrm{BLUE}(\boldsymbol{X\beta} \mid \mathscr{M}_2) \,\}$ *mean that every representation of the BLUE for* $\boldsymbol{X\beta}$ *under* $\mathscr{M}_1$ *remains the BLUE for* $\boldsymbol{X\beta}$ *under* $\mathscr{M}_2$. *Then the following statements are equivalent:*

1. $\{\, \mathrm{BLUE}(\boldsymbol{X\beta} \mid \mathscr{M}_1) \,\} \subset \{\, \mathrm{BLUE}(\boldsymbol{X\beta} \mid \mathscr{M}_2) \,\}$,
2. $\mathscr{C}(\boldsymbol{V}_2\boldsymbol{X}^{\perp}) \subset \mathscr{C}(\boldsymbol{V}_1\boldsymbol{X}^{\perp})$,
3. $\boldsymbol{V}_2 = \boldsymbol{V}_1 + \boldsymbol{XN}_1\boldsymbol{X}' + \boldsymbol{V}_1\boldsymbol{MN}_2\boldsymbol{MV}_1$, *for some* $\boldsymbol{N}_1$ *and* $\boldsymbol{N}_2$,
4. $\boldsymbol{V}_2 = \boldsymbol{XN}_3\boldsymbol{X}' + \boldsymbol{V}_1\boldsymbol{MN}_4\boldsymbol{MV}_1$, *for some* $\boldsymbol{N}_3$ *and* $\boldsymbol{N}_4$.

Notice that obviously

$$\{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_1)\,\} = \{\,\mathrm{BLUE}(\mathbf{X}\boldsymbol{\beta} \mid \mathscr{M}_2)\,\} \iff$$
$$\mathscr{C}(\mathbf{V}_2\mathbf{X}^{\perp}) = \mathscr{C}(\mathbf{V}_1\mathbf{X}^{\perp}).$$

For the equality between the BLUEs of $\mathbf{X}_1\boldsymbol{\beta}_1$ under two partitioned models, see Haslett and Puntanen (2010a).

Model with New Observations: Best Linear Unbiased Predictor (BLUP)

Consider the model $\mathscr{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, and let $\mathbf{y}_f$ denote an $m \times 1$ unobservable random vector containing *new observations*. The new observations are assumed to follow the linear model $\mathbf{y}_f = \mathbf{X}_f\boldsymbol{\beta} + \boldsymbol{\varepsilon}_f$, where $\mathbf{X}_f$ is a known $m \times p$ model matrix associated with new observations, $\boldsymbol{\beta}$ is the same vector of unknown parameters as in $\mathscr{M}$, and $\boldsymbol{\varepsilon}_f$ is an $m \times 1$ random error vector associated with new observations. Our goal is to predict the random vector $\mathbf{y}_f$ on the basis of $\mathbf{y}$. The expectation and the covariance matrix are

$$\mathrm{E}\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_f\end{pmatrix} = \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{X}_f\boldsymbol{\beta}\end{pmatrix}, \quad cov\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_f\end{pmatrix} = \begin{pmatrix}\mathbf{V} = \mathbf{V}_{11} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix},$$

which we may write as

$$\mathscr{M}_f = \left\{\begin{pmatrix}\mathbf{y}\\ \mathbf{y}_f\end{pmatrix}, \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{X}_f\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\mathbf{V} & \mathbf{V}_{12}\\ \mathbf{V}_{21} & \mathbf{V}_{22}\end{pmatrix}\right\}.$$

A linear predictor $\mathbf{Ay}$ is said to be unbiased for $\mathbf{y}_f$ if $\mathrm{E}(\mathbf{Ay}) = \mathrm{E}(\mathbf{y}_f) = \mathbf{X}_f\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Then the random vector $\mathbf{y}_f$ is said to be unbiasedly predictable. Now an unbiased linear predictor $\mathbf{Ay}$ is the best linear unbiased predictor, BLUP, if the Löwner ordering

$$cov(\mathbf{Ay} - \mathbf{y}_f) \le_{\mathrm{L}} cov(\mathbf{By} - \mathbf{y}_f)$$

holds for all $\mathbf{B}$ such that $\mathbf{By}$ is an unbiased linear predictor for $\mathbf{y}_f$.

The following theorem characterizes the BLUP; see, e.g., Christensen (2002, p 283), and Isotalo and Puntanen (2006, p 1015).

Theorem 5 (Fundamental BLUP equation) *Consider the linear model $\mathscr{M}_f$, where $\boldsymbol{X}_f\boldsymbol{\beta}$ is a given estimable parametric function. Then the linear estimator $\boldsymbol{Ay}$ is the best linear unbiased predictor (BLUP) for $\boldsymbol{y}_f$ if and only if $\boldsymbol{A}$ satisfies the equation*

$$\boldsymbol{A}(\boldsymbol{X} : \boldsymbol{V}\boldsymbol{X}^{\perp}) = (\boldsymbol{X}_f : \boldsymbol{V}_{21}\boldsymbol{X}^{\perp}).$$

In terms of Pandora's Box (Theorem 2), $\boldsymbol{Ay}$ is the BLUP for $\boldsymbol{y}_f$ if and only if there exists a matrix $\boldsymbol{L}$ such that $\boldsymbol{A}$ satisfies the equation

$$\begin{pmatrix}\boldsymbol{V} & \boldsymbol{X}\\ \boldsymbol{X}' & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{A}'\\ \boldsymbol{L}\end{pmatrix} = \begin{pmatrix}\boldsymbol{V}_{12}\\ \boldsymbol{X}_f'\end{pmatrix}.$$

The Mixed Model

A mixed linear model can be presented as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \text{or in short} \quad \mathscr{M}_{\mathrm{mix}} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \mathbf{D}, \mathbf{R}\},$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Z} \in \mathbb{R}^{n\times q}$ are known matrices, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown fixed effects, $\boldsymbol{\gamma}$ is an unobservable vector (q elements) of *random effects* with $cov(\boldsymbol{\gamma}, \boldsymbol{\varepsilon}) = \mathbf{0}_{q\times p}$ and

$$\mathrm{E}(\boldsymbol{\gamma}) = \mathbf{0}_q, \quad cov(\boldsymbol{\gamma}) = \mathbf{D}_{q\times q}, \quad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_n, \quad cov(\boldsymbol{\varepsilon}) = \mathbf{R}_{n\times n}.$$

This leads directly to

Theorem 6 *Consider the mixed model $\mathscr{M}_{mix} = \{\boldsymbol{y}, \boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \boldsymbol{D}, \boldsymbol{R}\}$. Then the linear estimator $\boldsymbol{By}$ is the BLUE for $\boldsymbol{X\beta}$ if and only if*

$$\boldsymbol{B}(\boldsymbol{X} : \boldsymbol{\Sigma}\boldsymbol{X}^{\perp}) = (\boldsymbol{X} : \boldsymbol{0}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{ZDZ}' + \boldsymbol{R}$. Moreover, $\boldsymbol{Ay}$ is the BLUP for $\boldsymbol{\gamma}$ if and only if

$$\boldsymbol{A}(\boldsymbol{X} : \boldsymbol{\Sigma}\boldsymbol{X}^{\perp}) = (\boldsymbol{0} : \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{X}^{\perp}).$$

In terms of Pandora's Box (Theorem 2), $\boldsymbol{Ay} = \mathrm{BLUP}(\boldsymbol{\gamma})$ if and only if there exists a matrix $\boldsymbol{L}$ such that $\boldsymbol{A}$ satisfies the equation

$$\begin{pmatrix}\boldsymbol{\Sigma} & \boldsymbol{X}\\ \boldsymbol{X}' & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{A}'\\ \boldsymbol{L}\end{pmatrix} = \begin{pmatrix}\boldsymbol{ZD}\\ \boldsymbol{0}\end{pmatrix}.$$

For the equality between the BLUPs under two mixed models, see Haslett and Puntanen (2010b, 2010c).

Acknowledgment

The research of the second author was supported in part by the Natural Sciences and Engineering Research Council of Canada.

About the Authors

Simo Puntanen is Adjunct Professor in the Department of Mathematics and Statistics, University of Tampere, Tampere, Finland. Currently he is the Book Review Editor of the *International Statistical Review*, and a member of the Editorial Board of the *Annals of the Institute of Statistical Mathematics*, *Communications in Statistics*, *Statistical Papers*, and *Mathematical Inequalities & Applications*. He is a founding member (with George P. H. Styan)

of the International Workshop on Matrices and Statistics, held regularly from 1990 onwards. He is a co-author (with George P. H. Styan and Jarkko Isotalo) of the book *Matrix Tricks for Linear Statistical Models: Our Personal Top Twenty* (Springer, 2010).

George P. H. Styan is Professor Emeritus of Mathematics and Statistics at McGill University in Montréal (Québec), Canada, and is currently the Abstracting Editor of *Current Index to Statistics*. In 2009, Professor Styan was named an Honorary Member of the Statistical Society of Canada "for his deep research at the interface of Matrix Theory and Statistics; for his remarkable editorial work within Canada and beyond, his mentoring of graduate and postdoctoral students; and for innumerable other scholarly and professional contributions to the international statistical community."

Cross References

►Autocorrelation in Regression
►Gauss-Markov Theorem
►General Linear Models
►Least Squares
►Linear Regression Models
►Ordered Statistical Data: Recent Developments
►Small Area Estimation
►Statistical Design of Experiments (DOE)
►Unbiased Estimators and Their Applications

References and Further Reading

Baksalary JK, Rao CR, Markiewicz A (1992) A study of the influence of the 'natural restrictions' on estimation problems in the singular Gauss–Markov model, J Stat Plann Infer 31:335–351

Christensen R (2002) Plane answers to complex questions: the theory of linear models, 3rd edn. Springer, New York

Haslett SJ, Puntanen S (2010a) Effect of adding regressors on the equality of the BLUEs under two linear models. J Stat Plann Infer 140:104–110

Haslett SJ, Puntanen S (2010b) Equality of BLUEs or BLUPs under two linear models using stochastic restrictions. Statistical Papers 51:465–475

Haslett SJ, Puntanen S (2010c) On the equality of the BLUPs under two linear mixed models. Metrika, DOI 10.1007/S00184–010–0308–6

Isotalo J, Puntanen S (2006) Linear prediction sufficiency for new observations in the general Gauss–Markov model. Commun Stat-Theor 35: 1011–1023

Mitra SK, Moore BJ (1973) Gauss–Markov estimation with an incorrect dispersion matrix. Sankhyā Series *A* 35:139–152

Puntanen S, Styan GPH (1989) The equality of the ordinary least squares estimator and the best linear unbiased estimator (with comments by Oscar Kempthorne and by Shayle R. Searle and with "Reply" by the authors). Am Stat 43:153–164

Puntanen S, Styan GPH, Werner HJ (2000) Two matrix-based proofs that the linear estimator *Gy* is the best linear unbiased estimator. J Stat Plann Infer 88:173–179

Rao CR (1967) Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In: Le Cam LM, Neyman J (eds) Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Berkeley, California, 1965/1966, vol 1. University of California Press, Berkeley, pp 355–372

Rao CR (1971) Unified theory of linear estimation. Sankhyā, Series A 33:371–394 (Corrigenda (1972), 34:194 and 477)

Rao CR (1974) Projectors, generalized inverses and the BLUE's. J R Stat Soc Series B 36:442–448

Haslett SJ, Puntanen S (2010b) Equality of BLUEs or BLUPs under two linear models using stochastic restrictions. Statistical Papers 51: 564–475

Haslett SJ, Puntanen S (2010c) On the equality of the BLUPs under two linear mixed models. Metrika, aailable online, DOI 10.1007/s00184-010-0308-8.

Zyskind G (1967) On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. Ann Math Stat 38:1092–1109

Beta Distribution

Arjun K. Gupta
Distinguished University Professor
Bowling Green State University, Bowling Green, OH, USA

A random variable X is said to have the beta distribution with parameters a and b if its probability density function is

$$f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1, \quad a > 0, b > 0 \tag{1}$$

where

$$B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1} du$$

denotes the beta function. The beta family, whose origin can be traced to 1676. in a letter from Sir Isaac Newton to Henry Oldenberg, has been utilized extensively in statistical theory and practice.

Originally defined on the unit interval, many generalizations of (1) have been proposed in the literature; see Karian and Dudewicz (2000) for a four parameter generalization defined over a finite interval, McDonald and Richards (1987a, b) for a generalization obtained by power transformation of X; Libby and Novick (1982) and Armero and Bayarri (1994) for generalizations obtained by dividing (1) by certain algebraic functions; Gordy (1998) for a generalization obtained by multiplying (1) by an exponential function; and, Nadarajah and Kotz (2003) for a generalization obtained by multiplying (1) by a Gauss hypergeometric function. For further details, the reader is referred to

Chap. 25 in Johnson et al. (1994) and Gupta and Nadarajah (2004).

Some properties of beta distribution (1) are listed here.

1. rth moment about zero

$$\mu'_r = \frac{B(a+r,b)}{B(a,b)} = \frac{\Gamma(a+r)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+r)}.$$

In particular,

$$E(X) = \frac{a}{a+b},$$

$$Var(X) = \frac{ab}{(a+b+1)(a+b)^2}.$$

2. Characteristic function

$$E\left[e^{itX}\right] = {}_1F_1(a; a+b; it)$$

where ${}_1F_1$ is the confluent hypergeometric function defined by

$${}_1F_1(\alpha;\beta;z) = \sum_{k=0}^{\infty} \frac{(\alpha)_k}{(\beta)_k}\frac{z^k}{k!}.$$

3. The random variable

$$Y = \frac{X}{1-X}$$

has the Pearson type VI distribution defined by

$$f_Y(y) = \frac{1}{B(a,b)}\frac{y^{a-1}}{(1+y)^{a+b}},\ y > 0.$$

About the Author

Professor Gupta is Distinguished University Professor, Department of Mathematics and Statistics Bowling Green State University. He was Chair of the Department (1985–1987). He is Elected Fellow of the Ohio Academy of Science (2001), ASA, ISI and Royal Statistical Society. He was awarded the All India Mathematical Society Golden Jubilee Award (1959). Professor Gupta was editor and associate editor of many international journals. Currently, he is Associate editor for *Communications in Statistics*. He has written or co-authored more than 450 papers and 16 books, including *Handbook of Beta Distribution and Its Applications* (with Saralees Nadarajah, eds., CRC, 2004).

Cross References

►Approximations to Distributions
►Bayesian Statistics
►F Distribution
►Location-Scale Distributions
►Multivariate Statistical Distributions
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Uniform Distribution in Statistics

References and Further Reading

Armero C, Bayarri MJ (1994) Prior assessments for prediction in queues. The Statistician 43:139–153

Gordy MB (1988) Computationally convenient distributional assumptions for commonvalue auctions. Comput Econ 12: 61–78

Gupta AK, Nadarajah S (2004) Handbook of beta distribution and its applications. Marcel Dekker, New York

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol 2, 2nd edn. Wiley, New York

Karian ZA, Dudewicz EJ (2000) Fitting statistical distributions: the generalized lambda distribution and generalized bootstrap methods. CRC, Boca Raton, Florida

Libby DL, Novick MR (1982) Multivariate generalized beta-distributions with applications to utility assessment. J Educ Stat 7:271–294

McDonald JB, Richards DO (1987a) Model selection: some generalized distributions. Commun Stat 16A:1049–1074

McDonald JB, Richards DO (1987b) Some generalized models with application to reliability. J Stat Plann Inf 16:365–376

Nadarajah S, Kotz S (2004) A generalized beta distribution II. InterStat

Bias Analysis

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

Methodological shortcomings of studies can lead to bias, in the sense of systematic (nonrandom) distortion of estimates from the studies. Well-recognized shortcomings (bias sources) include the following: Units may be selected for observation in a nonrandom fashion; stratifying on additional unmeasured covariates U may be essential for the X-Y association to approximate a target causal effect; inappropriate covariates may be entered into the analysis; and components of X or Y or Z may not be adequately measured.

In methodologic modeling or *bias analysis*, one models these shortcomings. In effect, one attempts to model the design and execution of the study, including features (such as selection biases and measurement errors) beyond investigator control. The process is thus a natural extension to imperfect experiments and observational studies of the design-based paradigm in experimental and survey statistics.

Bias analysis well established in engineering and policy research and are covered in many books, albeit in a wide variety of forms and specialized applications. Little and Rubin (2002) focus on missing-data problems; Eddy et al. (1992) focus on medical and health-risk assessment; and Vose (2000) covers general risk assessment. Models for specific biases have a long if scattered history in epidemiology, going back to Berkson (1946) and Cornfield et al. (1959); Greenland and Lash (2008) give a review. Nonetheless, methods for statistical inference from bias models have only recently begun to appear in observational health research (Robins et al. 2000; Graham 2000; Lash and Fink 2003; Lash et al. 2009; Phillips 2003; Greenland 2003a, 2005, 2009a, b; Gustafson 2003, 2005a, b; Fox et al. 2005).

Statistical Formulation

Many of the parameters in realistic bias models will not be *identifiable* (estimable) from the data, necessitating inferential approaches well beyond those of conventional statistics. The simplest approach is to fix those parameters at specific values, estimate effects assuming those values are correct, and see how effect estimates change as those values are varied. This process is an example of ►*sensitivity analysis*. One can also assign the parameters prior probability distributions based on background information, and summarize target estimates over these distributions or over the resulting posterior distribution.

Consider the problem of estimating the effect of X on Y, given a collection of antecedent covariates Z, as common in causal inference (see Causation and Causal Inference). Standard approaches estimate the regression of Y on X and Z, E(Y|x,z), and then taking the fitted (partial) regression of Y on X given Z as the effect of X on Y. Usually a parametric model r(x,z;β) for E(Y|x,z) is fit and the coefficient for X is taken as the effect (this approach is reflected in common terminology that refers to such coefficients as "main effects"); the logistic model for a binary Y is the most common epidemiologic example. Model fitting is almost always done as if

1. Within levels of X and Z, the data are a ►simple random sample and any missingness is completely random.
2. The causal effect of X on Y is accurately reflected by the association of X and Y given Z (i.e., there is no residual confounding – as might be reasonable to assume if X were randomized within levels of Z).
3. X, Y, and Z are measured without error.

In reality, (1) sampling and missing-data probabilities may jointly depend on X, Y, and Z in an unknown fashion, (2) stratifying or adjusting for certain unmeasured (and possibly unknown) covariates U might be essential for the association of X and Y to correspond to a causal effect of X on Y, and (3) some of the X, Y and Z components are mismeasured.

Selection Biases

Let V = (X,Y,Z). One approach to sampling (selection) biases posits a model s(v;σ) for the probability of selection given v, then uses this model in the analysis along with r(x,z;β), e.g., by incorporating s(v;σ) into the likelihood function (Eddy et al. 1992; Little and Rubin 2002; Gelman et al. 2003; Greenland 2009b) or by using $s(v;\sigma)^{-1}$ as a weighting factor (Robins et al. 1994, 2000; Copas and Li 1997; Scharfstein et al. 1999). The parameters β and σ are usually cannot be completely estimated from the data under analysis, so one must either posit various fixed values for σ and estimate β for each chosen σ (sensitivity analysis), or else give β σ a prior distribution and conduct a Bayesian analysis.

A third, somewhat informal approach between sensitivity and Bayesian analysis is Monte-Carlo risk analysis or Monte-Carlo sensitivity analysis (MCSA). This approach repeatedly samples σ from its prior distribution, resamples (bootstraps) the data, and re-estimates β using the sampled σ and data; it then outputs the distribution of results obtained from this repeated sampling-estimation cycle. MCSA can closely approximate Bayesian results under certain (though not all) conditions (Greenland 2001, 2005), most notably when β and σ are *a priori* independent and there is negligible prior information about β. These selection-modeling methods can be generalized (with many technical considerations) to handle arbitrary missing data (Little and Rubin 2002; Robins et al. 1994, 2000).

Confounding

Suppose U is a collection of unmeasured (latent) covariates required for identification of the effect of X on Y. One approach to problem (2) is to model the distribution of U and V with a probability model p(u,v;β,γ) = p(y|u,x,z;β)p(u,x,z;γ). Again, one can estimate β by likelihood-based or by weighting methods, but because U is unmeasured, the parameter (β,γ) will not be fully estimable from the data and so some sort of sensitivity analysis or prior distribution will be needed (e.g., Cornfield et al. 1959; Yanagawa 1984; Robins et al. 2000; Rosenbaum 2002; Greenland 2003a, 2003b, 2005, 2009b). Results will depend heavily on the prior specification given U. For example, U may be a specific unmeasured covariate (e.g., smoking status) with well studied relations to X, Y, and Z,

which affords straightforward Bayesian and MCSA analyzes (Steenland and Greenland 2004). On the other hand, U may represent an unspecified aggregation of latent confounders, in which case the priors and hence inferences are more uncertain (Greenland 2003a).

Measurement Error and Misclassification

Suppose that the collection of "true" values V = (X,Y,Z) has a corresponding collection of measurements or surrogates W (which might include multiple surrogates for X, Y, or Z). The measurement-error problem (problem 3) can then be expressed as follows: For some or all units, at least one of the V components is missing, but the measurements in W that correspond to the missing V components are present. If enough units are observed with both V and W complete, the problem can be handled by standard missing-data methods. For example, given a model for the distribution of V and W one can use likelihood-based methods (Little and Rubin 2002), or impute V components where absent and then fit the model $r(x,z;\beta)$ for $E(Y|x,z)$ to the completed data (Cole et al. 2006), or fit the model to the complete records using weights derived from all records using a model for missing-data patterns (Robins et al. 1994). Direct Bayesian approaches to measurement error are also available (Gustafson 2003; Greenland 2009a, b).

Alternatively, there are many measurement-error correction procedures that adjust the "naïve" β estimates obtained by fitting the regression using W as if it were V. This adjustment is usually accomplished with a model relating V to W fitted to the complete records, as in instrumental-variable (regression calibration) corrections and their extensions (Carroll et al. 2006). Some recent methods are based on assuming various subsamples with information on multiple surrogates are available (so W may be of much higher dimension than V and may have complex missing-data patterns) (e.g., Spiegelman et al. 2005).

All methods assume that missingness in V and W components is random, which is often quite implausible because noncooperation increases with demands on subjects, collection of some components may be demanding (e.g., as when W includes diet diaries or biomarkers), and cooperation may be related to unobserved true values or confounders. Thus selection modeling will be needed along with measurement modeling to account for this nonrandom ("nonignorable") missingness.

Further nonidentified modeling becomes a necessity if a component of V is never observed on any unit (or, more practically, if there are too few complete records to support large-sample missing-data or measurement-error procedures). Latent-variable methods are natural for this situation (Berkane 1997). For example, one could model the distribution of (V,W) or a sufficient factor from that distribution by a parametric model; the unobserved components of V are then the latent variables in the model. As before, the parameters will not be fully identified, making Bayesian methods a natural choice for summary inferences (Gustafson 2003, 2005a, b; Greenland 2005, 2009a, b).

Realistic specification for nonidentified measurement error models can become quite complex, with inferences displaying extreme sensitivity to parameter constraints or prior distributions. Nonetheless, methodologic modeling helps provide an honest accounting for the large uncertainty that can be generated by even modest measurement error.

Acknowledgments

Some of the above material is adapted from Greenland (2004).

About the Author

For biography *see* the entry ►Confounding and Confounder Control.

Cross References

►Bias Correction
►Confounding and Confounder Control
►Identifiability
►Measurement Error Models
►Sensitivity Analysis

References and Further Reading

Berkane M (ed) (1997) Latent variable modeling and applications to causality. Lecture Notes in Statistics (120), Springer, New York

Berkson J (1946) Limitations of the application of fourfold table analysis to hospital data. Biometrics Bull 2:47–53

Brumback BA, Hernán MA, Haneuse S, Robins JM (2004) Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. Stat Med 23: 749–767

Carroll RJ, Ruppert D, Stefanski LA, Crainicenau CM (2006) Measurement error in nonlinear models: a modern perspective, 2nd edn. Chapman and Hall/CRC Press, Boca Raton, FL

Cole SR, Chu H, Greenland S (2006) A simulation study of multiple-imputation for measurement error correction. Int J Epidemiol 35:1074–1081

Copas JB, Li HG (1997) Inference for non-random samples. J Roy Stat Soc B 59:55–96

Cornfield J, Haenszel W, Hammond WC, Lilienfeld AM, Shimkin MB, Wynder EL (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst 22:173–203

Eddy DM, Hasselblad V, Schachter R (1992) Meta-analysis by the confidence profile method. Academic, New York
Fox MP, Lash TL, Greenland S (2005) A method to automate probabilistic sensitivity analyses of misclassified binary variables. Int J Epidemiol 34:1370–1377
Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis, 2nd edn. Chapman and Hall/CRC, New York
Goetghebeur E, van Houwelingen HC (eds) (1998) Analyzing noncompliance in clinical trials. Stat Med 17:247–389
Graham P (2000) Bayesian inference for a generalized population attributable fraction. Stat Med 19:937–956
Greenland S (2001) Sensitivity analysis, Monte-Carlo risk analysis, and Bayesian uncertainty assessment. Risk Anal 21:579–583
Greenland S (2003a) The impact of prior distributions for uncontrolled confounding and response bias. J Am Stat Assoc 98:47–54
Greenland S (2003b) Quantifying biases in causal models: classical confounding versus collider-stratification bias. Epidemiology 14:300–306
Greenland S (2004) An overview of methods for causal inference from observational studies. Ch. 1. In: Gelman A, Meng XL (eds) Applied Bayesian modeling and causal inference from an incomplete-data perspective. New York, Wiley, pp 3–13
Greenland S (2005) Multiple-bias modeling for observational studies (with discussion). J Roy Stat Soc A 168:267–308
Greenland S (2009a) Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods. Int J Epidemiol 38(6):1662–1673
Greenland S (2009b) Relaxation penalties and priors for plausible modeling of nonidentified bias sources. Stat Sci 24(2): 195–210
Greenland S, Lash TL (2008) Bias analysis. Ch. 19. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Philadelphia, Lippincott, pp 345–380
Gustafson P (2003) Measurement error and misclassification in statistics and epidemiology. Chapman and Hall, New York
Gustafson P (2005a) On model expansion, model contraction, identifiability, and prior information: two illustrative scenarios involving mismeasured variables (with discussion). Stat Sci 20:111–140
Gustafson P (2005b) The utility of prior information and stratification for parameter estimation with two screening tests but no gold standard. Stat Med 24:1203–1217
Gustafson P, Le ND, Saskin R (2001) Case-control analysis with partial knowledge of exposure misclassification probabilities. Biometrics 57:598–609
Joseph L, Gyorkos T, Coupal L (1995) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 141:263–272
Lash TL, Fink AK (2003) Semi-automated sensitivity analysis to assess systematic errors in observational epidemiologic data. Epidemiology 14:451–458
Lash TL, Fox MP, Fink AK (2009) Applying quantitative bias analysis to epidemiologic data. Springer, New York
Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
Phillips CV (2003) Quantifying and reporting uncertainty from systematic errors. Epidemiology 14:459–466
Robins JM (1999) Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry DA (eds) Statistical models in epidemiology. Springer, New York, pp 95–134
Robins JM, Rotnitzky A, Zhao LP (1994) Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 89:846–866
Robins JM, Rotnitzky A, Scharfstein DO (2000) Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry DA (eds) Statistical models in epidemiology. Springer, New York, pp 1–94
Rosenbaum P (2002) Observational studies, 2nd edn. Springer, New York
Scharfstein DO, Rotnitsky A, Robins JM (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Stat Assoc 94:1096–1120
Spiegelman D, Zhao B, Kim J (2005) Correlated errors in biased surrogates: study designs and methods for measurement error correction. Stat Med 24:1657–1682
Steenland K, Greenland S (2004) Monte-Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. Am J Epidemiol 160: 384–392
Vose D (2000) Risk analysis. Wiley, New York
Yanagawa T (1984) Case-control studies: assessing the effect of a confounding factor. Biometrika 71:191–194

Bias Correction

Gauss M. Cordeiro
Professor
Universidade Federal Rural de Pernambuco, Recife, Brazil

Introduction

A central object in asymptotic likelihood theory is the calculation of the second-order biases of the maximum likelihood estimates (MLEs). To improve the accuracy of these estimates, substantial effort has gone into computing the cumulants of log-likelihood derivatives which are, however, notoriously cumbersome. The MLEs typically have biases of order $O(n^{-1})$ for large sample size n, which are commonly ignored in practice, the justification being that they are small when compared to the standard errors of the parameter estimates that are of order $O(n^{-1/2})$. For small samples sizes, however, these biases can be appreciable and of the same magnitude as the corresponding standard errors. In such cases, the biases cannot be neglected, and for turning feasible estimation of their size in practical applications, corresponding formulae for their calculation need to be established for a wide range of probability distributions and regression models.

Bias correction has been extensively studied in the statistical literature and there has been considerable interest in finding simple matrix expressions for second-order biases of MLEs in some classes of regression models which do not involve cumulants of log-likelihood derivatives. The methodology has been applied to several regression models. We focus on the following models: normal nonlinear models (Cook et al. 1986), generalized log-gamma regression model (Young and Bakir 1987), ▶generalized linear models (Cordeiro and McCullagh 1991), ARMA models (Cordeiro and Klein 1994), multivariate nonlinear regression models (Cordeiro and Vasconcellos 1997), generalized linear models with dispersion covariates (Botter and Cordeiro 1998), symmetric nonlinear regression models (Cordeiro et al. 2000), Student t regression model with unknown degrees of freedom (Vasconcellos and Silva 2005), beta regression models (Ospina et al. 2006) and a class of multivariate normal model where the mean vector and the covariance matrix have parameters in common (Patriota and Lemonte 2009). In general two parameter continuous distributions, Stósic and Cordeiro (2009) showed how to symbolically compute the biases of the MLEs bypassing the traditional computation of joint cumulants of log-likelihood derivatives.

The bias approximation may be used to produce a bias-corrected estimator by subtracting the bias approximation from the MLE. Alternatively, an examination of the form of the bias may suggest a re-parametrization of the model that results is less biased estimates.

General Formula

Consider that the total log-likelihood function $\ell(\theta)$, based on n observations not necessarily independent and identically distributed, is a function of a $p\times 1$ vector θ of unknown parameters. We assume that $\ell = \ell(\theta)$ is regular (Cox and Hinkley 1974) with respect to all θ derivatives up to and including those of third order. We introduce the following log-likelihood derivatives in which we reserve lower-case subscripts $r, s, t, \ldots$ to denote components of the vector θ: $U_r = \partial l/\partial\theta_r$, $U_{rs} = \partial^2 l/\partial\theta_r\partial\theta_s$, and so on. The standard notation is adopted for the cumulants of log-likelihood derivatives $\kappa_{rs} = E(U_{rs})$, $\kappa_{r,s} = E(U_r U_s)$, $\kappa_{rs,t} = E(U_{rs}U_t)$, etc, where all κ's refer to a total over the sample and are, in general, of order n. The elements of the information matrix K^{-1} are $\kappa_{r,s} = -\kappa_{rs}$ and let $\kappa^{r,s} = -\kappa^{rs}$ denote the corresponding elements of the inverse information matrix K^{-1}, which is of order $O(n^{-1})$.

The MLE $\widehat{\theta}$ of θ can be obtained as a solution of a system of nonlinear equations $\widehat{U}_r = 0$ for $r = 1, \ldots, p$. A general formula for the $O(n^{-1})$ bias of $\widehat{\theta}$ for a regular statistical model with p unknown parameters was given by Cox and Snell (1968) and Cordeiro and McCullagh (1991). Assuming standard regularity conditions (Cox and Hinkley 1974), we can expand $\widehat{U}_r = 0$ to obtain $U_r + \sum_s U_{rs}(\widehat{\theta}_s - \theta_s) + O_p(1) = 0$ and write in matrix notation $U = J(\widehat{\theta} - \theta) + O_p(1)$, where U is the score vector and J is the observed information matrix. Since $J = K + O_p(n^{1/2})$ we have $U = K(\widehat{\theta} - \theta) + O_p(1)$ and

$$\widehat{\theta} - \theta = K^{-1}U + O_p(n^{-1}). \tag{1}$$

Equation 1 is important to provide higher-order moments and cumulants of the estimate $\widehat{\theta}$. If we now expand $\widehat{U}_r$ up to terms of second-order, we obtain

$$U_r + \sum_s U_{rs}(\hat{\theta}_s - \theta_s) + \frac{1}{2}\sum_{s,t} U_{rst}(\hat{\theta}_s - \theta_s)(\hat{\theta}_t - \theta_t) + o_p(1) = 0$$

and then by calculating its expected value

$$\sum_s \kappa_{rs}E(\hat{\theta}_s - \theta_s) + \sum_s \mathrm{Cov}(U_{rs}, \hat{\theta}_s - \theta_s) + \frac{1}{2}\sum_{s,t}\kappa_{rst}(-\kappa^{st}) + o(1) = 0. \tag{2}$$

Up to terms of order $O(n^{-1})$, we can write $\mathrm{Cov}(U_{rs}, \widehat{\theta}_s - \theta_s) = \mathrm{Cov}\left(U_{rs}, -\sum_t \kappa^{st}U_t\right) = -\sum_t \kappa_{rs,t}\kappa^{st}$.

Let $B(\widehat{\theta}_a)$ be the $O(n^{-1})$ bias of the estimate $\widehat{\theta}_a$ for $a = 1, \ldots, p$. Inverting Eq. 2, we can write $B(\widehat{\theta}_a)$ as (Cox and Snell 1968)

$$B(\widehat{\theta}_a) = \sum_{r,s,t}\kappa^{ar}\kappa^{st}\left(\kappa_{rs,t} + \frac{1}{2}\kappa_{rst}\right) = \sum_{r,s,t}\kappa^{ar}\kappa^{st}\left(\kappa_{rs}^{(t)} - \frac{1}{2}\kappa_{rst}\right). \tag{3}$$

We can verify that the two alternative forms for $B(\widehat{\theta}_a)$ are equivalent using Bartlett identity. In general regression models, we can derive matrix expressions for the bias of the estimate $\widehat{\theta}$, say $B(\widehat{\theta})$, from Eq. 3 when the cumulants κ's are invariant under permutations of parameters (see Cordeiro and McCullagh 1991).

The estimate $\widehat{\theta}$ can be inserted in $B(\widehat{\theta})$ to define bias corrected estimate $\widetilde{\theta} = \widehat{\theta} - \widehat{B}(\widehat{\theta})$, where $\widehat{B}(\widehat{\theta})$ is the value of $B(\widehat{\theta})$ at $\widehat{\theta}$. The bias corrected estimate $\widetilde{\theta}$ is expected to have better sampling properties than the classical estimate $\widehat{\theta}$. In fact, several simulations presented in the literature (Botter and Cordeiro 1998; Cordeiro et al. 2000; Vasconcellos and Silva 2005; Ospina et al. 2006; Patriota and Lemonte 2009) have shown that the corrected estimates $\widetilde{\theta}$ have smaller biases than their corresponding uncorrected estimates, thus suggesting that these bias corrections have the effect of shrinking the corrected estimates toward to the true parameter values. However, we can not say that the bias corrected estimates offer always some improvement

over the MLEs, since they can have larger mean squared errors than the uncorrected estimates.

We give a simple example by taking n *iid* observations from a normal distribution $N(\mu, \sigma^2)$, where we are interested to calculate the n^{-1} biases of the MLEs of μ and σ. The elements of the information matrix are: $\kappa_{\mu,\mu} = n/\sigma^2$, $\kappa_{\mu,\sigma} = 0$ and $\kappa_{\sigma,\sigma} = 2n/\sigma^2$. The third cumulantes are easily obtained: $\kappa_{\mu\mu\mu} = \kappa_{\mu,\mu\mu} = \kappa_{\sigma,\mu\mu} = \kappa_{\sigma,\mu\sigma} = \kappa_{\mu,\sigma\sigma} = \kappa_{\mu\sigma\sigma} = 0$, $\kappa_{\mu\mu\sigma} = -\kappa_{\mu,\mu\sigma} = 2n/\sigma^3$, $\kappa_{\sigma,\sigma\sigma} = -6n/\sigma^3$ and $\kappa_{\sigma\sigma\sigma} = 10n/\sigma^3$. Thus, $B(\widehat{\mu}) = 0$ since $\widehat{\mu} = \Sigma y_i/n$ has no bias. Further, after some algebra, $B(\hat{\sigma}) = -3\sigma/4n$. This result is in agreement with the exact bias of $\widehat{\sigma} = \{\Sigma(y_i - \overline{y})^2/n\}^{1/2}$ given by $E(\hat{\sigma}) = \sqrt{\frac{2}{n}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\sigma$, which is obtained from the χ^2_{n-1} distribution of $(n-1)\hat{\sigma}^2/\sigma^2$. In fact, using Stirling expansion in $E(\hat{\sigma})$ yields $E(\hat{\sigma}) = \sigma\left\{1 - \frac{3}{4n} + O(n^{-2})\right\}$. The corrected estimate of σ is then $\tilde{\sigma} = \left(1 + \frac{3}{4n}\right)\hat{\sigma}$.

For a one-parameter model, the n^{-1} bias of $\widehat{\theta}$ follows from Eq. 3 by setting all parameters equal to θ. We obtain a formula first derived by Bartlett (1953)

$$B(\hat{\theta}) = \kappa^{\theta\theta^2}\left(\kappa_{\theta\theta,\theta} + \frac{1}{2}\kappa_{\theta\theta\theta}\right) = \kappa^{\theta\theta^2}\left(\kappa^{(\theta)}_{\theta\theta} - \frac{1}{2}\kappa_{\theta\theta\theta}\right).$$

Stósic and Cordeiro (2009) presented simple programs (scripts) that may be used with algebraic manipulation software Maple and Mathematica to calculate closed-form analytic expressions for the bias corrections B_μ and B_ϕ of the MLEs of the parameters μ and ϕ for arbitrary two-parameter continuous distributions through a straightforward application of Eq. 3.

While these symbolic computation software packages have currently the ability to deal with analytic expressions of formidable size and complexity, limitations still exist, and it turns out that the complexity of the formulae involved in calculating the cumulants of log-likelihood derivatives for some distributions exceed their capacity. It is precisely for this reason that they presented equivalent scripts for calculating the bias correction terms in both frameworks. For some very rare cases neither Maple nor Mathematica were able to produce closed form expressions, but even for these, the current scripts may be expected to produce results on future versions of the softwares (under the assumption that backward compatibility of the scripting language is maintained). It should be pointed out that the above fact does not diminish the usefulness of either of the software packages (in their current versions), as both have produced closed form expressions in a large majority of the tested continuous density functions. Moreover, in all cases where both packages came up with a closed form expression, the results were found to be identical stressing the extremely high level of confidence that may be attributed to analytic manipulations involved.

For both Maple and Mathematica, after specifying the form and the domain of the density function $f(y;\mu,\phi)$ as well as the assumptions to be made on μ and ϕ (e.g. $\mu \in \mathbb{R}$ or $\mu > 0$), the program first defines and analytically calculates the cumulants (κ's), the second-order cumulants are then subsequently inserted into intermediate expression for the information matrix, in order to find the inverse information matrix, which are then used together with the third-order cumulants to produce the final result from Eq. 3. From now on, we use the notation $\psi'(p)$ and $\psi''(p)$ for the derivatives of the digamma function $\psi(p) = d\log\{\Gamma(p)\}/dp$, $\gamma = 1 - \psi(2)$ for the Euler's constant and $\zeta(p)$ for the Riemann Zeta function. The examples below are obtained using these programs and agree with previously reported results in the literature:

1. Normal distribution with mean μ and variance ϕ^2

$$nB_\mu = 0, \quad nB_\phi = -\frac{3\phi}{4n}.$$

2. Inverse Gaussian distribution with mean μ and precision parameter ϕ

$$nB_\mu = 0, \quad nB_\phi = \frac{3\phi}{n}.$$

3. Gamma distribution with mean μ and shape parameter ϕ

$$B_\mu = 0, \quad B_\phi = -\frac{2 - \phi\psi'(\phi) + \phi^2\psi''(\phi)}{2n\left(\phi\psi'(\phi) - 1\right)^2}.$$

4. Weibull distribution with scale μ and shape ϕ (here $E(y) = \mu\Gamma(1 + 1/\phi)$)

$$B_\mu = \frac{\mu}{2\pi^4\phi^2 n}\left\{\pi^4(1 - 2\phi) + 6\pi^2\left[1 + \gamma^2 + 5\phi - 2\gamma(1 + 2\phi)\right] + 72(\gamma - 1)\phi\zeta(3)\right\},$$
$$B_\phi = \frac{18\phi\left(\pi^2 - 2\zeta(3)\right)}{\pi^4 n}.$$

5. Logistic distribution with mean μ and variance $\pi^2\phi^2/6$

$$B_\mu = 0, \quad B_\phi = -\frac{9}{4n}\frac{\phi\left(4\pi^2 + 3\right)}{\left(\pi^2 + 3\right)^2}.$$

6. Extreme value distribution with mean $\mu + \gamma\,\phi$ and variance $\pi^2\,\phi^2/6$

$$B_\mu = \frac{\phi\left(3\,(-5+4\,\gamma)\,\pi^2+\pi^4-36\,(-1+\gamma)\,\zeta(3)\right)}{4\pi^4},$$
$$B_\phi = \frac{-12\,\phi\,\left(\pi^2-3\,\zeta(3)\right)}{4\pi^4}.$$

7. Beta distribution with parameters μ and ϕ

$$\begin{aligned}B_\mu = {} & \frac{1}{2\left[\psi'(\mu)\,(\psi'(\phi)-\psi'(\mu+\phi))-\psi'(\phi)\,\psi'(\mu+\phi)\right]^2 n}\\ & \times\{-[\psi'(\mu+\phi)\,(\psi'(\mu+\phi)\,(\psi''(\mu)-\psi''(\phi))\\ & +\psi'(\mu)\,\psi''(\phi))]\\ & +\psi'(\phi)^2\,[-\psi''(\mu)+\psi''(\mu+\phi)]\\ & +\psi'(\phi)\,[2\,\psi'(\mu+\phi)\,\psi''(\mu)+\psi'(\mu)\,\psi''(\mu+\phi)]\},\end{aligned}$$

$$\begin{aligned}B_\phi = {} & \frac{1}{2\left[\psi'(\mu)\,(\psi'(\phi)-\psi'(\mu+\phi))-\psi'(\phi)\,\psi'(\mu+\phi)\right]^2 n}\\ & \times\{\psi'(\mu+\phi)^2\,[\psi''(\mu)-\psi''(\phi)]\\ & +2\,\psi'(\mu)\,\psi'(\mu+\phi)\,\psi''(\phi)\\ & +\psi'(\mu)^2\,[-\psi''(\phi)+\psi''(\mu+\phi)]\\ & +\psi'(\phi)\,[-(\psi'(\mu+\phi)\,\psi''(\mu))+\psi'(\mu)\,\psi''(\mu+\phi)]\}.\end{aligned}$$

8. Student t distribution with location parameter μ and dispersion parameter ϕ

$$B_\mu = 0,\quad B_\phi = \frac{-3\left(-3+2\,\nu+\nu^2\right)\phi}{4\,n\,\nu\,(5+\nu)}.$$

When ν tends to infinity, we obtain $B_\phi = -3\phi/(4n)$ as is well known in the normal case.

9. Generalized Rayleigh distribution with mean $\frac{\Gamma\left(\phi+\frac{3}{2}\right)}{\sqrt{\mu}\,\Gamma(\phi+1)}$ and variance $\frac{1}{\mu}\left[1+\phi-\frac{\Gamma\left(\phi+\frac{3}{2}\right)^2}{\Gamma(\phi+1)^2}\right]$

$$B_\mu = \frac{\mu\,[-3\,\psi'(1+\phi)+2\,(1+\phi)\,\psi'(1+\phi)^2-(1+\phi)\,\psi''(1+\phi)]}{2\,[-1+(1+\phi)\,\psi'(1+\phi)]^2},$$
$$B_\phi = \frac{-2+(1+\phi)\,\psi'(1+\phi)-(1+\phi)^2\,\psi''(1+\phi)}{2\,n\,[-1+(1+\phi)\,\psi'(1+\phi)]^2}.$$

10. Type I Gumbel distribution with mean $\frac{\gamma+\log(\phi)}{\mu}$ and variance $\frac{\pi^2}{6\,\mu^2}$

$$B_\mu = \frac{18\,\mu\,\left(\pi^2-2\,\zeta(3)\right)}{\pi^4\,n},$$
$$\begin{aligned}B_\phi = {} & \frac{3\,\phi}{2\,\pi^4\,n}\{2\left(-6+4\,\gamma+\gamma^2\right)\pi^2+\pi^4+2\,\pi^2\log(\phi)^2\\ & +4\log(\phi)\,[(2+\gamma)\,\pi^2-6\,\zeta(3)]\\ & -24\,(-1+\gamma)\,\zeta(3)\}.\end{aligned}$$

11. Type II Gumbel distribution with mean $\phi^{\frac{1}{\mu}}\,\Gamma\left(1-\frac{1}{\mu}\right)$ and variance

$$\phi^{\frac{2}{\mu}}\left[\Gamma\left(1-\frac{2}{\mu}\right)-\Gamma\left(1-\frac{1}{\mu}\right)^2\right]$$

$$B_\mu = \frac{18\,\mu\,\left(\pi^2-2\,\zeta(3)\right)}{\pi^4\,n},$$
$$\begin{aligned}B_\phi = {} & \frac{3\,\phi}{2\,\pi^4\,n}\{2\left(-6+4\,\gamma+\gamma^2\right)\pi^2+\pi^4+2\,\pi^2\log(\phi)^2\\ & +4\log(\phi)\,[(2+\gamma)\,\pi^2-6\,\zeta(3)]\\ & -24\,(-1+\gamma)\,\zeta(3)\}.\end{aligned}$$

12. Fisher-Tippett distribution with mode μ and variance $\pi^2\,\phi^2/6$

$$B_\mu = \frac{\phi\left(3\,(-5+4\,\gamma)\,\pi^2+\pi^4-36\,(-1+\gamma)\,\zeta(3)\right)}{\pi^4\,n},$$
$$B_\phi = \frac{-12\,\phi\,\left(\pi^2-3\,\zeta(3)\right)}{\pi^4\,n}.$$

For these twelve distributions tested, Maple fails to yield closed form analytic expressions for bias corrections B_μ and B_ϕ for the extreme value, Student t, type I Gumbel and Fisher-Tippett distributions, whereas Mathematica fails only for the ▸logistic distribution. Comparing the equations obtained with Maple and Mathematica with some results previously reported in the literature, we note that, in most cases, all of the terms fully agree with the previously reported results, and where discrepancies were observed, it was found that the current results are correct, and errors were identified in the previous publications. This fact builds confidence regarding the correctness of the presented scripts, and the ability of these software for analytic formulae manipulation, so that application of the scripts to other density functions may be expected to produce reliable closed form expressions.

It is worth emphasizing that there are other methods to obtain bias corrected estimates. In regular parametric problems, Firth (1993) developed the so-called "preventive" method, which also allows for the removal of the second-order bias. His method consists of modifying the original score function to remove the first-order term from the asymptotic bias of these estimates. In exponential families with canonical parameterizations, his correction scheme consists in penalizing the likelihood by the Jeffreys invariant priors. This is a preventive approach to bias adjustment which has its merits, but the connections between our results and his work are not pursued here. We should also stress that it is possible to avoid cumbersome and tedious algebra on cumulant calculations by using Efron's bootstrap; see Efron and Tibshirani (1993). We use

the analytical approach here since this leads to a nice formula. Moreover, the application of the analytical bias approximation seems to generally be the most feasible procedure to use and it continues to receive attention in the literature.

About the Author

For biography *see* the entry ►Bartlett and Bartlett-Type Corrections.

Cross References

►Bartlett and Bartlett-Type Corrections
►Bias Analysis
►Estimation: An Overview
►Jackknife
►Likelihood
►Target Estimation: A New Approach to Parametric Estimation

References and Further Reading

Bartlett MS (1953) Confidence intervals II. Biometrika 40: 306–317

Botter DA, Cordeiro GM (1998) Improved estimators for generalized linear models with dispersion covariates. J Stat Comput Sim 62:91–104

Cook DR, Tsai CL, Wei BC (1986) Bias in nonlinear regression. Biometrika 73:615–623

Cordeiro GM, Klein R (1994) Bias correction in ARMA Models. Stat Probab Lett 19:169–176

Cordeiro GM, McCullagh P (1991) Bias correction in generalized linear models. J R Stat Soc B 53:629–643

Cordeiro GM, Vasconcellos KLP (1997) Bias correction for a class of multivariate nonlinear regression models. Stat Probab Lett 35:155–164

Cordeiro GM, Ferrari SLP, Uribe-Opazo MA, Vasconcellos KLP (2000) Corrected maximum-likelihood estimation in a class of symmetric nonlinear regression models. Stat Probab Lett 46:317–328

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London

Cox DR, Snell EJ (1968) A general definition of residuals (with discussion). J R Stat Soc B 30:248–275

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, London

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80:27–38

Ospina R, Cribari–Neto F, Vasconcellos KLP (2006) Improved point and interval estimation for a beta regression model. Comput Stat Data Anal 51:960–981

Patriota AG, Lemonte AJ (2009) Bias correction in a multivariate regression model with genereal parameterization. Stat Probab Lett 79:1655–1662

Stósic B, Cordeiro GM (2009) Using Maple and Mathematica to derive bias corrections for two parameter distributions. J Stat Comput Sim 79:751–767

Vasconcellos KLP, Silva SG (2005) Corrected estimates for student *t* regression models with unknown degrees of freedom. J Stat Comput Sim 75:409–423

Young DH, Bakir ST (1987) Bias correction for a generalized log-gamma regression model. Technometrics 29:183–191

Binomial Distribution

Andreas N. Philippou[1], Demetrios L. Antzoulakos[2]
[1]Professor of Probability and Statistics
University of Patras, Patras, Greece
[2]Associate Professor
University of Piraeus, Piraeus, Greece

The binomial distribution is one of the most important distributions in Probability and Statistics and serves as a model for several real life problems. Special cases of it were first derived by Pascal (1679) and Bernoulli (1713).

Definition and genesis. Denote by X the number of successes in a sequence of n (≥ 1) independent trials of an experiment, and assume that each trial results in a *success* (S) or a *failure* (F) with respective probabilities p ($0 < p < 1$) and $q = 1-p$. The random variable (rv) X is said to have the *binomial distribution with parameters n and p*, and it is denoted by $B(n,p)$. The probability mass function (pmf) $f(x)$ of X is given by

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n, \qquad (1)$$

where $\binom{n}{x} = n!/x!(n-x)!$ for $0 \leq x \leq n$ and 0 otherwise.

In fact a typical element of the event $\{X = x\}$ is a sequence $SSFS \ldots SF$ of x S's and $n - x$ F's, having probability $p^x q^{n-x}$ because of P(S) = p and the independence of the trials. Since there are $\binom{n}{x}$ such distinct allocations of x S's and $n - x$ F's, the result follows.

The name of the distribution is due to the binomial theorem, which implies that $B(n,p)$ is a proper probability distribution, since $f(x) \geq 0$ for $x \in R$ and

$$\sum_{x=0}^{n} f(x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (q+p)^n = 1. \qquad (2)$$

The cumulative distribution function (cdf) of X is related with the incomplete beta function by the formula

$$F(x) = P(X \leq x) = \sum_{i=0}^{x} \binom{n}{i} p^i q^{n-i}$$
$$= (n-x)\binom{n}{x} \int_0^q t^{n-x-1}(1-t)^x dt, \quad x = 0, 1, 2, \ldots, n. \qquad (3)$$

Its ▶moment generating function (mgf) is

$$M(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x q^{n-x} = (q + pe^t)^n \qquad (4)$$

from which the mean and the variance follow as

$$\mu = E(X) = M'(0) = np, \quad \sigma^2 = Var(X) = E(X^2) - \mu^2 = M''(0) - (M'(0))^2 = npq. \qquad (5)$$

When $n = 1$, the binomial distribution $B(n, p)$ is known as the Bernoulli distribution.

By substitution, we get from (1) the recurrence relation

$$f(x) = \frac{(n - x + 1)p}{xq} f(x - 1), \quad x = 1, 2, \ldots, n, \qquad (6)$$

which along with the initial condition $f(0) = q^n$ is very useful for calculating binomial probabilities. It follows from (6) that if $(n + 1)p$ is not an integer, then $f(x)$ has a unique mode at $x = [(n + 1)p]$. If $(n + 1)p$ is an integer, then $f(x)$ has two modes, one at $x = (n + 1)p$ and one at $x = (n + 1)p - 1$.

The binomial distribution arises whenever a sample of size n is drawn randomly with replacement from a population containing just two types of elements.

Urn models. From an urn containing w white and b black balls, n balls are drawn randomly *with replacement*. Let X be the number of white balls drawn. Then

$$f(x) = P(X = x) = \binom{n}{x}\left(\frac{w}{w+b}\right)^x \left(\frac{b}{w+b}\right)^{n-x}, \quad x = 0, 1, 2, \ldots, n. \qquad (7)$$

In fact considering as a *success S the drawing of a white ball*, and noting that that the n balls are drawn randomly with replacement, it follows that X is the number of successes in a sequence of n independent trials of an experiment with success probability $p = w/(w + b)$. The result then follows from (1).

Assume now that the balls are drawn randomly *without replacement* and denote by Y the number of white balls drawn. The rv Y follows the hypergeometric distribution with pmf

$$g(x) = P(Y = x) = \frac{\binom{w}{x}\binom{b}{n-x}}{\binom{w+b}{n}}, \quad x = 0, 1, 2, \ldots, n. \qquad (8)$$

If, in addition, $w/(w + b) \to p$ $(0 < p < 1)$, as $w \to \infty$ and $b \to \infty$, it follows from (8) that

$$\lim_{w,b\to\infty} g(x) = \binom{n}{x}p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n. \qquad (9)$$

The last equation practically means that $g(x)$ is approximately equal $(\simeq)$ to the RHS of (7) for large w and b in comparison to n. The approximation is considered adequate if $w + b > 10n$.

The binomial distribution converges to the following distribution named after Poisson.

Poisson convergence (Poisson 1837). Let X_n be a sequence of rv's distributed as $B(n, p_n)$ and assume that as $n \to \infty$, $p_n \to 0$ and $np_n \to \lambda$ (> 0). Then

$$\lim_{n\to\infty}\binom{n}{x}p_n^x(1 - p_n)^{n-x} = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots . \qquad (10)$$

In practice, (10) means

$$\binom{n}{x}p^x q^{n-x} \simeq e^{-np}\frac{(np)^x}{x!} \qquad (11)$$

for large n and small p. The approximation is quite accurate if $n \geq 20$ and $p \leq 0.05$ or if $n \geq 100$ and $p \leq 0.1$.

The binomial distribution can be approximated by the normal distribution.

Normal approximation. Let X be a rv distributed as $B(n, p)$. Since X can be viewed as a sum of n independent Bernoulli rv's, a direct application of the Central Limit Theorem yields the following approximation for the cdf of X

$$F(x) = P(X \leq x) \simeq \Phi\left(\frac{x - np}{\sqrt{npq}}\right) \qquad (12)$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. If a ▶*continuity correction* is used we have the following improved approximation

$$F(x) = P(X \leq x) \simeq \Phi\left(\frac{x + 0.5 - np}{\sqrt{npq}}\right) \qquad (13)$$

The approximation is fairly good provided n and p are such that $npq > 20$.

The approximation of binomial probabilities by the normal distribution was proved by de Moivre (1738) for $p = 1/2$ and for arbitrary values of p by Laplace (1812).

We end this note by the following example.

Example. The manager of Alpha Airlines knows from past data that 10% of the passengers who buy Alpha tickets do not show up for travel. Based on this, Alpha Airlines sell 220 tickets when the available seats in their planes are only 200. Find the probability that each ticket holder showing up will find a seat available and hence there will be no complaints.

Solution. Denote by X the number of ticket holders who do not show up for travel. Then the required probability is $P(X \geq 20) = 1 - P(X \leq 19)$. Assuming that each of the 220 ticket holders has the same probability

$p = 0.1$ of not showing up, independently of the others, we have that X is distributed as $B(220, 0.1)$. Therefore, the *exact probability* of no complaints is $P(X \geq 20) = 1 - \sum_{x=0}^{19} \binom{220}{x} (0.1)^x (0.9)^{220-x} = 0.7057$. The *Poisson approximation* by (11) gives $P(X \geq 20) \simeq 1 - \sum_{x=0}^{19} e^{-22} 22^x / x! = 0.6938$. The *normal approximation* by (12) is $P(X \geq 20) \simeq 1 - \Phi\left((19-22)/\sqrt{19.8}\right) = \Phi(0.6742) = 0.7499$. Finally, the *normal approximation with continuity correction* by (13) is $P(X \geq 20) \simeq 1 - \Phi\left((19+0.5-22)/\sqrt{19.8}\right) = \Phi(0.5618) = 0.7129$.

About the Author

For biography *see* the entry ►Distributions of order K.

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Distributions of Order K
►Generalized Linear Models
►Geometric and Negative Binomial Distributions
►Inverse Sampling
►Modeling Count Data
►Multivariate Statistical Distributions
►Proportions, Inferences, and Comparisons
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Statistical Methods in Epidemiology
►Univariate Discrete Distributions: An Overview

References and Further Reading

Bernoulli J (1713) Ars conjectandi. Thurnisius, Basilea
de Moivre A (1738) The doctrine of chances, 2nd edn. Woodfall, London
Laplace PS (1812) Théorie Analytique des Probabilités, 3rd edn. 1820, Courcier Imprimeur, Paris. Reprinted by EJ Gabay, 1992, Paris
Pascal B (1679) Varia opera Mathematica D. Petri de Fermat, Tolossae
Poisson SD (1837) Récherches sur la probabilité des jugements en matiere criminelle et en matiere civile, precedees des regles generales du calcul des probabilites, Paris: Bachelier, Imprimeur-Libraire pour les Mathematiques, la Physique, etc.

Bioinformatics

Susan R. Wilson
Professor, Faculty of Medicine, Faculty of Science
University of New South Wales, Sydney, NSW, Australia

Bioinformatics is a relatively young, cross-disciplinary research area at the intersection of the biological sciences with the mathematical, statistical, and physical sciences and chemistry and information technology. In the past decade or so, there has been phenomenal growth of life science databases. For example, the most widely used nucleotide sequence database is Genbank that is maintained by the National Center for Biotechnology Information (NCBI) of the US National Library of Medicine; as of February 2008 it contained approximately 86 billion nucleotides from 83 million sequences. Its size continues to grow exponentially as more genomes are being sequenced. However, there is a very large gap (that will take a long time to fill) between our knowledge of the functioning of the genome and the generation (and storing) of raw genomic data. This overview touches briefly on those aspects of bioinformatics that will be of interest to statisticians.

The stated goal for many researchers is for developments in Bioinformatics to be focused at finding the fundamental laws that govern biological systems, as in physics. However, if such laws exist, they are a long way from being determined for biological systems. Instead the current aim is to find insightful ways to model limited components of biological systems and to create tools which biologists can use to analyze data. Examples include tools for statistical assessment of the similarity between two or more DNA sequences or protein sequences, for finding genes in genomic DNA, for quantitative analysis of functional genomics data, and for estimating differences in how genes are expressed in say different tissues, for analysis and comparison of genomes from different species, for phylogenetic analysis, and for DNA sequence analysis and assembly. Tools such as these involve statistical modeling of biological systems. Although the most reliable way to determine a biological molecule's structure or function is by direct experimentation, there is much that can be achieved *in vitro*, i.e., by obtaining the DNA sequence of the gene corresponding to an RNA or protein and analyzing it, rather than the more laborious finding of its structure or function by direct experimentation.

Much biological data arise from mechanisms that have a substantial probabilistic component, the most significant being the many random processes inherent in biological evolution, and also from randomness in the sampling process used to collect the data. Another source of variability or randomness is introduced by the biotechnological procedures and experiments used to generate the data. So the basic goal is to distinguish the biological "signal" from the "noise". Today, as experimental techniques are being developed for studying genome wide patterns, such as expression arrays, the need to appropriately deal with the inherent variability has been multiplied astronomically. For example, we have progressed from studying one or

a few genes in comparative isolation to being able to evaluate simultaneously thousands of genes. Not only must methodologies be developed which scale up to handle the enormous data sets generated in the post-genomic era, they need to become more sensitive to the underlying biological knowledge and understanding of the mechanisms that generate the data. For statisticians, research has reached an exciting and challenging stage at the interface of computational statistics and biology. The need for novel approaches to handle the new genome-wide data has coincided with a period of dramatic change in approaches to statistical methods and thinking. This "quantum" change has been brought about, or even has been driven by, the potential of ever more increasing computing power. What was thought to be intractable in the past is now feasible, and so new methodologies need to be developed and applied.

Unfortunately too many of the current practices in the biological sciences rely on methods developed when computational resources were very limiting and are often either (a) simple extensions of methods for working with one or a few outcome measures, and do not work well when there are thousands of outcome measures, or (b) ad-hoc methods (that are commonly referred to as "statistical" or "computational", or more recently "data mining"! methods (see ►Data Mining)) that make many assumptions for which there is often no (biological) justification. The challenge now is to creatively combine the power of the computer with relevant biological and stochastic process knowledge to derive novel approaches and models, using minimal assumptions, and which can be applied at genomic wide scales. Such techniques comprise the foundation of bioinformatic methods in the future.

Hidden Markov model (HMMs) is a tool that is popular in bioinformatics. Here, applications of HMMs are related to prediction of protein-coding regions in genome sequences, modeling families of related DNA or protein sequences or prediction of secondary structure elements from protein primary sequences. Biological data are the result of evolution, that is, the result of an incredibly complex, and currently unknown, stochastic process. Very simplified models of this process are often used, particularly for the construction of phylogenetic trees. Such evolutionary models have been developed as both discrete time and continuous time processes. A key requirement of these processes is that they be reversible, since statistical comparisons of, say, two contemporary species, can require one tracing up the tree of evolution to a common ancestor and then down the tree to the other species, and so the stochastic process for tracing must be the same in each direction.

Of recent years, gene expression data have become of increasing interest to statistical scientists. The technology has evolved too. Such data are being generated using technologies like microarrays, and very recently, next-generation sequencing. These data can be thought of as lying in very high dimensional space, and the resultant challenges are at the forefront of modern statistical science.

In the past, biologists looked at genes and proteins one at a time. Now we have the technology to start to look at all the elements – genes, mRNA, proteins, interactions, and so on - in a biological system, and to explore their relationships as the system functions in response to biological, and environmental, perturbations. This is known as systems biology. The long-term goal is to have (mathematical) models to describe the behavior of the system given any kind of perturbation, and then to be able to redesign systems by modifications, such as drugs, to have completely new properties. The key question is whether it is possible to take a systematic approach to mapping pathways. The preliminary step involves building a comprehensive "scaffold" of molecular interactions that broadly covers many aspects of cellular function and physiological responses. At the broad, overall level, statistical data mining approaches are used to search for patterns and relationships between variables, and Bayesian networks model conditional dependencies. At a more detailed level, Markov chains model predictions, loss and interconversion among molecular species and states. Today, systems biology is very much in its infancy, but with a challenging future that will also involve expansion to include many more levels of complexity.

An excellent starting point reference for statisticians is Ewens and Grant (2005). Bioconductor (http://www.bioconductor.org) is an open source and open development software project for the analysis of genomic data.

About the Author

For biography *see* the entry ►Biostatistics.

Cross References

►Biostatistics
►Forensic DNA: Statistics in
►Statistical Genetics

References and Further Reading

Ewens WJ, Grant G (2005) Statistical methods in bioinformatics, 2nd edn. Springer, New York

Biopharmaceutical Research, Statistics in

Christy Chuang-Stein
Vice President of the American Statistical Association (2009–2011), Head
Pfizer Inc., Kalamazoo, MI, USA

"*Statistics in Biopharmaceutical Research*" is the title of the on-line journal launched by the American Statistical Association in 2009. There are at least two other international peer-reviewed journals completely dedicated to the use of statistics to support the development of pharmaceutical products. They are *Journal of Biopharmaceutical Statistics* (Taylor and Francis Group) and *Pharmaceutical Statistics* (John Wiley and Sons). There are many books devoted to this area of statistical applications also, e.g., Senn (2008), Dmitrienko et al. (2007) and Dmitrienko et al. (2005). In the United States, pharmaceutical and biotech industries employ thousands of statisticians, either directly or indirectly. These statisticians support the discovery, development and commercialization of valuable medicines, medicines that have made substantial contributions to a longer life expectancy and a better quality of life in the past 50 years.

The development of pharmaceutical products is a long and high risk proposition. It takes an average of 15 years for a new compound to be discovered and eventually submitted to regulators for approval. As of 2008, the cost of developing a new drug is estimated to be between $800 million and $2 billion US dollars (Masia 2008). Biopharmaceutical research begins in the laboratory where chemists synthesize compounds and biologists screen the compounds for activities. Because of the large number of compounds, it is essential to develop an efficient algorithm-based process to conduct high-throughput screening for the maximum yield of promising compounds. Once a compound is judged to meet the level of required potency, it needs to go through formulation development so that the active ingredient could be delivered to the target site of actions in test subjects. The first test subjects are laboratory animals used to evaluate the effect of the compound on cardiovascular function, reproductive function, tumor development and the general wellbeing of offspring born to animals exposed to the compound. Most of the animal experiments are conducted according to the International Conference on Harmonisation (ICH) guidance M3(R2) (2009) on nonclinical safety studies. The need to use the smallest number of animals at this preclinical testing stage has led to the use of efficient experimental designs with repeated measures on each animal. In addition, data mining techniques (see ►Data Mining) are used widely to search for chemical and physical properties that tend to associate with compounds that turn out to be successful.

The majority of statistical support in biopharmaceutical research takes place during the clinical testing in humans. In the United States, this support has grown substantially since the 1962 Kefauver-Harris Amendment (Krantz 1966) that required drug sponsors to prove a product's safety and efficacy in controlled clinical trials before receiving marketing authorization. It is usually thought that the first properly randomized control trial in the twentieth century that was recorded involved the use of streptomycin for the treatment of pulmonary tuberculosis (MRC, 1948). Since that time, the number of clinical trials (both randomized and non-randomized) has skyrocketed as evidenced by the number of trials registered at the www.clinicaltrials.gov site in the United States.

The clinical development of a new treatment is often divided into three phases. All clinical trials need to follow ICH E6 guidance on good clinical practice (1996) and the 1964 Declaration of Helsinki. In Phase 1 trials, healthy volunteers are randomized to receive a single dose or multiple doses of the compound or a placebo to study the tolerance and pharmacokinetics of the new compound. The objective is to decide an acceptable dose range. For cytotoxic agents, Phase 1 trials are conducted in cancer patients with the objective to estimate the maximum tolerated dose. Because of the small number of subjects (e.g., 40–100) at this stage, safety evaluation focuses on identifying common side effects of the new treatment. If the tolerance and pharmacokinetic profiles based on the limited data are judged to be acceptable, testing will proceed to Phase 2.

In Phase 2, treatment using the new compound will be compared against a concurrent comparator (placebo, an approved product or the standard of care) in patients with the target disease. The primary objective is to gather safety and efficacy data in patients. Different dose strengths are typically used in these trials to help estimate the dose–response relationship. The latter often involves fitting an Emax model or logistic model. For oncology trials, Phase 2 studies can be a single arm study using the maximum tolerated dose estimated from Phase 1. Some researchers further divide Phase 2 into Phase 2a proof-of-concept and Phase 2b dose-ranging studies. The former often includes a single dose strength to verify the hypothesized mechanism while the latter uses different dose strengths and clinically relevant endpoints. Phase 1 and Phase 2 trials are designed to help a sponsor learn about the new treatment (Sheiner 1997). They are exploratory in nature and the analysis will focus on estimation instead of hypothesis testing.

This is a critical phase of product development. It is during this stage that important information on dose(s), dosing schedule(s), endpoints and the target population will be evaluated and decided upon. Statistics is heavily used to design trials, to analyze the results and to make Go or No-Go decisions.

If data from the Phase 2 development offer good reasons to believe that the new treatment has a positive benefit to risk balance and can bring value to patients, development will move into the confirmatory stage, or Phase 3. In general, Phase 3 trials are double-blind randomized trials if blinding is at all possible. These trials are typically large (hundreds to thousands of patients) with a longer duration. The primary objective is to confirm the presence of a treatment effect and to collect additional safety information in a more diverse population. In terms of efficacy assessment, it could be either superiority over a comparator or non-inferiority to an active control (ICH E10 2000). For life-threatening conditions, interim analyses are often conducted to stop a trial early for efficacy (positive outcome) or futility (negative outcome) for ethical reasons. Interim analyses are typically conducted by individuals independent of the study and reviewed by an Independent Data Monitoring Committee (FDA DMC guidance 2006). Except for those pre-specified in the protocol as possible mid-trial adaptations, changes are strongly discouraged at this stage. When multiple comparisons (multiple doses, multiple endpoints, multiple subgroups, interim efficacy analysis etc.) are conducted for inferential purposes, the significance levels for comparisons need to be properly adjusted so that the family-wide Type I error rate is strongly controlled. The adjustment method needs to be pre-specified and can't be changed once the trial results become known. In short, the design and analysis of Phase 3 trials need to be carefully planned and rigorously executed. Statistical principles, as articulated in ICH E9 (1998), should be followed with very few deviations. Deviations, when occurring, need to be justified and sensitivity analyses should be conducted to evaluate the impact of the deviations on conclusions. Statistics is the basis for inferential conclusions in these trials.

After a product receives a marketing authorization, testing typically continues for additional uses of the product. A new product could also be included in a head-to-head comparison against another marketed product for differentiation or comparative effectiveness research. Much of the work on comparative effectiveness is to support the valuation of a new product, particularly in regions where a government board decides if a new product is eligible for reimbursement and the price of the product under a national healthcare system. These efforts often involve pooling data from multiple studies and can rely on endpoints different from those used to make marketing authorization decision. The work requires statisticians to collaborate closely with health economists, health care providers, third party payers and patients. Systematic review (including ►meta-analysis) is often an integral part of such efforts.

Following several highly visible product withdrawals, the safety of pharmaceutical products has been a major source of public attention in recent years. Data from clinical trials, spontaneous reports of adverse reactions collected in pharmacovigilance databases and longitudinal patient information from healthcare or claims databases are increasingly used to explore possible product-induced injuries. Statistical techniques, based on the concept of proportionality (Almenoff et al. 2007), have been developed and applied extensively to look for possible safety signals.

The decrease in the overall productivity measured by the number of approved new molecular entities each year has led industry and regulators to look for better ways to conduct biopharmaceutical research. Examples include FDA's Critical Path Initiative (2004, http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/) and European Union's Innovative Medicines Initiative (2007) (http://www.imi.europa.eu/index_en.html). One outcome from this emphasis is the extensive research on adaptive trial designs over the last 5 years (Gaydos et al. 2009; Bornkamp et al. 2007; Bretz et al. 2006; Gallo et al. 2006 etc.). Research on adaptive design includes designs for dose-ranging studies and confirmatory studies. Central to the concept of adaptive designs is a more efficient use of data and a more agile response to accumulated evidence on the effect of a new treatment.

In biopharmaceutical research, statisticians are at the heart of evidence collection, synthesis and communication. Statisticians have enormous opportunities and face probably an equal number of challenges. We can expect both opportunities and challenges to increase in the twenty-first century. Statisticians need to be in tune with the dynamic environment, to help meet the needs of multiple customers, to cash in on the opportunities and rise to the challenges (Chuang-Stein et al. 2010)!

About the Author

Dr. Christy Chuang-Stein is Head of the Statistical Research and Consulting Center, Pfizer Inc, United States. She is a Vice President of the American Statistical Association (2009–2011) and a past Chair of the Biostatistics and Data Management Technical Group of the Pharmaceutical Research and Manufacturers of America (2003–2004).

She is a Founding Editor (with two others) of the journal *Pharmaceutical Statistics* (2002–2005). She is a Fellow of the American Statistical Association (elected in 1998). She has authored and co-authored more than 120 papers including a book on *Analysis of Clinical Trials Using SAS: A Practical Guide* (2005). She co-edited a book on *Pharmaceutical Statistics Using SAS* (2007). Dr. Chuang-Stein received the Donald E Francke Award for Overall Excellence in Journal Publishing (2001, 2006, 2009) and the Thomas Teal Award for Excellence in Statistics Publishing (2008, 2010) from the Drug Information Association. She also received the Excellence in Continuing Education Award from the American Statistical Association (2005). She has served as an Associate Editor for The *American Statistician* (1993–1999), *Journal of Biopharmaceutical Statistics* (2000–2002), *Wiley Encyclopedia of Clinical Trials* (2005–2008) and *Drug Information Journal* (1996–present).

Cross References

►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest
►Medical Research, Statistics in
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences

References and Further Reading

Almenoff JS, Pattishall EN, Gibbs TG, DuMouchel W, Evans SJ, Yuen N (2007) Novel statistical tools for monitoring the safety of marketed drugs. Clin Pharm Ther 82:157–166

Bornkamp B, Bretz F, Dmitrienko A, Enas G, Gaydos B, Hsu C-H, Konig F, Liu Q, Neuenschwande B, Parke T, Pinheiro J, Roy A, Sax R, Shen F (2007) Innovative approaches for designing and analyzing adaptive dose-ranging Trials. J Biopharm Stat 17: 965–995

Bretz F, Schmidli H, Racine A, Maurer W (2006) Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. Biometrical J 48(4):623–634

Chuang-Stein C, Bain R, Branson M, Burton C, Hoseyni C, Rockhold F, Ruberg S, Zhang J (2010) Statisticians in the pharmaceutical industry: the 21st century. J Biopharm Stat 2(2):145–152

Declaration of Helsinki, History. Available at http://cme.cancer.gov/c01/a02_02.htm

Dmitrienko A, Molenberghs G, Chuang-Stein C, Offen W (2005) Analysis of clinical trials using SAS: a practical guide. Cary NC, SAS

Dmitrienko A, Chuang-Stein C, D'Agostino R (2007) Pharmaceutical statistics using SAS. SAS, Cary NC, SAS

FDA Guide for Clinical Trial Sponsors – Establishment and Operation of Clinical Trial Data Monitoring Committees, March 27 2006. Available at http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm127073.pdf.

Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J (2006) Adaptive designs in clinical drug development. J Biopharm Stat 16:275–283

Gaydos B, Anderson K, Berry D, Burnham N, Chuang-Stein C, Dudinak J, Fardipour P, Gallo P, Givens S, Lewis R, Maca J, Pinheiro J, Pritchett Y, Krams M (2009) Good practices for adaptive clinical trials in pharmaceutical product development. Drug Inform J 43(5):539–556

International Conference on Harmonisation E6(R1), Guidance for Good Clinical Practice, Step 5, May 1996. Available at http://www.ich.org/LOB/media/MEDIA482.pdf

International Conference on Harmonisation E9, Statistical Principles for Clinical Trials, Step 5, Feb 1998. Available http://www.ich.org/LOB/media/MEDIA485.pdf

International Conference on Harmonisation E10, The Choice of Control and Relatede Issues in Clinical Trials, Step 5, July 2000. Available at http://www.ich.org/LOB/media/MEDIA486.pdf

International Conference on Harmonisation M3(R2): Maintenance of the ICH Guideline on Non-Clinical Safety Studies for the Conduct of Human Clinical Trials for Pharmaceuticals, Step 5, June 2009. Available at http://www.ich.org/LOB/media/MEDIA5544.pdf

Krantz JC Jr (1966) New drugs and the Kefauver-Harris amendment. J New Drugs 6:77–79

Masia N (2008) The cost of developing a new drug. In Focus on intellectual property rights. A US Department of State Publication, April 23 2008. Available at http://www.america.gov/st/econ-English/2008/April/20080429230904myleen0.5233981.html

Medical Research Council (1948) Streptomycin treatment of pulmonary tuberculosis. Br Med J 2:769–782

Senn S (2008) Statistical issues in drug development (statistics in practice), 2nd edn. Wiley, New York

Sheiner LB (1997) Learning versus confirming in clinical drug development. Clin Pharm Ther 61:275–291

Biostatistics

Susan R. Wilson
Professor, Faculty of Medicine, Faculty of Science
University of New South Wales, Sydney, NSW, Australia

The term Biostatistics is formed from the words biology and statistics, but these days more commonly refers to a somewhat narrower coverage of statistical methods needed in medicine and public health. In its broadest sense, biostatistics is the science of collecting and analyzing biological data to create knowledge about biological processes. The field of biostatistics (even as the term is generally used today) is wide and in its scope includes applications in clinical medicine, public health, epidemiology, genetics (genomics and all the other 'omics), health services, ►demography, and laboratory research. An essential part of the practice of biostatistics is collaboration between the statistician and the medical scientist or health professional.

To give some insight into the very many facets encompassed by biostatistics we use the five central principles

of applied statistics (of which Biostatistics arguably can be regarded as the largest branch) proposed by Cox (2007).

1. *Formulation of Objectives:* This can vary very widely, dependent on the biological, health and related environmental issues being studied. Such issues range over, for example, calculations of birth and death rates, and life expectancy, finding the genomic and environmental basis of complex disease, evaluating the efficacy of a new treatment/drug, determination of efficacious dose for reliable drug production in pharmaceutical research, improving the delivery of health care. It can be useful to distinguish between decision-making and inference. A Bayesian approach is usually to be preferred for decision making in a narrow sense. Inference of causality is often a major objective and is widely studied, particularly in epidemiology. In the early stage of an investigation, biostatisticians can be important in helping to focus on the specific formulation of concrete objectives.
2. *Study design:* This ranges from observational studies and sample surveys, to secondary analysis of collected data, through to experimental design. Major study types include the ecologic study, the cross-sectional study, the case-control study, the cohort study and clinical trials.
3. *Measurement:* The actual entities being measured depend on the objectives, and specific situation, and range from instrument measurements to quality of life measures. Basic criteria include relevance, precision and minimal (preferably no) bias. Care may be needed to ensure that data quality does not drop if too many measurements are being taken, particularly when dealing with human subjects. One also needs to avoid any confounding effect on the system due to the measurement process itself; a much-studied example is the placebo effect in human studies.
4. *Analysis of Biostatistical Data:* The three basic, common phases in any statistical data analysis are data editing and quality control, preliminary, often exploratory, analysis and graphics, and more detailed analysis. Of recent years there has been an explosion in statistical reasoning and methods for analysis of studies of human health. Examples include developments in epidemiological methods, clinical trials, survival analysis, statistical inference based on likelihood methods, ►statistical genetics. Although much analysis is based on probabilistic models, increasingly purely algorithmic approaches, such as cluster analysis (see ►Cluster Analysis: An Introduction), machine learning and ►bioinformatics, are being used. This is particularly the case in the analyses of extremely large data sets where the current challenge is dealing with the extremely large number of measurements that is many times the number of observations.
5. *Interpretation:* The borderline with analysis is not clear-cut, but it is fundamentally important that biostatisticians present the conclusions in a way that is readily interpretable by the relevant members of the health community.

The human context of biostatistics differentiates it from other areas of applied statistics. In particular, ethical issues may arise, patients often do not make ideal "experimental units" and establishment of causation (as in all areas of biology) can be problematic. To achieve evidence-based medical practice, the Cochrane Collaboration was established in 1993 with the aims of facilitating meta-analyses of randomized clinical trials, to disseminate results effectively, and update these regularly.

Many parts of the global field of biostatistics have become disciplines in their own right, such as statistical genetics, demography, actuarial science, methods for clinical trials, as well as bioassay. Often there is overlap with a part of another scientific area; one such example is the recently emerged discipline of bioinformatics.

Many journals are either entirely devoted to biostatistics (including *Biostatistics*, *Statistics in Medicine*, *Statistical Methods for Medical Research*) or have a substantial part devoted (including *Biometrics*, *Biometrical Journal*). There are specialist journals for specific subareas, (such as *Pharmaceutical Statistics*, *Journal of the Society for Clinical Trials*). Besides specialist societies dealing with a subarea, many of the major statistical societies have a Biostatistics (or ►Medical Statistics) section.

Biostatistical software has been expanding rapidly to handle developments of new methodology, as well as changing to meet the ongoing improvements in computer power and capability. There is an enormous, exponentially growing, number of biostatistics reference and textbooks, aimed at the great variety of backgrounds of those interested in biostatistics.

The *Encyclopedia of Biostatistics* (2005) offers a relatively definitive reference on the development and use of statistical methods for addressing the problems and critical questions that confront scientists, practitioners and policy makers involved in the medical and life sciences. Therein, broad sections of the subject were identified, covering specific biostatistical work (clinical trials, epidemiologic studies, clinical epidemiology, vital and health statistics, health services research, laboratory studies, biological models, health surveys and biomedical experiments), as well as particular branches of statistical methodology of special

biostatistical interest (►categorical data analysis, statistical models, longitudinal data analysis, multivariate analysis, survival analysis, sampling and experimental design, statistical computing), medical specialities with statistical applications, human genetics and genetic epidemiology.

The growing importance and application of ►biostatistics is reflected in the increasing number of statisticians employed in the healthcare sector, pharmaceutical industry and medical schools. The boundaries are vague as the discipline of biostatistics is broad, linking a theoretical discipline, namely mathematical statistics, to a diversity of applied sciences relevant to medicine and human health.

About the Author

Susan Wilson is Professor in the Faculties of Science and of Medicine at the University of New South Wales. She previously was Director, Centre for Bioinformation Science, Mathematical Sciences Institute, Australian National University. Professor Wilson is an elected Fellow of the Institute of Mathematical Statistics and of the American Statistical Association. She is a Past-President of the International Biometric Society.

Cross References

►Bioinformatics
►Biopharmaceutical Research, Statistics in
►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest
►Forensic DNA: Statistics in
►Medical Research, Statistics in
►Medical Statistics
►Statistical Genetics
►Statistical Methods in Epidemiology

References and Further Reading

Cox DR (2007) Applied statistics: a review. Ann Appl Stat 1:1–16
Armitage P, Colton T (eds) (2005) Encyclopedia of biostatistics, Wiley, Chicheste

Bivariate Distributions

Colin Rose
Director
Theoretical Research Institute, Sydney, NSW, Australia

Bivariate distributions allow one to model the relationship between two random variables, and thus they raise subject areas such as dependence, correlation and conditional distributions. We consider the continuous and discrete cases, separately.

Continuous Bivariate Distributions

Let (X, Y) denote two random variables defined on a domain of support $\Lambda \subset \mathbb{R}^2$, where we assume Λ is an open set in $\mathbb{R}^2$. Then a function $f : \Lambda \to \mathbb{R}_+$ is a joint bivariate pdf (*probability density function*) if it has the following properties:

$$f(x,y) > 0, \text{ for } (x,y) \in \Lambda$$
$$\int_\Lambda \int f(x,y)dxdy = 1 \quad (1)$$
$$P((X,Y) \in S) = \int_S \int f(x,y)dxdy, \text{ for any } S \subset \Lambda$$

The joint cdf (*cumulative distribution function*) is given by:

$$F(x,y) = P(X \le x, Y \le y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(v,w)dvdw \quad (2)$$

where $0 \le F(x,y) \le 1$. The probability content of a rectangular region $S = \{(x,y) : a < x < b,\ c < y < d\}$ can be expressed in terms of the cdf $F(x,y)$ as:

$$P(a < X < b,\ c < Y < d) = F(a,c) - F(a,d) - F(b,c) + F(b,d) \quad (3)$$

The *marginal pdf* of X, denoted $f_x(x)$, is:

$$f_x(x) = \int_y f(x,y)dy \quad (4)$$

and similarly, the *marginal pdf* of Y, denoted $f_y(y)$, is:

$$f_y(y) = \int_x f(x,y)dx \quad (5)$$

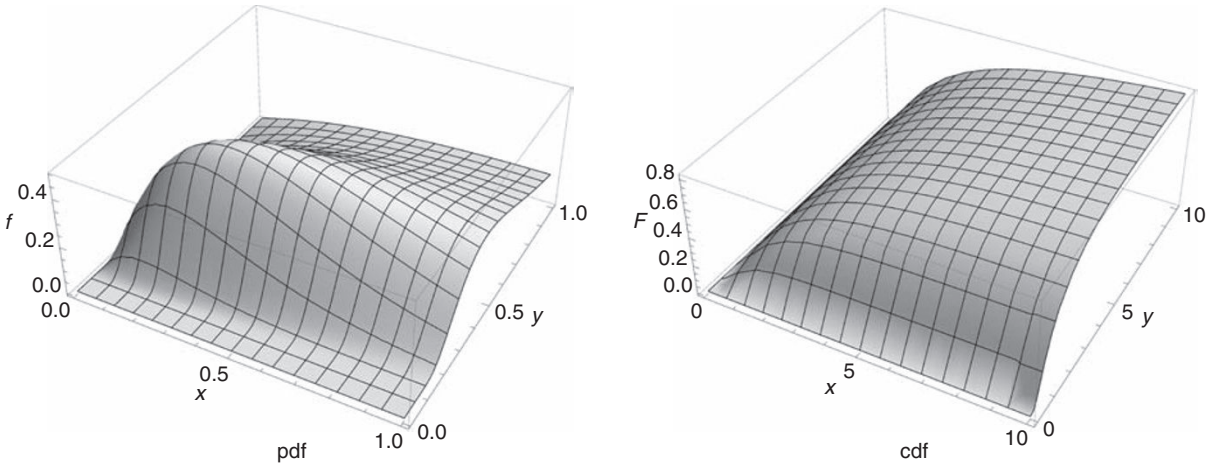
The *conditional pdf* of X given $Y = y$ is denoted by $f(x|Y = y)$ or, for short, $f(x|y)$. It is defined by

$$f(x|y) = \frac{f(x,y)}{f_y(y)}, \quad \text{provided } f_y(y) > 0 \quad (6)$$

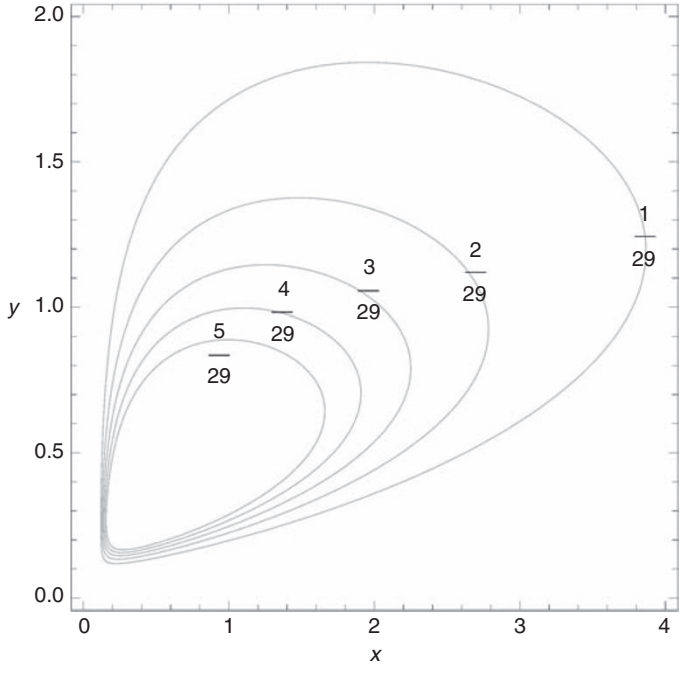
Table 1 lists some well-known bivariate pdf's. Whereas, in a univariate world, one speaks of *the* ►Gamma distribution (one single functional form), by contrast, in a bivariate world, there are a multitude of different bivariate Gamma distributions. This is because the term bivariate Gamma, or bivariate Exponential, or bivariate Uniform etc. is applied to essentially any bivariate distribution whose marginal pdf's are univariate Gamma, Exponential, or Uniform, respectively.

Bivariate Distributions. Table 1 Some bivariate continuous pdf's

Distribution	pdf	Domain	Parameters
Bivariate Normal (standardised)	$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)$	$(x,y) \in \mathbb{R}^2$	$-1 < \rho < 1$
Bivariate Normal (general)	$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2}{2(1-\rho^2)}\right)$	$(x,y) \in \mathbb{R}^2$	$-1 < \rho < 1$, $\sigma_x > 0, \sigma_y > 0$
Bivariate T (standardised)	$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{x^2 - 2\rho xy + y^2}{v(1-\rho^2)}\right)^{-1-\frac{v}{2}}$	$(x,y) \in \mathbb{R}^2$	$-1 < \rho < 1$, $v > 0$
Bivariate Cauchy	$f(x,y) = \frac{1}{2\pi}(1 + x^2 + y^2)^{-3/2}$	$(x,y) \in \mathbb{R}^2$	
Bivariate Exponential (Gumbel Type *I*)	$f(x,y) = e^{-x-y-xy\theta}((1+\theta x)(1+\theta y) - \theta)$	$x > 0,\ y > 0$	$-1 \le \theta \le 1$
Bivariate Exponential (Gumbel Type *II*)	$f(x,y) = \frac{e^{x+y} + \alpha(e^x - 2)(e^y - 2)}{e^{2(x+y)}}$	$x > 0, y > 0$	$-1 < \alpha < 1$
Bivariate Gamma (McKay)	$f(x,y) = \frac{c^{a+b}}{\Gamma[a]\Gamma[b]} x^{a-1}(y-x)^{b-1}e^{-cy}$	$0 < x < y < \infty$	$a, b, c > 0$
Bivariate Logistic (Gumbel)	$f(x,y) = \frac{2e^{-x-y}}{(1+e^{-x}+e^{-y})^3}$	$(x,y) \in \mathbb{R}^2$	
Bivariate Uniform (Morgenstern)	$f(x,y) = 1 + \alpha(2x-1)(2y-1)$	$0 \le x \le 1$, $0 \le y \le 1$	$-1 \le \alpha \le 1$



Bivariate Distributions. Fig. 1 The joint pdf $f(x,y) = \exp[(-1-x)/y]x/y^4$ of Example 1 (left) and the cdf $F(x,y)$ (right)



Bivariate Distributions. Fig. 2 Contours of the joint pdf $f(x,y) = \exp[(-1-x)/y]x/y^4$ of Example 1, illustrated when $f(x,y) = 1/29, 2/29, 3/39, 4/29, 5/29$

Example 1 Joint pdf

Consider the function $f(x,y) = e^{\frac{-1-x}{y}} x/y^4$ with domain of support $\Lambda = \{(x,y) : 0 < x < \infty, 0 < y < \infty\}$. Clearly, f is positive over its domain, and it integrates to unity over the domain. Thus, $f(x,y)$ can represent the joint pdf of a pair of random variables. Figure 1 plots $f(x,y)$ over part of its support, and the cdf over a somewhat wider region of its support …

A contour plot allows one to pick out specific contours along which $z = f(x,y)$ is constant. That is, each

contour joins points on the surface that have the same height z. Figure 2 plots all combinations of x and y such that $f(x,y) = \frac{1}{29}, \frac{2}{29}, \frac{3}{29}, \frac{4}{29}$ and $\frac{5}{29}$.

For extensive detail on continuous bivariate distributions, see Balakrishnan and Lai (2009).

Constructing Continuous Bivariate Distributions: Copulae

Copulae (see ▶Copulas) provide a method for constructing bivariate distributions from known marginal distributions.

Let the continuous random variable X have cdf $\Phi(x)$; similarly, let the continuous random variable Y have cdf $G(y)$. We wish to create a bivariate distribution $H(x,y)$ from these known marginals. The joint distribution function $H(x,y)$ is given by

$$H(x,y) = C(\Phi, G) \tag{7}$$

where C denotes the copula function. Then, the joint pdf $h(x,y)$ is given by

$$h(x,y) = \frac{\partial^2 H(x,y)}{\partial x\, \partial y} \tag{8}$$

Table 2 lists some examples of copulae.

With the exception of the independent case, each copula in Table 2 includes parameter α. This term induces a new parameter into the joint bivariate distribution $h(x,y)$, which gives added flexibility. In each case, setting parameter $\alpha = 0$ (or taking the limit $\alpha \to 0$, in the Frank case) yields the independent copula $C = \Phi G$ as a special case. For more detail on copulae, see, for instance, Nelson (2006). For many alternative ways to construct continuous bivariate distributions, see Balakrishnan and Lai (2009).

Discrete Bivariate Distributions

Let (X,Y) denote two random variables defined on a domain of support $\Lambda \subset \mathbb{R}^2$. Then a function $f : \Lambda \to \mathbb{R}_+$ is a joint pmf (*probability mass function*) if it has the following properties:

$$\begin{aligned} f(x,y) &= P(X = x, Y = y) > 0, \text{ for } (x,y) \in \Lambda \\ \sum_{\Lambda}\sum f(x,y) &= 1 \\ P((X,Y) \in S) &= \sum_{S}\sum f(x,y), \text{ for any } S \subset \Lambda \end{aligned} \tag{9}$$

The joint cdf is:

$$F(x,y) = P(X \le x, Y \le y) = \sum_{v \le x}\sum_{w \le y} f(v,w) \tag{10}$$

The *marginal pmf* of X, denoted $f_x(x)$, is:

$$f_x(x) = \sum_y f(x,y) \tag{11}$$

and similarly, the *marginal pmf* of Y, denoted $f_y(y)$, is:

$$f_y(y) = \sum_x f(x,y) \tag{12}$$

The *conditional pmf* of Y given $X = x$ is denoted by $f(y|X=x)$ or, for short, $f(y|x)$. It is defined by

$$f(y|x) = \frac{f(x,y)}{f_x(x)}, \quad \text{provided } f_x(x) > 0 \tag{13}$$

For extensive detail on discrete bivariate distributions, see Kocherlakota and Kocherlakota (1992).

Given discrete random variables defined on subsets of the non-negative integers $\{0,1,2,\ldots\}$, and $\vec{t} = (t_1, t_2) \in \mathbb{R}^2$, the bivariate *probability generating function* (pgf) is:

$$\Pi\left(\vec{t}\right) = E\left[t_1^X t_2^Y\right] \tag{14}$$

The pgf provides a way to determine the probabilities:

$$P(X = r, Y = s) = \frac{1}{r!s!} \frac{\partial^{r+s}\Pi\left(\vec{t}\right)}{\partial t_1^r \partial t_2^s}\Big|_{\vec{t}=\vec{0}} \tag{15}$$

Example 2 Joint pmf

Let random variables X and Y have joint pmf $f(x,y) = \frac{x+1-y}{54}$ with domain of support $\Lambda = \{(x,y) : x \in \{3,5,7\}, y \in \{0,1,2,3\}\}$, as per Table 3.

This is a well-defined pmf since all the probabilities are positive, and they sum to 1. Figure 3 plots the joint pmf and the joint cdf. For computational details, see Rose and Smith (2002).

Example 3 A bivariate Poisson

Let Z_0, Z_1 and Z_2 be mutually stochastically independent univariate Poisson random variables, with non-negative parameters λ_0, λ_1 and λ_2, respectively, and pmf's $g_i(z_i)$ for $i \in \{0,1,2\}$:

$$g_i(z_i) = \frac{e^{-\lambda_i}\lambda_i^{z_i}}{z_i!} \text{ defined on } z_i \in \{0,1,2,\ldots\} \tag{16}$$

Due to independence, the joint pmf of (Z_0, Z_1, Z_2) is $g_0(z_0)g_1(z_1)g_2(z_2)$. Then, a non-trivial *bivariate Poisson* distribution is obtained as the joint distribution of X and Y where:

$$X = Z_1 + Z_0 \text{ and } Y = Z_2 + Z_0 \tag{17}$$

The joint pmf of X and Y can be found via the method of transformations (see, for instance, Rose and Smith (2002,

Bivariate Distributions. Table 2 Some examples of copulae

Copula	Formula	Restrictions
Independent	$C = \Phi G$	
Morgenstern	$C = \Phi G(1 + \alpha(1-\Phi)(1-G))$	$-1 \le \alpha \le 1$
Ali–Mikhail–Haq	$C = \dfrac{\Phi G}{1-\alpha(1-\Phi)(1-G)}$	$-1 \le \alpha < 1$
Frank	$C = -\dfrac{1}{\alpha}\log\left[1 + \dfrac{\left(e^{-\alpha\Phi}-1\right)\left(e^{-\alpha G}-1\right)}{e^{-\alpha}-1}\right]$	$\alpha \neq 0$

Bivariate Distributions. Table 3 Joint pmf of $h(x,y) = \dfrac{x+1-y}{54}$

	$Y=0$	$Y=1$	$Y=2$	$Y=3$
$X=3$	$\frac{4}{54}$	$\frac{3}{54}$	$\frac{2}{54}$	$\frac{1}{54}$
$X=5$	$\frac{6}{54}$	$\frac{5}{54}$	$\frac{4}{54}$	$\frac{3}{54}$
$X=7$	$\frac{8}{54}$	$\frac{7}{54}$	$\frac{6}{54}$	$\frac{5}{54}$

p. 244)). Doing so yields the bivariate Poisson pmf as:

$$\begin{aligned} f(x,y) &= P(X=x, Y=y) \\ &= e^{(-\lambda_0-\lambda_1-\lambda_2)} \sum_{i=0}^{x} \frac{\lambda_0^i \lambda_1^{x-i} \lambda_2^{y-i}}{i!(x-i)!(y-i)!} \\ &= e^{(-\lambda_0-\lambda_1-\lambda_2)} \lambda_0^x (-\lambda_2)^{-x} \lambda_2^y \frac{U\left(-x, 1-x+y, -\frac{\lambda_1\lambda_2}{\lambda_0}\right)}{x!y!} \end{aligned} \tag{18}$$

with domain of support $\{(x,y) : x \in \{0,1,2,\ldots\}, y \in \{0,1,2,\ldots\}\}$, and where $U(a,b,c)$ denotes the confluent hypergeometric function.

Product Moments

The bivariate *raw moment* $\acute{\mu}_{r,s}$ is defined as:

$$\acute{\mu}_{r,s} = E\left[X^r Y^s\right]. \tag{19}$$

With $s = 0$, $\acute{\mu}_{r,0}$ denotes the r^{th} raw moment of X. Similarly, with $r = 0$, $\acute{\mu}_{0,s}$ denotes the s^{th} raw moment of Y. More generally, $\acute{\mu}_{r,s}$ is known as a *product* raw moment or joint raw moment.

The bivariate *central moment* $\mu_{r,s}$ is defined as

$$\mu_{r,s} = E[(X - E[X])^r (Y - E[Y])^s] \tag{20}$$

The *covariance* of X and Y, denoted $\mathrm{Cov}(X,Y)$, is defined as $\mu_{1,1}$, namely:

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])] \tag{21}$$

The *correlation* between X and Y is defined as

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \tag{22}$$

where it can be shown, by the Cauchy–Schwarz inequality, that $-1 \le \rho \le 1$.

Let $\vec{t} = (t_1, t_2) \in \mathbb{R}^2$ denote two dummy variables. Then the bivariate ▶*moment generating function* (mgf) $M_{X,Y}\left(\vec{t}\right)$ is a function of $\vec{t}$, defined by

$$M\left(\vec{t}\right) = E[\exp(t_1 X + t_2 Y)] \tag{23}$$

provided the expectation exists for all $t_i \in (-c, c)$, for some constant $c > 0$, $i = 1, 2$. If it exists, the mgf $M\left(\vec{t}\right)$ can be used to generate the product raw moments $\acute{\mu}_{r,s} = E\left[X^r Y^s\right]$ as follows:

$$\acute{\mu}_{r,s} = E\left[X^r Y^s\right] = \frac{\partial^{r+s} M\left(\vec{t}\right)}{\partial t_1^r \partial t_2^s}\Big|_{\vec{t}=\vec{0}} \tag{24}$$

The *cumulant generating function* is the natural logarithm of the mgf. The bivariate *characteristic function* is similar to (23) and given by

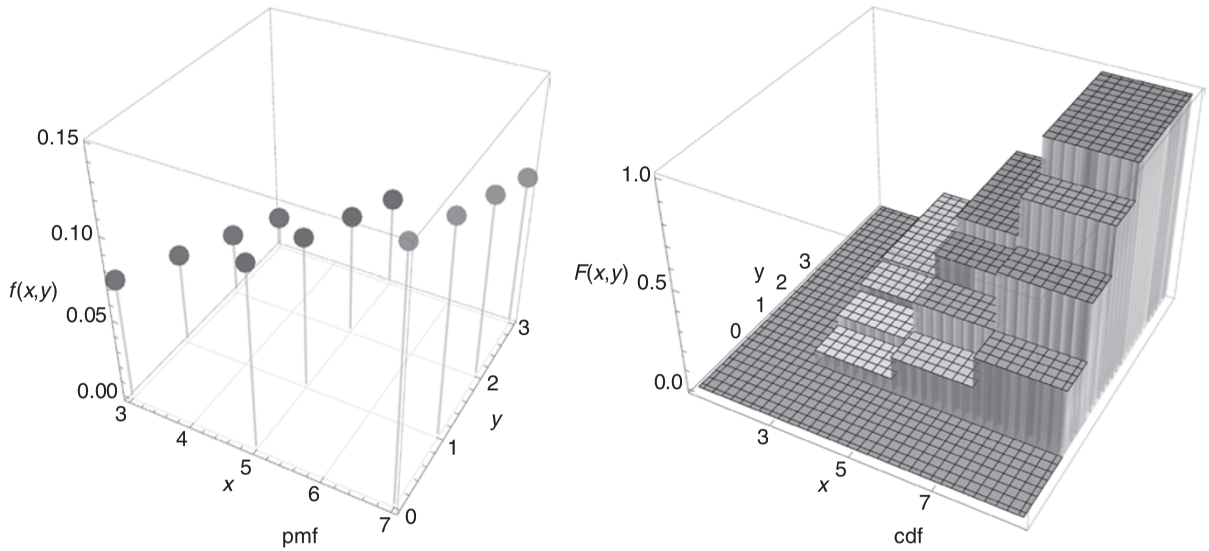
$$C\left(\vec{t}\right) = E[\exp(i\,(t_1 X + t_2 Y))] \tag{25}$$

where i denotes the unit imaginary number.

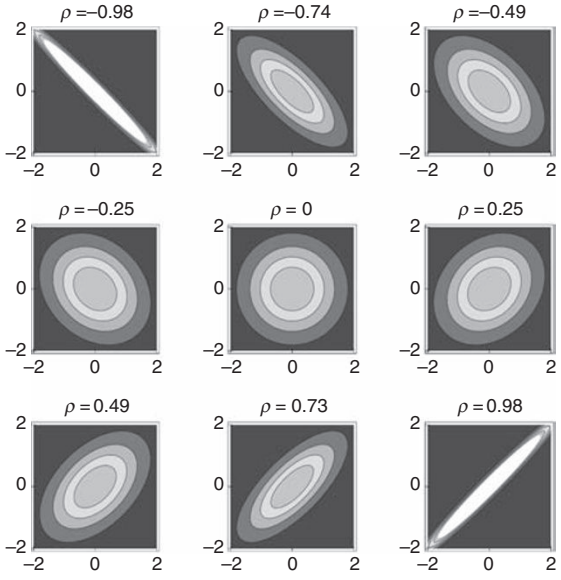
Dependence

Let random variables X and Y have joint pdf $f(x,y)$, with marginal density functions $f_x(x)$ and $f_y(y)$. Then X and Y are said to be *mutually stochastically independent* if and only if

$$f(x,y) = f_x(x) f_y(y) \tag{26}$$



Bivariate Distributions. Fig. 3 Joint pmf of $h(x, y) = \frac{x+1-y}{54}$ of Example 2 (*left*) and the joint cdf (*right*)



Bivariate Distributions. Fig. 4 Contour plots of the bivariate Normal pdf, for different values of ρ

i.e., if and only if the joint pdf is equal to the product of the marginal pdf's. If X and Y are mutually stochastically independent, then, amongst other properties, $\mathrm{Cov}(X, Y) = 0$. Independence implies zero covariance, but the converse is not true: that is, zero covariance does *not* imply that X and Y are independent.

Figure 4 illustrates contour plots for a bivariate Normal pdf with zero means and variance–covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Here, ρ denotes the correlation coefficient between X and Y. Each plot corresponds to a specific value of ρ. In the top left corner, $\rho = -0.98$ (almost perfect negative correlation), whereas in the bottom right corner, $\rho = 0.98$ (almost perfect positive correlation). The middle plot corresponds to the case of zero correlation. In any given plot, the edge of each shaded region represents the contour line, and each contour is a two-dimensional ellipse along which the bivariate Normal pdf $f(x, y)$ is constant.

Balakrishnan and Lai (2009, Chaps. 3 and 4) provide detail on alternative and more sophisticated concepts of dependence.

About the Author

Dr. Colin Rose is director of the Theoretical Research Institute (Sydney). He is the co-author (with M.D. Smith) of the Springer text *Mathematical Statistics with Mathematica*. His current area of research is exact computational methods in mathematical statistics, in particular, with application to the mathStatica project (winner of the Best Software Contribution at CompStat Berlin). He is an Editor of the *Journal of Statistical Software*, and a recent guest editor of the *Mathematica Journal*.

Cross References

►Copulas
►Multivariate Statistical Distributions
►Statistical Distributions: An Overview
►Univariate Discrete Distributions: An Overview

References and Further Reading

Balakrishnan N, Lai CD (2009) Continuous bivariate distributions, 2nd edn. Springer, New York
Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Marcel Dekker, New York
Kocherlakota S, Kocherlakota K (1992) Bivariate discrete distributions. Wiley, New York
Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York
Rose C, Smith MD (2002) Mathematical statistics with Mathematica. Springer, New York

Bootstrap Asymptotics

Rudolf Beran
Distinguished Professor Emeritus
University of California-Davis, Davis, CA, USA

The bootstrap (see ►Bootstrap Methods), introduced by Efron (1979), merges simulation with formal model-based statistical inference. A statistical model for a sample X_n of size n is a family of distributions $\{P_{\theta,n}: \theta \in \Theta\}$. The parameter space Θ is typically metric, possibly infinite-dimensional. The value of θ that identifies the true distribution from which X_n is drawn is unknown. Suppose that $\hat{\theta}_n = \hat{\theta}_n(X_n)$ is a consistent estimator of θ. The bootstrap idea is

(a) Create an artificial *bootstrap world* in which the true parameter value is $\hat{\theta}_n$ and the sample X_n^* is generated from the fitted model $P_{\hat{\theta}_n,n}$. That is, the conditional distribution of X_n^*, given the data X_n, is $P_{\hat{\theta}_n,n}$.

(b) Act as if a sampling distribution computed in the fully known bootstrap world is a trustworthy approximation to the corresponding, but unknown, sampling distribution in the model world.

For example, consider constructing a confidence set for a parametric function $\tau(\theta)$, whose range is the set T. As in the classical pivotal method, let $R_n(X_n, \tau(\theta))$ be a specified *root*, a real-valued function of the sample and $\tau(\theta)$. Let $H_n(\theta)$ be the sampling distribution of the root under the model. The *bootstrap distribution* of the root is $H_n(\hat{\theta}_n)$, a random probability measure that can also be viewed as the conditional distribution of $R_n\left(X_n^*, \tau(\hat{\theta}_n)\right)$ given the sample X_n. An associated *bootstrap confidence set* for $\tau(\theta)$, of nominal coverage probability β, is then $C_{n,B} = \left\{t \in T: R_n(X_n, t) \le H_n^{-1}(\beta, \hat{\theta}_n)\right\}$. The quantile on the right can be approximated, for instance, by Monte Carlo techniques. The intuitive expectation is that the coverage probability of $C_{n,B}$ will be close to β whenever $\hat{\theta}_n$ is close to θ.

When does the bootstrap approach work? Bootstrap samples are perturbations of the data from which they are generated. If the goal is to probe how a statistical procedure performs on data sets similar to the one at hand, then repeating the statistical procedure on bootstrap samples stands to be instructive. An exploratory rationale for the bootstrap appeals intellectually when empirically supported probability models for the data are lacking. Indeed, the literature on "statistical inference" continues to struggle with an uncritical tendency to view data as a *random* sample from a statistical model *known* to the statistician apart from parameter values. In discussing the history of probability theory, Doob (1972) described the mysterious interplay between probability models and physical phenomena: "But deeper and subtler investigations had to await until the blessing and curse of direct physical significance had been replaced by the bleak reliability of abstract mathematics."

Efron (1979) and most of the subsequent bootstrap literature postulate that the statistical model $\{P_{\theta,n}: \theta \in \Theta\}$ for the data is credible. "The bootstrap works" is taken to mean that bootstrap distributions, and interesting functionals thereof, converge in probability to the correct limits as sample size n increases. The convergence is typically established pointwise for each value of θ in the parameter space Θ. A template argument: Suppose that Θ is metric and that (a) $\hat{\theta}_n \to \theta$ in $P_{\theta,n}$-probability as $n \to \infty$; (b) for any sequence $\{\theta_n \in \Theta\}$ that converges to θ, $H_n(\theta_n) \Rightarrow H(\theta)$. Then $H_n(\hat{\theta}_n) \Rightarrow H(\theta)$ in $P_{\theta,n}$-probability. Moreover, any weakly continuous functional of the bootstrap distribution converges in probability to the value of that functional at the limit distribution.

Such equicontinuity reasoning, in various formulations, is widespread in the literature on bootstrap convergence. For statistical models of practical interest, considerable insight may be needed to devise a metric on Θ such that the template sufficient conditions both hold. Some early papers on bootstrap convergence after Efron (1979) are Bickel and Freedman (1981), Hall (1986), Beran (1987). Broader references are the books and monographs by Hall (1992), Mammen (1992), Efron and Tibshirani (1993), Davison and Hinkley (1997) and the review articles in the bootstrap issue of *Statistical Science* **18** (2003).

These references leave the impression that bootstrap methods often work, in the sense of correct pointwise asymptotic convergence or pointwise second-order accuracy, at every θ in the parameter space Θ. Counterexamples to this impression have prompted further investigations. One line of research has established necessary and sufficient conditions for correct pointwise convergence of bootstrap distributions as n tends to infinity [cf. Beran (1997), van Zwet and Zwet (1999)].

In another direction, Putter (1994) showed: Suppose that the parameter space Θ is complete metric and that (a) $H_n(\theta) \Rightarrow H(\theta)$ for every $\theta \in \Theta$ as $n \to \infty$; (b) $H_n(\theta)$ is continuous in θ, in the topology of weak convergence, for every $n \ge 1$; (c) $\hat{\theta}_n \to \theta$ in $P_{\theta,n}$-probability for every $\theta \in \Theta$ as $n \to \infty$. Then $H_n(\hat{\theta}_n) \Rightarrow H(\theta)$ in $P_{\theta,n}$-probability for "almost all" $\theta \in \Theta$. The technical definition of "almost all" is a set of Baire category II. While "almost all" θ may sound harmless, the failure of bootstrap convergence on a tiny set in the parameter space typically stems from non-uniform convergence of bootstrap distributions over neighborhoods of that set. When that is the case, pointwise limits are highly deceptive.

To see this concretely, let $\hat{\theta}_{n,S}$ denote the ▶James-Stein estimator for an unknown p-dimensional mean vector θ on which we have n i.i.d. observations, each having a $N(0, I_p)$ error. Let $H_n(\theta)$ be the sampling distribution of the root $n^{1/2}(\hat{\theta}_{n,S} - \theta)$ under this model. As n tends to infinity with $p \ge 3$ fixed, we find (cf. Beran (1997)):

(a) The natural bootstrap distribution $H_n(\bar{X}_n)$, where $\bar{X}_n$ is the sample mean vector, converges correctly almost everywhere on the parameter space, except at $\theta = 0$. A similar failure occurs for the bootstrap distribution $H_n(\hat{\theta}_{n,S})$.

(b) The weak convergences of the sampling distribution $H_n(\theta)$ and of the two bootstrap distributions just described are *not* uniform over neighborhoods of the point of bootstrap failure, $\theta = 0$.

(c) The exact quadratic risk of the James-Stein estimator strictly dominates that of $\bar{X}_n$ at *every* θ, especially at $\theta = 0$. If the dimension p is held fixed, the region of substantial dominance in risk shrinks towards $\theta = 0$ as n increases. The asymptotic risk of the James-Stein estimator dominates that of the sample mean only at $\theta = 0$. That the dominance is strict for every finite $n \geq 1$ is missed by the non-uniform limit. Apt in describing non-uniform limits is George Berkeley's celebrated comment on infinitesimals: "ghosts of departed quantities."

In the James-Stein example, correct pointwise convergence of bootstrap distributions as n tends to infinity is an inadequate "bootstrap works" concept, doomed by lack of uniform convergence. The example provides a leading instance of an estimator that dominates classical counterparts in risk and fails to bootstrap naively. The message extends farther. Stein (1956, first section) already noted that multiple shrinkage estimators, which apply different shrinkage factors to the summands in a projective decomposition of the mean vector, are "better for most practical purposes." Stein (1966) developed multiple shrinkage estimators in detail. In recent years, low risk multiple shrinkage estimators have been constructed implicitly through regularization techniques, among them adaptive penalized least squares with quadratic penalties, adaptive submodel selection, or adaptive symmetric linear estimators. Naive bootstrapping of such modern estimators fails as it does in the James-Stein case.

Research into these difficulties has taken two paths: (a) devising bootstrap patches that fix *pointwise* convergence of bootstrap distributions as the number of replications n tends to infinity [cf. Beran (1997) for examples and references to the literature]; (b) studying bootstrap procedures under asymptotics in which the dimension p of the parameter space increases while n is held fixed or increases. Large p bootstrap asymptotics turn out to be uniform over usefully large subsets of the parameter space and yield effective bootstrap confidence sets around the ►James-Stein estimator and other regularization estimators [cf. Beran (1995), Beran and Dümbgen (1998)]. The first section of Stein (1956) foreshadowed the role of large p asymptotics in studies of modern estimators.

About the Author

Rudolf Beran was Department Chair at UC Davis (2003–2007) and at UC Berkeley (1986–1989). He received in 2006 the Memorial Medal of the Faculty of Mathematics and Physics, Charles University, Prague, in recognition of "distinguished and wide-ranging achievements in mathematical statistics, ... devoted service to the international statistical community, and a long-lasting collaboration with Czech statisticians." During 1997–1999 he held an Alexander von Humboldt U.S. Senior Scientist Award at Heidelberg University. He has authored or co-authored over 100 papers in international journals and published lecture notes (with G. R. Ducharme) on *Asymptotic Theory for Bootstrap Methods in Statistics* (Publications CRM, Université de Montréal, 1991).

Cross References

►Bootstrap Methods

►Functional Derivatives in Statistics: Asymptotics and Robustness

References and Further Reading

Beran R (1987) Prepivoting to reduce level error of confidence sets. Biometrika 74:457–468

Beran R (1995) Stein confidence sets and the bootstrap. Statistica Sinica 5:109–127

Beran R (1997) Diagnosing bootstrap success. Ann Inst Stat Math 49:1–24

Beran R, Dümbgen L (1998) Modulation of estimators and confidence sets. Ann Stat 26:1826–1856

Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. Ann Stat 9:1196–1217

Davison AC, Hinkley DV (1997) Bootstrap methods and their application. Cambridge University Press, Cambridge

Doob JL (1972) William Feller and twentieth century probability. In: Le Cam LM, Neyman J, Scott EL (eds) Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability. II, University of California Press, Berkeley and Los Angeles, pp xv–xx

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York

Hall P (1986) On the bootstrap and confidence intervals. Ann Stat 14:1431–1452

Hall P (1992) The bootstrap and Edgeworth expansion. Springer, New York

Mammen E (1992) When does bootstrap work? Lecture Notes in Statistics 77. Springer, New York

Putter H (1994) Consistency of resampling methods. PhD dissertation, Leiden University, Leiden

Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Neyman J (ed) Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability I, University of California Press, Berkeley and Los Angeles, pp 197–206

Stein C (1966) An approach to the recovery of inter-block information in balanced incomplete block designs. In: David FN (ed) Festschrift for Jerzy Neyman. Wiley, New York, pp 351–364

van Zwet EW, Zwet WR (1999) A remark on consistent estimation. Math Method Stat 8:277–284

Bootstrap Methods

Michael R. Chernick[1], Wenceslao González-Manteiga[2], Rosa M. Crujeiras[2], Erniel B. Barrios[3]
[1]Lankenau Institute for Medical Research, Wynnewood, PA, USA
[2]University of Santiago de Compostela, Santiago de Compostela, Spain
[3]Professor and Dean
University of the Philippines, Quezon City, Philippines

Introduction

Use of the bootstrap idea goes back at least to Simon (1969) who used it as a tool to teach statistics. But the properties of the bootstrap and its connection to the ▶jackknife and other resampling methods, was not realized until Efron (1979). Similar resampling methods such as the jackknife and subsampling go back to the late 1940s and 1960s respectively (Quenouille (1949) for the jackknife and Hartigan (1969) and McCarthy (1969) for subsampling). In 1979 the impact that the bootstrap would have was not really appreciated and the motivation for Efron's paper was to better understand the jackknife and its properties. But over the past 30 years it has had a major impact on both theoretical and applied statistics with the applications sometimes leading the theory and vice versa. The impact of Efron's work has been so great that he was awarded with the President's Medal of Science by former President George W. Bush and Kotz and Johnson (1992) included the 1979 Annals of Statistics paper in their three volume work on breakthroughs in statistics. After the publication of Efron's paper, Simon's interest in bootstrapping was revitalized and he and Peter Bruce formed the company Resampling Stats which publicized the methodology and provided elementary software for teaching and basic data analysis (see Simon and Bruce (1991)). The bootstrap is not simply another statistical technique but is rather a general approach to statistical inference with very broad applicability and very mild modeling assumptions.

There are now a number of excellent books that specialize in bootstrap or resampling in general. Included in this list are Efron (1982), Efron and Tibshirani (1993), Davison and Hinkley (1997), Chernick (1999, 2007), Hall (1992), Manly (1997), Lunneborg (2000), Politis et al. (1999), Shao and Tu (1995) and Good (1998, 2004). Many other texts devote chapters to the bootstrap including a few introductory statistics texts. There are even a couple of books that cover subcategories of bootstrapping (Westfall and Young (1993) and Lahiri (2003b)).

The Basic Idea

Formally, denote by $\mathbf{X} = (X_1, \ldots, X_n)$ a random sample from X with unknown distribution F and consider a random variable $T = T(X_1, \ldots, X_n; F)$ which may be as simple as $T = \overline{X} - \mu$, with $\mu = \int x dF(x)$, or a more complicated one such as a nonparametric kernel density estimator of the density f given by $T = \hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$, where h denotes the bandwidth parameter and K is a kernel function. The main goal in statistical inference is to determine the sampling distribution of T, namely $\mathbb{P}_F(T(\mathbf{X}, F) \leq x)$. If F_n denotes the empirical distribution of X, from the sample $\mathbf{X}$, then the bootstrap version of T is given by $T^* = T\left(X_1^*, \ldots, X_n^*; F_n\right)$ where $\mathbf{X}^* = \left(X_1^*, \ldots, X_n^*\right)$ is a random sample from F_n. This is known as the naive bootstrap. We may also replace F by a smoothed version $\tilde{F}_n(x) = \int_{-\infty}^{x} \hat{f}_h(t)dt$ (called the smooth or generalized bootstrap) or by a parametric estimate of F say $F_{\hat{\theta}_n}$ (parametric bootstrap). See Efron (1979, 1982) as introductory references, and Silverman and Young (1987) and Dudewicz (1992) for detailed coverage on the smooth bootstrap.

As it has been said, the main goal is to estimate the sampling distribution of T. The bootstrap estimator of $\mathbb{P}_F(T(\mathbf{X}, F) \leq x)$ is given by $\mathbb{P}_{F_n}(T(\mathbf{X}^*, F_n) \leq x) = \mathbb{P}^*(T(\mathbf{X}^*, F_n) \leq x)$, where $\mathbb{P}^*$ is the associated probability in the bootstrap world (the distribution associated with sampling with replacement from F_n in the naive bootstrap case). Since in most cases this probability cannot be computed, Monte Carlo methods are used in order to obtain an approximation, based on B bootstrap replicates $\mathbf{X}^{*j}$, with $j = 1, \ldots B$. The estimate of $\mathbb{P}^*(T(\mathbf{X}^*, F_n) \leq x)$ is just

$$\widehat{\mathbb{P}^*}(T(\mathbf{X}^*, F_n) \leq x) = \frac{\#\{j; T(\mathbf{X}^{*j}, F_n) \leq x\}}{B},$$

where # denotes the cardinal of the set. Once we have approximated the statistic's distribution, other inference problems such as bias and variance estimation, confidence interval construction or hypothesis testing, etc. can be tackled. With confidence intervals, the variable of interest is usually given by $T = \hat{\theta} - \theta$, where θ is an unknown parameter of the distribution and $\hat{\theta}$ is the corresponding estimator. The simplest case is based on the direct approximation of the distribution of T, known as the percentile method. Several refinements such as bias-corrected percentile, percentile-t (also called bootstrap-t) or other corrections based on Edgeworth expansions (see ▶Edgeworth Expansion), have been proposed. For more complete coverage of bootstrap confidence intervals, see Hall (1988, 1992), Efron and Tibshirani (1993), Chernick (2007) or Davison and Hinkley (1997).

Generally speaking, the bootstrap methodology aims to reproduce from the sample the mechanism generating the data which may be a probability distribution, a regression model, a time series, etc. Nowadays, bootstrap methods have been applied to solve different inference problems, including bandwidth selection in curve estimation (Cao 1993), distribution calibration in ▶empirical processes or empirical regression processes (Stute et al. 1998) or inference for incomplete data detailed below, among others.

Extension to Dependent Data

For the sake of simplicity, we may distinguish two perspectives. First, data involved in the statistic may be directly observed, where the basic bootstrap resampling procedures introduced above can be applied. Secondly, data may exhibit a complex generating mechanism. As a special case, consider a parametric regression model

$$y_i = m_\theta(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $m_\theta(\cdot)$ is the regression function and ε_i denotes the associated ith error. For fixed design and parametric regression function, we may proceed by resampling the residuals $e_i = y_i - m_{\hat{\theta}}(x_i)$, where $\hat{\theta}$ is a parameter estimator. Naive bootstrap samples $\tilde{e}_i^*$ drawn from the empirical distribution of the centered residuals $\{e_i - \bar{e}\}$ are used to get the bootstrap regression model $y_i^* = m_{\hat{\theta}}(x_i) + \tilde{e}_i^*$. This approach is called model-based bootstrapping.

Efron also introduced the bootstrap in this context. Each of the bootstrap samples can provide an estimate of the regression parameter (possibly a vector of parameters) following the same estimation procedure that was used with the original fitted model (e.g., ordinary ▶least squares). From all the bootstrap replicates, we get a Monte Carlo approximation to the bootstrap distribution of the regression parameter and this is then used to make inferences about the parameter based on this approximation to the sampling distribution for the parameter estimate. In model-based inference, Paparoditis and Politis (2005) underscored the importance of the choice of residuals. For example, to maximize power in bootstrap-based hypothesis testing, residuals are obtained using a sequence of parameter estimators that converge to the true parameter value under both the null and alternative hypotheses.

Different modifications of this simple idea allow for adapting to random design, heterocedastic models or situations where the regression function is not totally specified or is unknown, such as in nonparametric regression. From the first references, specially in parametric regression, by Bickel and Freedman (1981) and Freedman (1981), several advances have been introduced in this context. For the nonparametric case, see Chap. 14 in Schimek (2000).

Although bootstrap originally started with independent sample observations, just as in the regression models described above, there are extensions to dependent data: ▶time series, ▶spatial statistics, ▶point processes, spatio-temporal models and more. For example, similar to the ideas of bootstrap in regression models, given an explicit dependence structure such an autoregressive model, we may write:

$$y_i = m(y_{i-1}, \ldots, y_{i-p}) + \varepsilon_i$$

and proceed by resampling from the residuals. When an explicit parametric equation is not available, an alternative is block bootstrap, which consists of resampling blocks of subsamples, trying to capture the dependence in the data. Bootstrap replicates obtained by these methods may not be stationary even though the original data came from a stationary process. In order to solve this problem, Politis and Romano (1994a) propose the stationary bootstrap. These bootstrap procedures can be adapted for predicting the future values of the process. An overview of bootstrap methods for estimation and prediction in time series can be found in Cao (1999). The idea of block bootstrap has also been extended to the spatial setting (see Lahiri 2003b)

There is theoretical justification for bootstrapping time series. As an example, Politis and Romano (1994b) established convergence of certain sums of stationary time series that can facilitate bootstrap resampling. The block bootstrap was among the early proposals for time series data. While the method is quite straightforward, there are associated problems like getting independence between blocks while maintaining the dependence structure within the block. The size of the block is a crucial quantity that should be determined to assure success in block bootstrapping. The AR-sieve method was also introduced as a residual-based method similar to the model-based approach. Local bootstrap was also introduced but in the context of a local regression framework (nonparametric) and to account for the nonparametric model, resampling allows the empirical distribution to vary locally in the time series.

Bühlman (2002) compared different methods for time series bootstrapping. The block bootstrap is recognized as the most general and simplest generalization of the original independent resamples but is sometimes criticized for the possible artifacts it may exhibit when blocks are linked together. Blocking can potentially introduce some dependence structure in addition to those naturally existing in the data. The AR-sieve is less sensitive to selection of a

model than the block to the block length. The local bootstrap for ▶nonparametric estimation is observed to yield slower rates of convergence. Generally, the AR-sieve is relatively advantageous among the bootstrap approaches for time series data.

Recently, the bootstrap has been introduced in more complex and complicated models. In modeling non-stationary volatility, Xu (2008) used an autoregression around a polynomial trend with stable autoregressive roots to illustrate how non-stationary volatility affects the consistency, convergence rates and asymptotic distributions of the estimators. Westerlund and Edgerton (2007) proposed a bootstrap test for the null hypothesis of cointegration in ▶panel data. Dumanjug et al. (2009) developed a block bootstrap method in a spatial-temporal model.

Diversity of Applications

The use of bootstrap as a tool for calibrating the distribution of a statistic has been extended to most topics in statistical inference. Not trying to be exhaustive, it is worth it considering the immersion of bootstrap in the following fields of study:

1. Incomplete data. When dealing with censored data, the empirical estimator of the distribution is replaced by the ▶Kaplan-Meier estimator (Efron 1981).
2. Missing information. If some observations are missing or imputed, bootstrap estimators must be suitably adapted (Efron 1994).
3. Hypothesis testing in regression models. When the goal is to check whether a parametric regression model m_θ fits the data, the distribution of a test statistic $T = D(\hat{m}_h, m_{\hat{\theta}})$, where $\hat{m}_h$ is a nonparametric estimator of the regression function m and D is a distance measure, must be calibrated. In this context, a broad literature can be cited, such as Härdle and Mammen (1993) or Cao and González-Manteiga (1993). For a recent review on the topic, see González-Manteiga and Crujeiras (2008).
4. Small area inference. Bootstrap has also shown a great development in finite populations (Shao and Tu 1995), specially in recent years with the appearance of small area models. See Hall and Maiti (2006) and Lahiri (2003a).
5. Bootstrap has also recently been used in learning theory and high dimensional data. As an example, see the application of bootstrap in the regression model with functional data (Ferraty et al. 2009) or the bootstrap for variable choice in regression or classification models (Hall et al. 2009).

The continuing development of bootstrap methods has been motivated by the increasing progress in computational efficiency in recent years. Other computer-intensive methods are the ▶Markov Chain Monte Carlo, usually known as MCMC [see Smith and Roberts (1993), for instance] and subsampling (Politis et al. 1999).

Some Historical Development

The bootstrap's popularity rests in its relaxation of distribution assumptions that can be restrictive and its wide variety of applications as described above. We see that the development of the bootstrap evolved as follows. Efron introduced it in (1979) with some theoretical and heuristic development. Theoretical activity followed quickly with Athreya, Bickel, Freedman, Singh, Beran and Hall providing notable contributions in the early 1980s. Efron realized its practical value early on and efforts to make the scientific community aware of its potential were the Diaconis and Efron (1983) article in Scientific American and the article by Efron and Tibshirani (1986).

So by the early 1990s enough theory and successful applications had developed to lead to an explosion of papers, mostly applied and some extending the theory. The literature was so large that Chernick (1999) contains more than 1,600 references. An excellent and nearly up-to-date survey article on the bootstrap is Lahiri (2006).

When and Why Bootstrap Can Fail to Work

For the bootstrap to *work*, the bootstrap estimates must be consistent. But even when the first results on consistency of the bootstrap estimate of a mean were derived, Bickel, Freedman and others realized that there were cases where the bootstrap is inconsistent. Two notable examples are (1) the sample mean when the population distribution has an infinite variance but the ordinary sample mean appropriately normalized still converges to a stable law and (2) the maximum of a sample when the population distribution is in the domain of attraction of an extreme value distribution. These examples are covered in Chap. 9 of Chernick (2007).

These results on bootstrap inconsistency were already published in the 1980s and led to a concern about what the real limitations of the bootstrap are. The volume edited by LePage and Billard (1992) and the monograph by Mammen (1992) address these concerns. Consistency is one requirement but what about the small sample properties? This was addressed beautifully using simulation as illustrated in Shao and Tu (1995) and Efron (1983). The work of Efron

and others on small sample accuracy of bootstrap estimates of error rates in classification is summarized in Chap. 2 of Chernick (2007).

Remedies for Failure

Efron's bootstrap principle states that the nonparametric bootstrap mimics sampling from a population by letting the empiric distribution F_n play the role of the unknown population distribution F and letting the bootstrap distribution F_n^* play the role of F_n. This is to say in words what was described using formal mathematics earlier in this article.

Efron thought it was natural to take the size of a bootstrap sample to be n but others saw no reason why a value $m < n$ could not be used. For many problems where consistency was shown $m = n$ works fine. Bickel and Ren (1996) introduced a bootstrap approach called the m-out-of-n bootstrap which takes $m < n$ and works well in some problems. A recent advance in bootstrapping was the proof that for the two examples previously described, where the bootstrap is inconsistent when $m = n$, the m-out-of-n bootstrap is consistent provided m tends to infinity at a slower rate than n and slow enough for $m/n \to 0$. Prior to the formal introduction of the m-out-of-n bootstrap, Athreya (1987) showed that for heavy-tailed distributions a trimmed mean could converge to the population mean with the advantage that the sampling distribution of the trimmed mean has second moments. Using this idea he proved consistency of the m-out-of-n bootstrap for the mean in the infinite variance case. Fukuchi (1994) did the same for extremes. Both results require m and n to approach infinity at a rate that would have $m/n \to 0$. The m-out-of-n bootstrap has been further studied by Bickel et al. (1997) and Politis et al. (1999). Zelterman (1993) found a different way to modify the bootstrap to make it consistent for the extreme values. This is all summarized in Chap. 9 of Chernick (2007).

The theoretical developments from 1995 to the present have been in the area of (1) modifying the bootstrap to fix inconsistency in order to widen its applicability and (2) extending the theory to dependent situations (as previously mentioned). Lahiri (2003a) is the ideal reference for a detailed account of these developments with dependent data.

Problems and Refinements for Bootstrap Confidence Intervals

Bootstrap confidence intervals have been a concern and Efron recognized early on that getting the asymptotic coverage nearly correct in small samples required more sophistication than his simple percentile method bootstrap. So the bias corrected bootstrap was developed to do that. However, in Schenker (1985) the example of variance estimation for a particular chi square population distribution showed that even the BC method had coverage problems in small samples. Efron (1987) introduced the BCa method which remedied the problem discovered by Schenker.

However, in recent years variance estimation for other examples with skewed or heavy-tailed distributions has shown all bootstrap confidence interval methods to be problematic in small samples. A large Monte Carlo investigation, Chernick and LaBudde (2010), compares the coverage of various bootstrap confidence intervals for the variance estimate from a variety of population distributions when sample sizes are small. They also provide an idea of rates of convergence by showing how the coverage improves as the sample size gets large. An interesting surprise is that in some situations for small sample sizes the lower order bootstrap work better than the higher order ones. This is because they are simpler and do not involve estimating biases and acceleration constants which depend on third order moments of the distribution. For the lognormal population they show that at the small sample sizes (20–100) the coverage error is shockingly high for all methods.

About the Authors

Dr. Chernick is the author of *Bootstrap Methods: A Practitioner's Guide* (Wiley 1999), with the second edition title *Bootstrap Methods: A Guide for Practitioners and Researchers* (Wiley 2007). He is also the coauthor of *Introductory Biostatistics for the Health Sciences: Modern Methods including Bootstrap* (Wiley 2002). He is currently Manager of Biostatistical Services at the Lankenau Institute for Medical Research. He is the winner of the Jacob Wolfowitz Prize in 1983 and is a past President of the Southern California Chapter of the American Statistical Association. He has taught at California State University and the University of Southern California. He is elected Fellow of the American Statistical Association (2001).

Wenceslao González-Manteiga is a full-time professor and main researcher at the Department of Statistics and Operations Research (University of Santiago de Compostela, Spain). He received his Ph.D. degree in Mathematics at this University. He was Editor-in-Chief of *TEST* for 6 years and the *International Journal of the Spanish Statistical Society*. Over the years, he has collaborated with researchers in many fields, visiting several universities and has published about 140 papers in journals such as *Technometrics, Chemometrics, Journal of the American Statistical Association, Annals of Statistics* and *Computational*

Statistics and Data Analysis. He has also collaborated in several projects in applied statistics.

Rosa M. Crujeiras is an associate professor and the Department of Statistics and Operations Research (University of Santiago de Compostela, Spain). With a dissertation thesis on spectral methods for spatial data, she obtained her Ph.D. degree in Mathematics at this University in January 2007. She has been a post-doctoral researcher at the Université catholique de Louvain (Belgium). She has published several papers in international journals, such as *Environmetrics, Theory of Probability and its Applications* or *Computational Statistics and Data Analysis*.

Dr. Erniel Barrios is a Professor and Dean, School of Statistics, University of the Philippines Diliman He was a board member of the Philippine Statistical Association. He served as visiting researcher at The Ohio State University (USA) and the Asian Development Bank Institute (Japan). He has authored and co-authored over 50 papers and a forthcoming book on small area estimation. He served as Associate Editor of *Social Science Diliman*, a journal published by University of the Philippines Diliman (2001–2002) and is incoming Editor of *The Philippine Statistician* published by the Philippine Statistical Association.

Cross References

►Bootstrap Asymptotics
►Exact Goodness-of-Fit Tests Based on Sufficiency
►Functional Derivatives in Statistics: Asymptotics and Robustness
►Jackknife
►Markov Chain Monte Carlo
►Monte Carlo Methods in Statistics
►Multiple Imputation
►Statistical Inference for Stochastic Processes
►Statistics: An Overview
►Target Estimation: A New Approach to Parametric Estimation

References and Further Reading

Andrews D (2000) Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. Econometrica 68:399–405

Athreya KB (1987) Bootstrap estimation of the mean in the infinite variance case. Annals of Statistics 15:724–731

Beran R (1997) Diagnosing bootstrap success. Ann Inst Stat Math 49:1–24

Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. Ann Stat 6:1196–1217

Bickel PJ, Gotze F, van Zwet WR (1997) Resampling fewer than n observations: gains, loses, and remedies for losses. Statistica Sinica 7:1–32

Bickel PJ, Ren JJ (1996) The m out of n bootstrap and goodness of fit tests with doubly censored data. In: Huber PJ, Rieter H (eds) Robust statistics, data analysis, and computer-intensive methods. Lecture Notes in statistics, vol. 109. Springer–verlag, New York, pp 35–48

Bühlman P (1997) Sieve bootstrap for time series. Bernoulli 3:123–148

Bühlman P (2002) Bootstrap for time series. Stat Sci 17:52–72

Cao R (1993) Bootstrapping the mean integrated squared error. J Multivariate Anal 45:137–160

Cao R (1999) An overview of bootstrap methods for estimating and predicting in time series. Test 8:95–116

Cao R, González-Manteiga W (1993) Bootstrap methods in ion smoothing. J Nonparametric Stat 2:379–388

Chernick MR (1999) Bootstrap methods: a practitioners guide. Wiley, New York

Chernick MR (2007) Bootstrap methods: a guide for practitioners and researchers, 2nd edn. Wiley, Hoboken

Chernick MR, LaBudde R (2010) Revisiting qualms about bootstrap confidence intervals. Am J Math Manage Sci, To appear

Davison AC, Hinkley DV (1997) Bootstrap methods and their applications. Cambridge University Press, Cambridge

Diaconis P, Efron B (1983) Computer-intensive methods in statistics. Sci Am 248:116–130

Dudewicz EJ (1992) The generalized bootstrap. In: Jockel K-H, Rothe G, Sendler W (eds) Bootstrapping and related techniques, Proceedings Trier FRG. Lecture Notes in Economics and Mathematical Systems, vol. 376. Springer, Berlin, pp 31–37

Dumanjug C, Barrios E, Lansangan J (2010) Bootstrap procedures in a spatial-temporal model. J Stat Comput Sim 80:809–822

Efron B (1979) Bootstrap methods. Another look at the jackniffe. Ann Stat 7:1–26

Efron B (1981) Censored data and the Bootstrap. J Am Stat Assoc 374:312–319

Efron B (1982) The jacknife, the bootstrap and other resampling plans. SIAM, Philadelphia

Efron B (1983) Estimating the error rate of a prediction rule: improvements on cross-validation. J Am Stat Assoc 78:316–331

Efron B (1987) Better bootstrap confidence intervals (with discussion). J Am StatAssoc 82:316–331

Efron B (1994) Missing data, imputation and the bootstrap. J Am Stat Assoc 89:463–479

Efron B (2000) The bootstrap and modern statistics. J Am Stat Assoc 95:1293–1296

Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. Stat Sci 1:54–77

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, London

Ferraty F, Van Keilegom I, Vieu P (2010) On the validity of the bootstrap in nonparametric functional regression. Scand J Stat 37:286–306

Freedman DA (1981) Bootstrapping regression models. Ann Stat 6:1218–1228

Fukuchi JI (1994) Bootstrapping extremes of random variables. PhD dissertation. Iowa State University, Ames

González-Manteiga W, Cao R (1993) Testing the hypothesis of a general linear model using nonparametric regression estimation. Test 2:223–249

González-Manteiga W, Crujeiras RM (2008) A review on goodness-of-fit tests for regression models. Pyrenees Int Workshop Stat Probab Oper Res: SPO 2007:21–59

Good P (1998) Resampling methods: a practical guide to data analysis. Birkhauser, Boston
Good P (2004) Permutation, parametric, and bootstrap tests of hypotheses, 3rd edn. Springer, New York
Hall P (1988) Theoretical comparison of bootstrap confidence intervals. Ann Stat 16:927–953
Hall P (1992) The bootstrap and Edgeworth expansion. Springer, Berlin
Hall P, Maiti T (2006) On parametric bootstrap methods for small area prediction. J R Stat Soc, Series B Stat Method 68:221–238
Hall P, Lee ER, Park BU (2009) Bootstrap-based penalty choice for the Lasso achieving oracle performance. Statistica Sinica 19:449–471
Härdle W, Mammen E (1993) Comparing nonparametric versus parametric regression fits. Ann Stat 21:1926–1947
Hartigan JA (1969) Using subsample values as typical values. J Am Stat Assoc 64:1303–1317
Kotz S, Johnson NL (1992) Breakthroughs in statistics vol. II: methodology and distribution. Springer, New York
Lahiri P (2003a) On the impact of bootstrap in survey sampling and small-area estimation. Stat Sci – Rev J Inst Math Stat 18: 199–210
Lahiri SN (2003b) Resampling methods for dependent data. Springer Series in Statistics, Springer, New York
Lahiri SN (2006) Bootstrap methods: a review. In: Fan J, Koul HL (eds) Frontiers in statistics. Imperial College Press, London, pp 231–265
LePage R, Billard L (eds) (1992) Exploring the limits of bootstrap. Wiley, New York
Lunneborg CE (2000) Data analysis by resampling: concepts and applications. Brooks/Cole, Pacific Grove
Mammen E (1992) When does the bootstrap work? Asymptotic results and simulations. Springer, Heidelberg
Manly BFJ (1997) Randomization, bootstrap and Monte Carlo methods in biology, 2nd edn. Chapman and Hall, London
McCarthy PJ (1969) Pseudo-replication: half-samples. Int Stat Rev 37:239–263
Paparoditis E, Politis D (2005) Bootstrap hypothesis testing in regression models. Stat Probab Lett 74:356–365
Politis DN, Romano JP (1994a) The stationary bootstrap. J Am Stat Assoc 89:1303–1313
Politis DN, Romano J (1994b) Limit theorems for weakly dependent Hilbert Space valued random variables with application to the stationary bootstrap. Statistica Sinica 4:461–476
Politis DN, Romano JP, Wolf M (1999) Subsampling. Springer, New York
Quenouille MH (1949) Approximate tests of correlation in time series. J R Stat Soc B 11:18–84
Schenker N (1985) Qualms about bootstrap confidence intervals. J Am Stat Assoc 80:360–361
Schimek MG (2000) Smoothing and regression. Wiley series in probability and statistics: applied probability and statistics, Wiley, New York
Shao J, Tu DS (1995) The jackknife and bootstrap. Springer Series in Statistics, Springer, New York
Silverman BW, Young GA (1987) The bootstrap: smooth or not to smooth? Biometrika 74:469–479
Simon JL (1969) Basic research methods in social science. Random House, New York
Simon JL, Bruce P (1991) Resampling: a tool for everyday statistical work. Chance 4:22–32
Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. J R Stat Soc B Methodol 55:3–23
Stute W, Gonzlez-Manteiga W, Presedo-Quindimil MA (1998) Bootstrap approximations in model checks for regression. J Am Stat Asoc 93:141–149
Tukey JW (1958) Bias and confidence in not quite large samples (abstract). Ann Math Stat 29:614
Westerland J, Edgerton D (2007) A panel bootstrap cointegration test. Econ Lett 97:185–190
Westfall P, Young SS (1993) Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley, New York
Xu K (2008) Bootstrapping autoregression under nonstationary volatility. Econ J 11:1–26
Zelterman D (1993) A semiparametric bootstrap technique for simulating extreme order statistics. J Am Stat Assoc 88:477–485

Borel–Cantelli Lemma and Its Generalizations

Tapas Kumar Chandra[1], Fejzi Kolaneci[2]
[1]Professor
Indian Statistical Institute, Calcutta, India
[2]Professor
University of New York, Tirana, Albania

The celebrated Borel–Cantelli Lemma is important and useful for proving the ▶laws of large numbers in the strong form. Consider a sequence of random events $\{A_n\}$ on a probability space (Ω, F, P), and we are interested in the question of whether infinitely many random events occur or if possibly only a finite number of them occur.

The upper limit of the sequence $\{A_n\}$ is the random event defined by

$$\{A_n \; i.o.\} = \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \,,$$

which occurs if and only if an infinite number of events A_n occur. This i.o. stands for "infinitely often."

Below we shall use the fact that if $\{A_n\}$ is a sequence of random events, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) \le \sum_{n=1}^{\infty} P(A_n). \quad (*)$$

The Borel–Cantelli Lemma

Lemma 1 *If* $\sum_{n=1}^{\infty} P(A_n) < \infty$, *then* $P(\limsup_{n\to\infty} A_n) = 0$. *If the random events* $A_1, A_2, \cdots, A_n, \cdots$ *are independent and* $\sum_{n=1}^{\infty} P(A_n) = \infty$, *then* $P(\limsup_{n\to\infty} A_n) = 1$.

Intuitively, $P(\limsup_{n\to\infty} A_n)$ is the probability that the random events A_n occur "infinitely often" and will be denoted by $P(A_n\ i.o.)$.

Proof

First part

Note that $\limsup_{n\to\infty} A_n \subset \bigcup_{k=m}^{\infty} A_k$ for each $m \geq 1$. So for each $m \geq 1$.

$$P\left(\limsup_{n\to\infty} A_n\right) \leq P\left(\bigcup_{k=m}^{\infty} A_k\right) \leq \sum_{k=m}^{\infty} P(A_k) \qquad \text{by } (*).$$

Since $\sum_{n=1}^{\infty} P(A_n)$ is convergent, the tails $\sum_{k=m}^{\infty} P(A_k) \to 0$ as $m \to \infty$. Letting $m \to \infty$, we get $P(\limsup_{n\to\infty} A_n) = 0$.

Second part

We show that

$$1 - P(\limsup_{n\to\infty} A_n) = P((\limsup_{n\to\infty} A_n)') = 0,$$

where A' denotes the complement of A. To this end, it is enough to show that

$$P\left(\bigcap_{k=m}^{\infty} A_k'\right) = 0 \text{ for each } m \geq 1,$$

since then by De Morgan's Rule

$$P\left((\limsup_{n\to\infty} A_n)'\right) = P\left(\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k'\right) \leq \sum_{n=1}^{\infty} P\left(\bigcap_{k=n}^{\infty} A_k'\right)$$
$$= 0 \text{ by } (*).$$

Fix such an m. Since $1 - x \leq \exp(-x)$ for each real x, we have for each $j \geq 1$

$$P\left(\bigcap_{k=m}^{\infty} A_k'\right) \leq P\left(\bigcap_{k=m}^{m+j} A_k'\right) = \prod_{k=m}^{m+j}(1 - P(A_k))$$
$$\leq \exp\left(-\sum_{k=m}^{m+j} P(A_k)\right).$$

As $\sum_{n=1}^{\infty} P(A_n)$ diverges, so does $\sum_{k=m}^{\infty} P(A_k)$ which implies that $\sum_{k=m}^{m+j} P(A_k) \to \infty$ as $j \to \infty$. As $\lim_{x\to\infty} \exp(-x) = 0$, we get upon letting $j \to \infty$ that

$$P\left(\bigcap_{k=m}^{\infty} A_k'\right) = 0.$$

□

Generalizations

The first part of the Borel-Cantelli Lemma was generalized in Barndorff-Nielsen (1961).

Lemma 2 *Let $\{A_n\}$ be a sequence of random events satisfying the conditions*

$$\lim_{n\to\infty} P(A_n) = 0 \quad \text{and} \quad \sum_{n=1}^{\infty} P(A_n A_{n+1}') < \infty.$$

Then

$$P(\limsup_{n\to\infty} A_n) = 0.$$

Lemma 2 holds true if the random events $A_n' A_{n+1}$ are substituted with $A_n A_{n+1}'$.

It should be noted that the hypothesis in Lemma 2 is weaker than the hypothesis in Lemma 1.

A further generalization of Lemma 2 was obtained in Stepanov (2006).

Lemma 3 *Let $\{A_n\}$ be a sequence of random events satisfying the condition $\lim_{n\to\infty} P(A_n) = 0$. Assume that there exists $m \geq 0$ such that*

$$\sum_{n=1}^{\infty} P(A_n' A_{n+1}' \cdots A_{n+m-1}' A_{n+m}) < \infty.$$

Then

$$P(\limsup_{n\to\infty} A_n) = 0.$$

Observe that the hypothesis in Lemma 3 when $m \geq 2$ is weaker than the hypothesis in Lemma 2.

Many attempts were made in order to weaken the independence condition in the second part of the Borel-Cantelli Lemma. This condition means mutual independence of random events $A_1, A_2, \cdots, A_n$ for every n. Erdös and Rényi (1959) discovered that the independence condition can be replaced by the weaker condition of pairwise independence of the random events $A_1, A_2, \cdots, A_n$ for every n. Indeed they also proved the result which is the special case '$l = 1$' in the following lemma due to Kochen and Stone (1964):

Lemma 4 *If $\{A_n\}$ is a sequence of random events satisfying the conditions*

$$\sum_{n=1}^{\infty} P(A_n) = \infty \quad \text{and} \quad \liminf_{n\to\infty} \frac{\sum_{i,k=1}^{n} P(A_i \cap A_k)}{\left[\sum_{k=1}^{n} P(A_k)\right]^2} = l,$$

then

$$P(\limsup_{n\to\infty} A_n) \geq \frac{1}{l}.$$

A further extension due to Chandra (2008) is given below.

Lemma 5 *Let* $\{A_n\}$ *be a sequence of random events such that* $\sum_{n=1}^{\infty} P(A_n) = \infty$. *Let*

$$\liminf_{n\to\infty} \frac{\sum_{1\le i\le j\le n}(P(A_i \bigcap A_j) - a_{ij})}{(\sum_{1\le k\le n} P(A_k))^2} = L$$

where $a_{ij} = (c_1 P(A_i) + c_2 P(A_j))P(A_{j-i}) + c_3 P(A_i)P(A_j)$ *for* $1 \le i < j$, $c_1 \ge 0, c_2 \ge 0, c_3 \in \mathcal{R}$ *being constants (L may depend on* c_1, c_2, c_3*). Assume that L is finite. Then* $c + 2L \ge 1$ *and*

$$P\left(\limsup_{n\to\infty} A_n\right) \ge (c + 2L)^{-1}$$

where $c = 2(c_1 + c_2) + c_3$.

As a special case of Lemma 5, we have the following result.

Lemma 6 *Let* $\{A_n\}$ *be a sequence of random events such that* $\sum_{n=1}^{\infty} P(A_n) = \infty$. *If for some constants* $c_1 \ge 0, c_2 \ge 0$, *and* $c_3 \in \mathcal{R}$ *there exists an integer* $N \ge 1$ *such the* $P(A_i \bigcap A_j) \le a_{ij}$ *whenever* $N \le i < j$ *where the* a_{ij} *are as in Lemma 5, then* $c \ge 1$ *and* $P(\limsup_{n\to\infty} A_n) \ge 1/c$, *c being as in Lemma 5.*

Petrov (1995) found conditions that are necessary and sufficient for the equality

$$P(\limsup_{n\to\infty} A_n) = p, \quad \text{where } 0 \le p \le 1,$$

as well as for the inequality

$$P(\limsup_{n\to\infty} A_n) \ge p, \quad \text{where } 0 < p \le 1.$$

For a different type of extensions of the Borel-Cantelli lemma, see Chen (1978) and Serfling (1975).

About the Authors

Professor Chandra's area of specialization spans Statistical Inference, Asymptotic Theory of Statistics, Limit Theorems and Large Deviations. He earned Ph.D. degree in Statistics in 1981 from The Indian Statistical Institute. He has published two books and about 40 papers in international journals. One important contribution is the demonstration of the superiority of the Score Test over a large family of tests including the Likelihood Ratio test and Wald's test under local alternatives, thereby settling an old conjecture of Professor C.R. Rao.

Dr. Fejzi Kolaneci is Professor and Chair of the Department of Mathematics, University of New York Tirana, Albania. He is Past Editor-in-Chief of AJNTS, *Albanian Journal of Natural and Technical Sciences*. He has written several textbooks, including *Probability Theory And Statistics*, *Mathematical Analysis*, and *Differential Equations*. Professor Kolaneci is past Secretary of Natural and Technical Sciences of Albanian Academy of Sciences, and past Dean of the Teaching Faculty of University "Fan S. Noli." He is member of London Mathematical Society, and member of Society for Industrial and Applied Mathematics.

Cross References

►Almost Sure Convergence of Random Variables
►Stochastic Global Optimization
►Strong Approximations in Probability and Statistics

References and Further Reading

Balakrishnan N, Stepanov A (2010) Generalization of Borel-Cantelli lemma. Math Sci 35(1), http://www.appliedprobability.org/content.aspx?Group=tms&Page=tmsabstracts
Barndorff-Nielsen O (1961) On the rate of growth of the partial maxima of a sequence of independent identically distributed random variables. Math Scan 9:383–394
Chandra TK (2008) The Borel–Cantelli lemma under dependence conditions. Stat Probabil Lett 78:390–395
Chen LHY (1978) A short note on the conditional Borel-Cantelli lemma. Ann Probab 8:699–700
Erdös P, Rényi A (1959) On Cantor's series with convergent $\sum 1/q_n$. Ann Univ Sci Budapest Sec Math 2:93–109
Kochen SB, Stone CJ (1964) A note on the Borel-Cantelli lemma, Illinois. J Math 8(2):248–251
Petrov VV (1995) Limit theorems of probability theory. Oxford University Press, Oxford
Serfling RJ (1975) A general Poisson approximation theorem. Ann Prob 3:726–731
Stepanov A (2006) Generalization of Borel-Cantelli lemma. eprint: arXiv:math/0605007v1

Box–Cox Transformation

Takashi Daimon
Hyogo College of Medicine, Hyogo, Japan
Osaka University Hospital, Osaka, Japan

Box and Cox (1964) proposed a family of power transformations in order to improve additivity, normality, and homoscedasticity of observations. The Box–Cox transformation, which was a modification of a family of power transformations introduced by Tukey (1957), was named for their work. For each value of a real or vector valued transformation parameter λ, let $\psi(y, \lambda)$ be a strictly monotone increasing transformation for a positive y in

some interval. Tukey's power transformations takes the following form:

$$\psi(y,\lambda) = \begin{cases} y^{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0. \end{cases}$$

To take account of the discontinuity at $\lambda = 0$ in the above equation, the original form of the Box–Cox transformation takes the following form:

$$\psi^{\mathrm{BC}}(y,\lambda) = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0. \end{cases}$$

They also proposed an extended form of the Box–Cox transformation, "shifted" power transformation which could deal with situations where y is negative but bounded below:

$$\psi(y,\boldsymbol{\lambda}) = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1}-1}{\lambda}, & \lambda_1 \neq 0, \\ \log(y+\lambda_2), & \lambda_1 = 0, \end{cases}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^{\mathrm{T}}$. However, since the range of the distribution is determined by the unknown shift parameter λ_2, the asymptotic results of maximum likelihood theory may not apply. Consequently, there have existed some alternative versions of the Box–Cox transformation which could handle a negative y. For example, Manly (1976) proposed the exponential transformation:

$$\psi(y,\lambda) = \begin{cases} \dfrac{\exp(\lambda y)-1}{\lambda}, & \lambda \neq 0, \\ y, & \lambda = 0. \end{cases}$$

John and Draper (1980) presented the so-called modulus transformation:

$$\psi(y,\boldsymbol{\lambda}) = \begin{cases} \operatorname{sign}(y)\dfrac{(|y|+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0, \\ \operatorname{sign}(y)\log(|y|+1), & \lambda = 0, \end{cases}$$

where

$$\operatorname{sign}(y) = \begin{cases} 1, & y \geq 0, \\ -1, & y < 0. \end{cases}$$

Bickel and Doksum (1981) suggested another modification:

$$\psi(y,\lambda) = \frac{|y|^{\lambda}\operatorname{sign}(y)-1}{\lambda}, \quad \lambda \neq 0.$$

Yeo and Johnson (2000) proposed another power transformation family motivated by the above modified modulus transformation:

$$\psi(y,\lambda) = \begin{cases} \dfrac{(y+1)^{\lambda}-1}{\lambda}, & y \geq 0, \lambda \neq 0, \\ \log(y+1), & y \geq 0, \lambda = 0, \\ -\dfrac{(1-y)^{2-\lambda}-1}{2-\lambda}, & y < 0, \lambda \neq 2, \\ -\log(1-y), & y < 0, \lambda = 2. \end{cases}$$

The main objective in the analysis using Box–Cox transformation is to make inference on the transformation parameter λ. Box and Cox (1964) applied the maximum likelihood as well as Bayesian methods for estimating the transformation parameter, but there have been many approaches to other inferences including hypothesis testing on the transformation parameter (see Sakia 1992 for details, which gave a comprehensive review on the Box–Cox transformation).

The Box–Cox transformation can be applied to a regressor, a combination of regressors, and/or to the response variable in a linear or nonlinear regression. For example let us consider the following linear functional form:

$$\psi^{\mathrm{BC}}(y,\lambda_0) = \beta_0 + \sum_{j=1}^{q}\beta_j\psi^{\mathrm{BC}}(x_j,\lambda_j) + \varepsilon,$$

where $\psi(y,\lambda_0)$ and $\psi(x_j,\lambda_j)$ represent the transformed response variable and explanatory variables, respectively, where λ_j $(j = 0,1,\ldots,p)$ are the transformation parameters, and ε represents the errors. When using such a functional form with the Box–Cox transformation, it is helpful to explore the underlying relationship in cases where the determination of the functional form need not be based on a priori rationale in any research field.

In addition, for example, for nonlinear regressions (See ►Nonlinear Regression) we can consider the following form (see Carroll and Ruppert 1988):

$$\psi^{\mathrm{BC}}(y,\lambda) = \psi^{\mathrm{BC}}(f(x,\beta),\lambda) + \varepsilon,$$

where $f(x,\beta)$ is a functional form (possibly, corresponding to a theoretical model) that has explanatory variables x and is nonlinear with respect to a real or vector valued parameter β. It is noted that the both sides in the above equation have the same Box–Cox transformation. Thus the objective here is to reduce ►heteroscedasticity and autocorrelation of the error structure as well as non-normality of the error (or residual) itself, rather than the determination of the functional form as above mentioned.

The Box–Cox transformation has been widely utilized. However, when using it we note that this transformation seldom fulfills the basic assumptions required for statistical

inference such as linearity, normality and homoscedasticity simultaneously as originally suggested by Box and Cox (1964).

About the Author

Dr. Takashi Daimon was an Assistant Professor of the Division of Drug Evaluation and Informatics, School of Pharmaceutical Sciences, University of Shizuoka. Currently he is the Chief of Data Center and a Specially Appointed Lecturer of the Medical Center for Translational Research at Osaka University Hospital. He is an Elected Member of the International Statistical Institute and an Associate Editor of the *Japanese Journal of Biometrics*.

Cross References

►Preprocessing in Data Mining
►Skewness
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Quality Control

References and Further Reading

Bickel PJ, Doksum KA (1981) An analysis of transformations revisited. J Am Stat Assoc 76:296–311
Box GEP, Cox DR (1964) An analysis of transformations. J Roy Stat Soc, Ser B 26:211–252
Carroll RJ, Ruppert D (1988) Transformations and weighting in regression. Chapman & Hall, London
John JA, Draper NR (1980) An alternative family of transformations. Appl Stat 29:190–197
Manly BF (1976) Exponential data transformation. Statistician 25:37–42
Sakia RM (1992) The Box–Cox transformation technique: a review. Statistician 41:169–178
Tukey JW (1957) The comparative anatomy of transformations. Ann Math Stat 28:602–632
Yeo I-K, Johnson RA (2000) A new family of power transformations to improve normality or symmetry. Biometrika 87(4):954–959

Box–Jenkins Time Series Models

John Boland
Associate Professor
University of South Australia, Adelaide, SA, Australia

Introduction

We are going to examine the Autoregressive Moving Average (ARMA) process for identifying the serial correlation attributes of a stationary time series (see Boland 2008; Box and Jenkins 1970). Another name for the processes that we will undertake is the Box–Jenkins (BJ) Methodology, which describes an iterative process for identifying a model and then using that model for forecasting. The Box–Jenkins methodology comprises four steps:

- Identification of process
- Estimation of parameters
- Verification of model
- Forecasting

Identification of Process

Assume we have a (at least weakly) stationary time series, i.e., no trend, seasonality, and it is homoscedastic (constant variance). Stationarity will be discussed further in section Stationarity. The general form of an ARMA model is

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{1}$$

where $\{X_t\}$ are identically distributed random variables $\sim(0, \sigma_X^2)$ and $\{Z_t\}$ are white noise, i.e., independent and identically distributed (iid) $\sim(0, \sigma_Z^2)$. ϕ_i and θ_j are the coefficients of polynomials satisfying

$$\begin{aligned} \phi(y) &= 1 - \phi_1 y - \cdots - \phi_p y^p \\ \theta(y) &= 1 + \theta_1 y + \cdots + \theta_q y^q, \end{aligned} \tag{2}$$

where $\phi(y), \theta(y)$ are the autoregressive and moving average polynomials respectively. Define the backward shift operator $B^j X_t = X_{t-j}$, $j = 0, 1, 2, \ldots$ and we may then write (2) in the form

$$\phi(B) X_t = \theta(B) Z_t \tag{3}$$

defining an $ARMA(p, q)$ model. If $\phi(B) = 1$, we then have a moving average model of order q, designated $MA(q)$. Alternatively, if we have $\theta(B) = 1$, we have an autoregressive model of order p, designated $AR(p)$. The question is, how do we identify whether we have an $MA(q), AR(p)$ or $ARMA(p, q)$? To do so, we can examine the behavior of the autocorrelation and partial autocorrelation functions.

Autocorrelation and Partial Autocorrelation Functions

We need some definitions to begin with. Suppose two variables X and Y have means μ_X, μ_Y respectively. Then the covariance of X and Y is defined to be

$$Cov(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\}. \tag{4}$$

If X and Y are independent, then

$$\begin{aligned} Cov(X, Y) &= E\{(X - \mu_X)(Y - \mu_Y)\} \\ &= E(X - \mu_X) E(Y - \mu_Y) = 0. \end{aligned} \tag{5}$$

If X and Y are not independent, then the covariance may be positive or negative, depending on whether high

values of X tend to happen coincidentally with high or low values of Y. It is usual to standardise the covariance by dividing by the product of their respective standard deviations, creating the correlation coefficient. If X and Y are random variables for the same stochastic process at different times, then the covariance coefficient is called the autocovariance coefficient, and the correlation coefficient is called the autocorrelation coefficient. If the process is stationary, then the standard deviations of X and Y will be the same, and their product will be the variance of either.

Let $\{X_t\}$ be a stationary time series. The *autocovariance function* (ACVF) of $\{X_t\}$ is $\gamma_X(h) = Cov(X_{t+h}, X_t)$, and the *autocorrelation function* (ACF) of $\{X_t\}$ is

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = Corr(X_{t+h}, X_t). \tag{6}$$

The autocovariance and autocorrelation functions can be estimated from observations of $X_1, X_2, \ldots, X_n$ to give the sample autocovariance function (SAF) and the sample autocorrelation function (SACF), the latter defined by

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}. \tag{7}$$

Thus the SACF is a measure of the linear relationship between time series separated by some time period, denoted by the lag k. Similar to the correlation coefficient of linear regression, r_k will take a value between +1 and –1, and the closer to ±1, the stronger the relationship. What relationship are we talking about? Consider a lag 1 value close to +1 as an example. This means that there is a strong relationship between X_t and X_{t-1}, X_{t-1} and $X_{t-2}, \ldots, X_{t-k+1}$ and X_{t-k}, and so on. The interesting thing is that what can happen in practice is that because of this serial correlation, it can appear that X_t has a strong relationship with X_{t-k}, k time units away from X_t, when in fact it is only because of this interaction. To sort out this potential problem, one estimates the partial autocorrelation function (PACF). The partial autocorrelation between X_t and X_{t-k} is the correlation between them after their mutual linear dependency on the intervening variables $X_{t-1}, \ldots, X_{t-k+1}$ has been removed. The sample PACF (SPACF) is given by the Yule–Walker equations,

$$\begin{bmatrix} 1 & r_1 & r_2 & \cdots & r_{k-2} & r_{k-1} \\ r_1 & 1 & r_1 & \cdots & r_{k-3} & r_{k-2} \\ . & . & . & \cdots & . & . \\ . & . & . & \cdots & . & . \\ r_{k-1} & r_{k-2} & r_{k-3} & \cdots & r_1 & 1 \end{bmatrix} \begin{bmatrix} \hat{\phi}_{k1} \\ \hat{\phi}_{k2} \\ . \\ . \\ \hat{\phi}_{kk} \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ . \\ . \\ r_k \end{bmatrix}. \tag{8}$$

The value of $\hat{\phi}_{kk}$ gives the estimate of the PACF at lag k. These equations can be solved using Cramer's Rule to obtain:

$$\hat{\phi}_{mm} = \frac{r_m - \sum_{j=1}^{m-1} \hat{\phi}_{m-1,j} r_{m-j}}{1 - \sum_{j=1}^{m-1} \hat{\phi}_{m-1,j} r_j}. \tag{9}$$

Once we have calculated these estimates for a stationary time series, we can use them to give an indication whether we should fit an $AR(p)$, $MA(q)$, or $ARMA(p,q)$ model. The criteria are in general:

- When the SACF dies down gradually and the SPACF has insignificant spikes at lags greater than p we should fit an $AR(p)$.
- When the SACF has a significant spike at lag q and the SPACF dies down gradually, we should fit an $MA(q)$.
- If both die down gradually, we fit an $ARMA(p,q)$. In this case, we will have to progressively increase p, q until we get a suitable model.

The last point brings up an interesting question; how do we decide between competing models? In fact, the situation is often not as simple as these criteria make it seem. Sometimes it is difficult to decide between for instance, an $AR(3)$ and an $ARMA(1,1)$ model. An aid in identifying the appropriate model comes from the principle of parsimony, using criteria from Information Theory. The ►Akaike's Information Criterion (AIC) is one such measure (Akaike 1973). The goal is to pick the model that minimises

$$AIC = -\frac{2}{T}\{\ln(\text{likelihood}) + l\}. \tag{10}$$

Here, l is the number of parameters fitted and T the number of data values. There is a competing criterion, that penalises the number of parameters fitted even more, called the (Schwarz) Bayesian Information Criterion (BIC) (Schwarz 1978),

$$BIC = -\frac{2}{T}\ln(\text{likelihood}) + \frac{l\ln(T)}{T}. \tag{11}$$

Moving Average Process

In a moving average process $MA(q)$, the present value of the series is written as the weighted sum of past shocks:

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}, \qquad (12)$$

where $Z_t \sim WN(0, \sigma_Z^2)$.

We find immediately that

$$E(X_t) = 0,$$
$$Var(X_t) = \sigma_Z^2 (1 + \sum_{i=1}^{q} {\theta_1}^2). \qquad (13)$$

Autoregressive Process

The general form of an $AR(p)$ process is:

$$\phi(B) X_t = Z_t$$
$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p) X_t = Z_t$$
$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t.$$

A first order autoregressive process $AR(1)$ is referred to as a Markov Chain (see ▶Markov Chains). It can, through successive substitutions, be written as:

$$\begin{aligned} X_t &= \phi X_{t-1} + Z_t \\ &= \phi(\phi X_{t-2} + Z_{t-1}) + Z_t \\ &= Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \phi^3 Z_{t-3} + \cdots. \end{aligned}$$

From this we write $Var(X_t) = \sigma_Z^2 (1 + \sum_{i=1}^{\infty} \phi^{2i}) = \dfrac{\sigma_Z^2}{1 - \phi^2}$.

Any $AR(p)$ process can be rewritten as an infinite order moving average process.

Stationarity

If neither the mean μ_t nor the autocovariances $\gamma_X(h)$ are dependent on t, then the process is said to be *weakly stationary*. This means that the autocovariances depend only on the length of time separating the observations h. A process is strictly stationary if for any values of $h_1, h_2, \ldots, h_n$, the joint distribution of $(X_t, X_{t+h_1}, X_{t+h_2}, \ldots, X_{t+h_n})$ depends not on t, but only on the time intervals between the variables. A sufficient condition for negating weak stationarity (and thus strict stationarity) is failure of the unit root test. A stochastic process has a unit root if its characteristic equation has 1 as a root.

For $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t$, the characteristic equation is given by:

$$m^p - \phi_1 m^{p-1} - \phi_2 m^{p-2} - \cdots - \phi_{p-1} m - \phi_p = 0.$$

If $m = 1$ is a root of this characteristic equation, the process has a unit root or is termed integrated of order 1, denoted $I(1)$. A first difference of the time series will be stationary.

If the characteristic equation has a unit root of multiplicity r then the process is integrated of order r, denoted $I(r)$. The time series differenced r times will be stationary. The process defined by $X_t = X_{t-1} + Z_t$ has a unit root and this process defines a ▶random walk. The process is customarily started at zero when $t = 0$, so $X_1 = Z_1$, and $X_t = \sum_{i=1}^{t} Z_i$, so we obtain $Var(X_t) = t\sigma_Z^2$, dependent on t.

Conditional Heteroscedastic Modelling

There is one particular type of non-stationarity that is receiving increasing attention. In financial markets, and other applications such as modelling wind farm output (Boland et al. 2007), a phenomenon has been identified wherein the SACF of Z_t shows no residual autocorrelation, but the SACF of ${Z_t}^2$ does. This property means that the noise is uncorrelated but not independent. This is reflective of a process that retains conditional heteroscedasticity, wherein there is what is termed as volatility clustering. Periods of high volatility can be followed by periods of low volatility. The volatility is generally modelled by Autoregressive Conditional Heteroscedastic (ARCH) or Generalised ARCH (GARCH) models, or utilising a concept called realised volatility (see Tsay 2005 and references therein).

About the Author

Associate Professor John Boland has extensive expertise in environmental modelling, specialising in time series and statistical modelling of climate variables and modelling systems under uncertainty. He has published 70 papers in many areas of environmental modelling and is Associate Editor for two international journals: *Renewable Energy*, and *Case Studies in Business, Industry and Government Statistics*. He, along with colleagues, is present or past holder of three Australian Research Council (ARC) Linkage and three Discovery grants. One Linkage and two Discovery Grants have been in the area of water cycle management. The present one in this area is focusing on using advanced mathematical techniques to better understand the structure of the Murray Darling Basin and its management options. Presently, he holds a Linkage and a Discovery grant in the energy area, focusing on the future of the electricity supply network with significant increases in the penetration of renewable energy into the grid. All these projects focus on stochastic modelling and risk management, particularly utilising conditional value at risk. His expertise in modelling systems under uncertainty has led him to being selected to run a one day Masterclass in Managing Data for Energy Storage and Renewable Energy Supply at the Energy Storage Forum, Beijing, March 2010.

Cross References

►Business Forecasting Methods
►Forecasting with ARIMA Processes
►Forecasting: An Overview
►Intervention Analysis in Time Series
►Mathematical and Statistical Modeling of Global Warming
►Moving Averages
►Seasonality
►Structural Time Series Models
►Time Series

References and Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In Petrov BN, Csaki F (eds) Second international symposium on information theory. Akademia Kiado, Budapest, pp 267–281

Boland J, Gilbert K, Korolkowicz M (10–13 December 2007) Modelling wind farm output variability. MODSIM07, Christchurch, New Zealand

Boland J (2008) Time series and statistical modeling of solar radiation. In Badescu V (ed) Recent advances in solar radiation modeling. Springer, Berlin, pp 283–312

Box G, Jenkins G (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco, CA

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Tsay RS (2005) Analysis of financial time series, 2nd edn. Wiley, New York

Brownian Motion and Diffusions

Vincenzo Capasso
Professor of Probability and Mathematical Statistics
University of Milan, Milan, Italy

Brownian Motion and the Wiener Process

A small particle (e.g., a pollen corn) suspended in a liquid is subject to infinitely many collisions with atoms, and therefore it is impossible to observe its exact trajectory. With the help of a microscope it is only possible to confirm that the movement of the particle is entirely chaotic. This type of movement, discovered under similar circumstances by the botanist Robert Brown, is called Brownian motion. As its mathematical inventor Einstein already observed, it is necessary to make approximations, in order to describe the process. The formalized mathematical model defined on the basis of these is called a Wiener process. Henceforth, we will limit ourselves to the study of the one-dimensional Wiener process in $\mathbb{R}$, under the assumption that the three components determining its motion in space are independent.

Definition 1 A real-valued process $(W_t)_{t\in\mathbb{R}_+}$ is a *Wiener process* if it satisfies the following conditions:

1. $W_0 = 0$ almost surely.
2. $(W_t)_{t\in\mathbb{R}_+}$ is a process with independent increments.
3. $W_t - W_s$ is normally distributed with $N(0, t-s)$, $(0 \le s < t)$.

Remark 1 From point 3 of Definition 1 it becomes obvious that every Wiener process is time homogeneous.

Proposition 1 *If $(W_t)_{t\in\mathbb{R}_+}$ is a Wiener process, then*

1. $E[W_t] = 0$ *for all* $t \in \mathbb{R}_+$,
2. $K(s,t) = Cov[W_t, W_s] = \min\{s,t\}$, $s,t \in \mathbb{R}_+$.

Proof 1. Fixing $t \in \mathbb{R}$, we observe that $W_t = W_0 + (W_t - W_0)$ and thus $E[W_t] = E[W_0] + E[W_t - W_0] = 0$. The latter is given by the fact that $E[W_0] = 0$ (by 1 of Definition 1) and $E[W_t - W_0] = 0$ (by 3 of Definition 1).

2. Let $s,t \in \mathbb{R}_+$ and $Cov[W_t, W_s] = E[W_t W_s] - E[W_t]E[W_s]$, which (by point 1) gives $Cov[W_t, W_s] = E[W_t W_s]$. For simplicity, if we suppose that $s < t$, then

$$E[W_t W_s] = E[W_s(W_s + (W_t - W_s))] = E\left[W_s^2\right] + E[W_s(W_t - W_s)].$$

Since $(W_t)_{t\in\mathbb{R}_+}$ has independent increments, we obtain

$$E[W_s(W_t - W_s)] = E[W_s]E[W_t - W_s]$$

and by 3 of Definition 1 it follows that this is equal to zero, thus

$$Cov[W_t, W_s] = E\left[W_s^2\right] = Var[W_s].$$

If we now observe that $W_s = W_0 + (W_s - W_0)$ and hence $Var[W_s] = Var[W_0 + (W_s - W_0)]$, then, by the independence of the increments of the process, we get

$$Var[W_0 + (W_s - W_0)] = Var[W_0] + Var[W_s - W_0].$$

Therefore, by points 1 and 3 of Definition 1 it follows that

$$Var[W_s] = s = \inf\{s,t\},$$

which completes the proof. □

Remark 2 By 1 of Definition 1, it follows, for all $t \in \mathbb{R}_+$, $W_t = W_t - W_0$ almost surely and by 3 of the same definition, that W_t is distributed as $N(0,t)$. Thus

$$P(a \le W_t \le b) = \frac{1}{\sqrt{2\pi t}} \int_a^b e^{-\frac{x^2}{2t}} dx, \qquad a \le b.$$

Remark 3 The Wiener process is a Gaussian process. In fact, if $n \in \mathbb{N}^*$, $(t_1, \ldots, t_n) \in \mathbb{R}_+^n$ with $0 = t_0 < t_1 < \ldots < t_n$

and $(a_1, \ldots, a_n) \in \mathbb{R}^n, (b_1, \ldots, b_n) \in \mathbb{R}^n$, such that $a_i \leq b_i, i = 1, 2, \ldots, n$, it can be shown that

$$P(a_1 \leq W_{t_1} \leq b_1, \ldots, a_n \leq W_{t_n} \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} g(0|x_1, t_1) g(x_1|x_2, t_2 - t_1) \cdots g(x_{n-1}|x_n, t_n - t_{n-1}) dx_n \cdots dx_1, \quad (1)$$

where

$$g(x|y,t) = \frac{e^{-\frac{|x-y|^2}{2t}}}{\sqrt{2\pi t}}.$$

Proposition 2 *If $(W_t)_{t\in\mathbb{R}_+}$ is a Wiener process, then it is also a martingale* (see ▶Martingales).

Proof The proposition follows from the fact that $(W_t)_{t\in\mathbb{R}_+}$ is a zero mean process with independent increments. □

Theorem 1 *Every Wiener process $(W_t)_{t\in\mathbb{R}_+}$ is a Markov process.*

Proof The theorem follows directly from the fact that $(W_t)_{t\in\mathbb{R}_+}$ is a process with independent increments. □

Remark 4 Since Brownian motion is continuous in probability, it admits a separable and progressively measurable modification.

Theorem 2 (Kolmogorov's continuity theorem). *Let $(X_t)_{t\in\mathbb{R}_+}$ be a separable real-valued stochastic process. If there exist positive real numbers r, c, ϵ, δ such that*

$$\forall h < \delta, \forall t \in \mathbb{R}_+, \qquad E[|X_{t+h} - X_t|^r] \leq ch^{1+\epsilon}, \quad (2)$$

then, for almost every $\omega \in \Omega$, the trajectories are continuous in $\mathbb{R}_+$.

Theorem 3 *If $(W_t)_{t\in\mathbb{R}_+}$ is a real-valued Wiener process, then it has continuous trajectories almost surely.*

Proof Let $t \in \mathbb{R}_+$ and $h > 0$. Because $W_{t+h} - W_t$ is normally distributed as $N(0, h)$, putting $Z_{t,h} = \frac{W_{t+h} - W_t}{\sqrt{h}}$, $Z_{t,h}$ has standard normal distribution. Therefore, it is clear that there exists an $r > 2$ such that $E[|Z_{t,h}|^r] > 0$, and thus $E[|W_{t+h} - W_t|^r] = E[|Z_{t,h}|^r] h^{\frac{r}{2}}$. If we write $r = 2(1+\epsilon)$, we obtain $E[|W_{t+h} - W_t|^r] = ch^{1+\epsilon}$, with $c = E[|Z_{t,h}|^r]$. The assertion then follows by Kolmogorov's continuity theorem. □

Theorem 4 *If $(W_t)_{t\in\mathbb{R}_+}$ is a real-valued Wiener process, then*

1. $P(\sup_{t\in\mathbb{R}_+} W_t = +\infty) = 1$
2. $P(\inf_{t\in\mathbb{R}_+} W_t = -\infty) = 1$

Theorem 5 *If $(W_t)_{t\in\mathbb{R}_+}$ is a real-valued Wiener process, then,*

$$\forall h > 0, \qquad P\left(\max_{0\leq s\leq h} W_s > 0\right) = P\left(\min_{0\leq s\leq h} W_s < 0\right) = 1.$$

Moreover, for almost every $\omega \in \Omega$ the process $(W_t)_{t\in\mathbb{R}_+}$ has a zero (i.e., crosses the spatial axis) in $[0, h]$, for all $h > 0$.

Theorem 6 *Almost every trajectory of the Wiener process $(W_t)_{t\in\mathbb{R}_+}$ is differentiable almost nowhere.*

Proposition 3 (scaling property). *Let $(W_t)_{t\in\mathbb{R}_+}$ be a Wiener process. Then the time-scaled process $(\tilde{W}_t)_{t\in\mathbb{R}_+}$ defined by*

$$\tilde{W}_t = tW_{1/t}, \; t > 0, \qquad \tilde{W}_0 = 0$$

is also a Wiener process.

Proposition 4 (Strong law of large numbers). *Let $(W_t)_{t\in\mathbb{R}_+}$ be a Wiener process. Then*

$$\frac{W_t}{t} \to 0, \qquad as \qquad t \to +\infty, \qquad a.s.$$

Proposition 5 (Law of iterated logarithms). *Let $(W_t)_{t\in\mathbb{R}_+}$ be a Wiener process. Then*

$$\limsup_{t\to+\infty} \frac{W_t}{\sqrt{2t \ln\ln t}} = 1, \qquad a.s.,$$
$$\liminf_{t\to+\infty} \frac{W_t}{\sqrt{2t \ln\ln t}} = -1, \qquad a.s.$$

As a consequence, for any $\epsilon > 0$, there exists a $t_0 > 0$, such that for any $t > t_0$ we have

$$-(1+\epsilon)\sqrt{2t \ln\ln t} \leq W_t \leq (1+\epsilon)\sqrt{2t \ln\ln t}, \qquad \text{a.s.}$$

Existence of the Wiener process is guaranteed by the following fundamental theorem (see, e.g., Billingsley 1968).

Theorem 7 (Donsker). *Let $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ be a sequence of independent identically distributed random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$, with mean 0 and finite, positive variance σ^2:*

$$E[\xi_n] = 0, \qquad E\left[\xi_n^2\right] = \sigma^2.$$

Let $S_0 = 0$ and, for any $n \in \mathbb{N} \setminus \{0\}$, let $S_n = \xi_1 + \xi_2 + \cdots + \xi_n$. Then the sequence of random processes defined by

$$X_n(t, \omega) = \frac{1}{\sigma\sqrt{n}} S_{[nt]}(\omega) + (nt - [nt]) \frac{1}{\sigma\sqrt{n}} \xi_{[nt]+1}(\omega)$$

for any $t \in \mathbb{R}_+$, $\omega \in \Omega$, $n \in \mathbb{N}$, weakly converges to a Wiener process.

Diffusion Processes

Markov Processes

Definition 2 *Let $(X_t)_{t\in\mathbb{R}_+}$ be a stochastic process on a probability space, valued in $(E,\mathcal{B})$ and adapted to the increasing family $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ of σ-algebras of subsets of $\mathcal{F}$. $(X_t)_{t\in\mathbb{R}_+}$ is a* Markov process *with respect to $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ if the following condition is satisfied:*

$$\forall B\in\mathcal{B},\forall(s,t)\in\mathbb{R}_+\times\mathbb{R}_+, s<t:\quad P(X_t\in B|\mathcal{F}_s)=P(X_t\in B|X_s)\ a.s. \tag{3}$$

Remark 5 If, for all $t\in\mathbb{R}_+$, $\mathcal{F}_t=\sigma(X_r,0\le r\le t)$, then condition (3) becomes

$$P(X_t\in B|X_r,0\le r\le s)=P(X_t\in B|X_s)\ \text{a.s.}$$

for all $B\in\mathcal{B}$, for all $(s,t)\in\mathbb{R}_+\times\mathbb{R}_+$, and $s<t$.

Proposition 6 *Under the assumptions of Definition 2, the following two statements are equivalent:*

1. *For all $B\in\mathcal{B}$ and all $(s,t)\in\mathbb{R}_+\times\mathbb{R}_+, s<t$: $P(X_t\in B|\mathcal{F}_s)=P(X_t\in B|X_s)$ almost surely.*
2. *For all $g:E\to\mathbb{R}$, $\mathcal{B}$-$\mathcal{B}_{\mathbb{R}}$-measurable such that $g(X_t)\in L^1(P)$ for all t, for all $(s,t)\in\mathbb{R}_+^2, s<t$: $E[g(X_t)|\mathcal{F}_s]=E[g(X_t)|X_s]$ almost surely.*

Theorem 8 *Every real stochastic process $(X_t)_{t\in\mathbb{R}_+}$ with independent increments is a Markov process.*

Proposition 7 Consider a real valued Markov process $(X_t)_{t\in\mathbb{R}_+}$, and let

$$p(s,x,t,A)=P(X_t\in A|X_s=x),\quad 0\le s<t<\infty, x\in\mathbb{R}, A\in\mathcal{B}_{\mathbb{R}}.$$

p is a *Markov transition probability function*, i.e., it is a non-negative function defined for $0\le s<t<\infty, x\in\mathbb{R}, A\in\mathcal{B}_{\mathbb{R}}$, which satisfies the following properties

1. For all $0\le s<t<\infty$, for all $A\in\mathcal{B}_{\mathbb{R}}$, $p(s,\cdot,t,A)$ is $\mathcal{B}_{\mathbb{R}}$-measurable.
2. For all $0\le s<t<\infty$, for all $x\in\mathbb{R}$, $p(s,x,t,\cdot)$ is a probability measure on $\mathcal{B}_{\mathbb{R}}$.
3. p satisfies the Chapman–Kolmogorov equation:

$$p(s,x,t,A)=\int_{\mathbb{R}}p(s,x,r,dy)p(r,y,t,A)\quad\forall x\in\mathbb{R}, s<r<t.$$

Definition 3 A Markov process $(X_t)_{t\in[t_0,T]}$ is said to be *homogeneous* if the transition probability functions $p(s,x,t,A)$ depend on t and s only through their difference $t-s$. Therefore, for all $(s,t)\in[t_0,T]^2$, $s<t$, for all $u\in[0,T-t]$, for all $A\in\mathcal{B}_{\mathbb{R}}$, and for all $x\in\mathbb{R}$:

$$p(s,x,t,A)=p(s+u,x,t+u,A)\qquad\text{a.s.}$$

Remark 6 If $(X_t)_{t\in\mathbb{R}_+}$ is a homogeneous Markov process with transition probability function p, then, for all $(s,t)\in\mathbb{R}_+{}^2, s<t$, for all $A\in\mathcal{B}_{\mathbb{R}}$, and for all $x\in\mathbb{R}$, we obtain

$$p(s,x,t,A)=p(0,x,t-s,A)\ \text{a.s.}$$

We may then define a one-parameter transition function

$$p(\tilde{t},x,A):=p(0,x,t-s,A)$$

with $\tilde{t}=(t-s), x\in\mathbb{R}, A\in\mathcal{B}_{\mathbb{R}}$.

Semigroups Associated with Markov Transition Probability Functions

Let $BC(\mathbb{R})$ be the space of all continuous and bounded functions on $\mathbb{R}$, endowed with the norm $\|f\|=\sup_{x\in\mathbb{R}}|f(x)|$ $(<\infty)$, and let $p(s,x,t,A)$ be a transition probability function $(0\le s<t\le T, x\in\mathbb{R}, A\in\mathcal{B}_{\mathbb{R}})$. We consider the operator

$$T_{s,t}:BC(\mathbb{R})\to BC(\mathbb{R}),\qquad 0\le s<t\le T,$$

defined by assigning, for all $f\in BC(\mathbb{R})$,

$$(T_{s,t}f)(x)=\int_{\mathbb{R}}f(y)p(s,x,t,dy).$$

If $s=t$, then

$$p(s,x,s,A)=\begin{cases}1 & \text{if } x\in A,\\ 0 & \text{if } x\notin A.\end{cases}$$

Therefore,

$$T_{t,t}=I\ \text{(identity)}. \tag{4}$$

Moreover, we have that

$$T_{s,t}T_{t,u}=T_{s,u},\qquad 0\le s<t<u. \tag{5}$$

In fact, if $f\in BC(\mathbb{R})$ and $x\in\mathbb{R}$,

$$\begin{aligned}&(T_{s,t}(T_{t,u}f))(x)\\ &=\int_{\mathbb{R}}(T_{t,u}f)(y)p(s,x,t,dy)\\ &=\int\int_{\mathbb{R}^2}f(z)p(t,y,u,dz)p(s,x,t,dy)\\ &=\int_{\mathbb{R}}f(z)\int_{\mathbb{R}}p(t,y,u,dz)p(s,x,t,dy)\\ &\qquad\text{(by Fubini's theorem)}\\ &=\int_{\mathbb{R}}f(z)p(s,x,u,dz)\\ &\qquad\text{(by the Chapman–Kolmogorov equation)}\\ &=(T_{s,u}f)(x).\end{aligned}$$

Definition 4 The family $\{T_{s,t}\}_{0\le s\le t\le T}$ is a *semigroup associated with the transition probability function* $p(s,x,t,A)$ (or with its corresponding Markov process).

Definition 5 If $(X_t)_{t\in\mathbb{R}_+}$ is a Markov process with transition probability function p and associated semigroup $\{T_{s,t}\}$, then the operator

$$\mathcal{A}_s f = \lim_{h\downarrow 0} \frac{T_{s,s+h}f - f}{h}, \qquad s \geq 0, f \in BC(\mathbb{R})$$

is called the *infinitesimal generator of the Markov process* $(X_t)_{t\geq 0}$. Its domain $\mathcal{D}_{\mathcal{A}_s}$ consists of all $f \in BC(\mathbb{R})$ for which the above limit exists uniformly (and therefore in the norm of $BC(\mathbb{R})$) (see, e.g., Feller 1971).

Remark 7 From the preceding definition we observe that

$$(\mathcal{A}_s f)(x) = \lim_{h\downarrow 0} \frac{1}{h} \int_{\mathbb{R}} [f(y) - f(x)] p(s, x, s+h, dy).$$

Definition 6 Let $(X_t)_{t\in\mathbb{R}_+}$ be a Markov process with transition probability function $p(s, x, t, A)$, and $\{T_{s,t}\}$ $(s, t \in \mathbb{R}_+, s \leq t)$ its associated semigroup. If, for all $f \in BC(\mathbb{R})$, the function

$$(t, x) \in \mathbb{R}_+ \times \mathbb{R} \to (T_{t,t+\lambda} f)(x) = \int_{\mathbb{R}} p(t, x, t+\lambda, dy) f(y) \in \mathbb{R}$$

is continuous for all $\lambda > 0$, then we say that the process satisfies the *Feller property*.

Theorem 9 *If $(X_t)_{t\in\mathbb{R}_+}$ is a Markov process with right-continuous trajectories satisfying the Feller property, then, for all $t \in \mathbb{R}_+$, $\mathcal{F}_t = \mathcal{F}_{t^+}$, where $\mathcal{F}_{t^+} = \bigcap_{t'>t} \sigma(X(s), 0 \leq s \leq t')$, and the filtration $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ is right-continuous.*

Remark 8 It can be shown that $\mathcal{F}_{t^+}$ is a σ-algebra.

Example 1 Wiener processes are processes with the Feller property, or simply *Feller processes*.

If we consider the time-homogeneous case, a Markov process $(X_t)_{t\in\mathbb{R}_+}$ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, will be defined in terms of a transition kernel $p(t, x, B)$ for $t \in \mathbb{R}_+$, $x \in \mathbb{R}$, $B \in \mathcal{B}_{\mathbb{R}}$, such that

$$p(h, X_t, B) = P(X_{t+h} \in B | \mathcal{F}_t) \qquad \forall t, h \in \mathbb{R}_+, B \in \mathcal{B}_{\mathbb{R}},$$

given that $(\mathcal{F}_t)_{t\in\mathbb{R}_+}$ is the natural filtration of the process. Equivalently, if we denote by $BC(\mathbb{R})$ the space of all continuous and bounded functions on $\mathbb{R}$, endowed with the *sup norm*,

$$E[g(X_{t+h})|\mathcal{F}_t] = \int_E g(y) p(h, X_t, dy) \quad \forall t, h \in \mathbb{R}_+, g \in BC(\mathbb{R}).$$

In this case the transition semigroup of the process is a one-parameter contraction semigroup $(T(t), t \in \mathbb{R}_+)$ on $BC(\mathbb{R})$ defined by

$$T(t)g(x) := \int_E g(y) p(t, x, dy) = E[g(X_t)|X_0 = x], \quad x \in \mathbb{R},$$

for any $g \in BC(\mathbb{R})$. The infinitesimal generator will be time independent. It is defined as

$$\mathcal{A}g = \lim_{t\to 0+} \frac{1}{t}(T(t)g - g)$$

for $g \in \mathcal{D}(\mathcal{A})$, the subset of $BC(\mathbb{R})$ for which the above limit exists, in $BC(\mathbb{R})$, with respect to the sup norm. Given the above definitions, it is obvious that for all $g \in \mathcal{D}(\mathcal{A})$,

$$\mathcal{A}g(x) = \lim_{t\to 0+} \frac{1}{t} E[g(X_t)|X_0 = x], \qquad x \in \mathbb{R}.$$

If $(T(t), t \in \mathbb{R}_+)$ is the contraction semigroup associated with a time-homogeneous Markov process, it is not difficult to show that the mapping $t \to T(t)g$ is right-continuous in $t \in \mathbb{R}_+$ provided that $g \in BC(\mathbb{R})$ is such that the mapping $t \to T(t)g$ is right continuous in $t = 0$. Then, for all $g \in \mathcal{D}(\mathcal{A})$ and $t \in \mathbb{R}_+$,

$$\int_0^t T(s)g\,ds \in \mathcal{D}(\mathcal{A})$$

and

$$T(t)g - g = \mathcal{A} \int_0^t T(s)g\,ds = \int_0^t \mathcal{A}T(s)g\,ds = \int_0^t T(s)\mathcal{A}g\,ds$$

by considering Riemann integrals. The following, so-called *Dynkin's formula*, establishes a fundamental link between ►Markov processes and ►martingales (see Rogers and Williams 1994, p 253).

Theorem 10 *Assume $(X_t)_{t\in\mathbb{R}_+}$ is a time-homogeneous Markov process on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, with transition kernel $p(t, x, B)$, $t \in \mathbb{R}_+$, $x \in \mathbb{R}$, $B \in \mathcal{B}_{\mathbb{R}}$. Let $(T(t), t \in \mathbb{R}_+)$ denote its transition semigroup and $\mathcal{A}$ its infinitesimal generator. Then, for any $g \in \mathcal{D}(\mathcal{A})$, the stochastic process*

$$M(t) := g(X_t) - g(X_0) - \int_0^t \mathcal{A}g(X_s)\,ds$$

is an $\mathcal{F}_t$-martingale.

The next proposition shows that a Markov process is indeed characterized by its infinitesimal generator via a martingale problem (see, e.g., Rogers and Williams 1994, p 253).

Theorem 11 (Martingale problem for Markov processes). *If an RCLL (right continuous with left limits) Markov process $(X_t)_{t\in\mathbb{R}_+}$ is such that*

$$g(X_t) - g(X_0) - \int_0^t \mathcal{A}g(X_s)\,ds$$

is an $\mathcal{F}_t$-martingale for any function $g \in \mathcal{D}(\mathcal{A})$, where $\mathcal{A}$ is the infinitesimal generator of a contraction semigroup on

E, then X_t is equivalent to a Markov process having $\mathcal{A}$ as its infinitesimal generator.

Markov Diffusion Processes

Definition 7 A Markov process on $\mathbb{R}$ with transition probability function $p(s,x,t,A)$ is called a *diffusion process* if

1. For all $\epsilon > 0$, for all $t \geq 0$, and for all $x \in \mathbb{R}$: $\lim_{h\downarrow 0} \frac{1}{h} \int_{|x-y|>\epsilon} p(t,x,t+h,dy) = 0$.
2. There exist $a(t,x)$ and $b(t,x)$ such that, for all $\epsilon > 0$, for all $t \geq 0$, and for all $x \in \mathbb{R}$,

$$\lim_{h\downarrow 0} \frac{1}{h} \int_{|x-y|<\epsilon} (y-x)p(t,x,t+h,dy) = a(t,x),$$
$$\lim_{h\downarrow 0} \frac{1}{h} \int_{|x-y|<\epsilon} (y-x)^2 p(t,x,t+h,dy) = b(t,x).$$

$a(t,x)$ is the *drift coefficient* and $b(t,x)$ the *diffusion coefficient* of the process.

Proposition 8 *If $(X_t)_{t\in\mathbb{R}_+}$ is a diffusion process with transition probability function p and drift and diffusion coefficients $a(x,t)$ and $b(x,t)$, respectively, and if $\mathcal{A}_s$ is the infinitesimal generator associated with p, then we have that*

$$(\mathcal{A}_s f)(x) = \frac{\partial f}{\partial x} a(s,x) + \frac{1}{2}\frac{\partial^2 f}{\partial x^2} b(s,x), \qquad (6)$$

provided that f is bounded and twice continuously differentiable.

Proposition 9 *A Wiener process is a time homogeneous diffusion process, with drift zero and diffusion coefficient equal to* 1.

Its infinitesimal generator is then

$$(\mathcal{A}f)(x) = \frac{1}{2}\frac{\partial^2 f}{\partial x^2}, \qquad (7)$$

About the Author

For biography *see* the entry ►Axioms of Probability.

Cross References

►Almost Sure Convergence of Random Variables
►Central Limit Theorems
►First Exit Time Problem
►Itô Integral
►Khmaladze Transformation
►Lévy Processes
►Markov Processes
►Non-Uniform Random Variate Generations
►Random Walk
►Statistical Inference for Stochastic Processes
►Statistical Modeling of Financial Markets
►Stochastic Difference Equations and Applications
►Stochastic Processes
►Stochastic Processes: Applications in Finance and Insurance
►Stochastic Processes: Classification
►Strong Approximations in Probability and Statistics
►Testing Exponentiality of Distribution

References and Further Reading

Ash RB, Gardner MF (1975) Topics in stochastic processes. Academic, London
Billingsley P (1968) Convergence of probability measures. Wiley, New York
Breiman L (1968) Probability. Addison-Wesley, Reading, MA
Capasso V, Bakstein D (2005) An introduction to continuous-time stochastic processes theory, models, and applications to finance, biology, and medicine, Birkhäuser, Boston, MA
Doob JL (1953) Stochastic processes. Wiley, New York
Dynkin EB (1965) Markov processes, vols 1–2. Springer, Berlin
Feller W (1971) An introduction to probability theory and its applications. 2nd edn. vol 2. Wiley, New York
Friedman A (1975) Stochastic differential equations and applications. Academic, London
Gihman II, Skorohod AV (1974–1979) The theory of stochastic processes, vols 1–3. Springer, Berlin
Ikeda N, Watanabe S (1989) Stochastic differential equations and diffusion processes. North-Holland, Kodansha
Karatzas I, Shreve SE (1991) Brownian motion and stochastic calculus. Springer, New York
Karlin S, Taylor HM (1975) A first course in stochastic processes. Academic, New York
Karlin S, Taylor HM (1981) A second course in stochastic processes. Academic, New York
Rogers LCG, Williams D (1994) Diffusions, Markov processes and martingales, vol 1. Wiley, New York
Taira K (1988) Diffusion processes and partial differential equations. Academic, New York

Business Forecasting Methods

Rob J. Hyndman
Professor of Statistics
Monash University, Melbourne, VIC, Australia

Forecasting, Planning, and Goals

Forecasting is a common statistical task in business where it helps inform decisions about scheduling of production, transportation and personnel, and provides a guide to long-term strategic planning. However, business forecasting is often done poorly and is frequently confused with planning and goals. They are three different things.

Forecasting is about predicting the future as accurately as possible, given all the information available including historical data and knowledge of any future events that might impact the forecasts.

Goals are what you would like to happen. Goals should be linked to forecasts and plans, but this does not always occur. Too often, goals are set without any plan for how to achieve them, and no forecasts for whether they are realistic.

Planning is a response to forecasts and goals. Planning involves determining the appropriate actions that are required to make your forecasts match your goals.

Forecasting should be an integral part of the decision-making activities of management, as it can play an important role in many areas of a company. Modern organizations require short-, medium-, and long-term forecasts, depending on the specific application.

Short-term forecasts are needed for scheduling of personnel, production, and transportation. As part of the scheduling process, forecasts of demand are often also required.

Medium-term forecasts are needed to determine future resource requirements in order to purchase raw materials, hire personnel, or buy machinery and equipment.

Long-term forecasts are used in strategic planning. Such decisions must take account of market opportunities, environmental factors, and internal resources.

An organization needs to develop a forecasting system involving several approaches to predicting uncertain events. Such forecasting systems require the development of expertise in identifying forecasting problems, applying a range of forecasting methods, selecting appropriate methods for each problem, and evaluating and refining forecasting methods over time. It is also important to have strong organizational support for the use of formal forecasting methods if they are to be used successfully.

Commonly Used Methods

Typically, businesses use relatively simple forecasting methods that are often not based on statistical modelling. However, the use of statistical forecasting is growing and some of the most commonly used methods are listed below.

Time Series Methods

Let the historical time series data be denoted by $y_1, \ldots, y_n$, and the forecast of y_{n+h} be given by $\hat{y}_{n+h|n}$, $h > 0$.

- Naïve forecasting is where the forecasts of all future values of a time series are set to be equal to the last observed value: $\hat{y}_{n+h|n} = y_n$, $h = 1, 2, \ldots$. If the data follow a random walk process ($y_t = y_{t-1} + e_t$ where e_t is white noise – a series of iid random variables with zero mean), then this is the optimal method of forecasting. Consequently, it is popular for stock price and stock index forecasting, and for other time series that measure the behavior of a market that can be assumed to be efficient.
- Simple exponential smoothing was developed in the 1950s (Brown 1959) and has been widely used ever since. Forecasts can be computed recursively as each new data point is observed:

$$\hat{y}_{t+1|t} = \alpha y_t + (1-\alpha)\hat{y}_{t|t-1},$$

where $0 < \alpha < 1$. (Longer-term forecasts are constant: $\hat{y}_{t+h|t} = \hat{y}_{t+1|t}$, $h \geq 2$.) Consequently, only the most recent data point and most recent forecast need to be stored. This was an attractive feature of the method when computer storage was expensive. The method has proved remarkably robust to a wide range of time series, and is optimal for several processes including the ARIMA(0,1,1) process (Chatfield et al. 2001).
- Holt's linear method (Holt 1957) is an extension of simple exponential forecasting that allows a locally linear trend to be extrapolated. Forecasts are given by $\hat{y}_{t+h|t} = \ell_t + hb_t$ where

$$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}),$$
$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1},$$

and the two parameters α and β must lie in $[0,1]$. Here ℓ_t denotes the level of the series and b_t the slope of the trend at time t.
- For seasonal data, a popular method is the Holt–Winters' method, also introduced in Holt (1957), which extends Holt's method to include seasonal terms. Then $\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t-m+h_m^+}$ where

$$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}),$$
$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1},$$
$$s_t = \gamma(y_t - \ell_t) + (1-\gamma)s_{t-m},$$

$h_m^+ = [(h-1) \mod m] + 1$, and the three parameters α, β and γ all lie in [0,1].

There is also a multiplicative version of the Holt–Winters' method, and damped trend versions of both Holt's linear method and Holt–Winters' method (Makridakis et al. 1998). None of these methods are explicitly based on

underlying time series models, and as a result the estimation of parameters and the computation of prediction intervals is often not done. However, all the above methods have recently been shown to be optimal for some state space models (Hyndman et al. 2008), and maximum likelihood estimation of parameters, statistical model selection and computation of prediction intervals is now becoming more widespread.

Other time series models sometimes used in business forecasting include ARIMA models, GARCH models (especially in finance), structural models and ►neural networks.

Explanatory Models for Forecasting

The use of explanatory models in business forecasting does not have such a long history as the use of time series methods.

- Linear regression modeling (see ►Linear Regression Models) is now widely used (e.g., Pardoe 2006) where a variable to be forecast is modeled as a linear combination of potential input variables:

$$y_t = \sum_{j=1}^{J} c_j x_{j,t} + e_t,$$

where e_t denotes an iid error term with zero mean. An interesting application of regression models to forecasting is given by Byron and Ashenfelter (1995) who use a simple regression models to predict the quality of a Grange wine using simple weather variables. However, it is far more common for regression modelling to be used to explain historical variation than for it to be used for forecasting purposes.

- In some domains, the use of nonparametric additive models for forecasting is growing (e.g., Hyndman and Fan 2010). Here, the model is often of the form

$$y_t = \sum_{j=1}^{J} f_j(x_{j,t}) + e_t,$$

where f_j is a smooth nonlinear function to be estimated nonparametrically.

- In advertising, there is a well-developed culture of using distributed lag regression models (e.g., Hanssens et al. 2001) such as

$$y_t = \sum_{j=1}^{J} \alpha \lambda^j x_{t-j} + e_t,$$

where x_t denotes advertising expenditure in month t, $0 < \lambda < 1$ and $\alpha > 0$.

Data Mining Methods for Business Forecasting

Outside of traditional statistical modelling, an enormous amount of forecasting is done using data mining methods (see ►Data Mining). Most of these methods have no formal statistical model, prediction intervals are not computed, and there is limited model checking. But some of the data-mining methods have proven powerful predictors in some contexts, especially when there is a vast quantity of available data. Predictive methods include neural networks, support vector machines, and regression trees. Many of the best-known business predictive algorithms are based on data-mining methods including the prediction of Netflix ratings (see ►Data Mining Time Series Data, Forecasting Principles).

About the Author

For biography *see* the entry ►Forecasting: An Overview.

Cross References

►Business Statistics
►Data Mining Time Series Data
►Exponential and Holt-Winters Smoothing
►Forecasting Principles
►Forecasting: An Overview
►Time Series

References and Further Reading

Brown RG (1959) Statistical forecasting for inventory control. McGraw-Hill, New York

Byron RP, Ashenfelter O (1995) Predicting the quality of an unborn Grange. Econ Rec 71(212):40–53

Chatfield C, Koehler AB, Ord JK, Snyder RD (2001) A new look at models for exponential smoothing. J Roy Stat Soc, Ser D: Statistician 50(2):147–159

Hanssens DM, Parsons LJ, Schultz RL (2001) Market response models: econometric and time series analysis, 2nd edn. Kluwer Academic, Boston, MA

Holt CC (1957) Forecasting trends and seasonals by exponentially weighted averages, O.N.R. Memorandum 52/1957, Carnegie Institute of Technology

Hyndman RJ, Fan S (2010) Density forecasting for long-term peak electricity demand, IEEE Transactions on Power Systems, 25(2):1142–1153

Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008) Forecasting with exponential smoothing: the state space approach. Springer, Berlin

Makridakis S, Wheelwright SC, Hyndman RJ (1998) Forecasting: methods and applications, 3rd edn. Wiley, New York

Pardoe I (2006) Applied regression modeling: a business approach, Wiley, Hoboken, NJ

Business Intelligence

Jasna Soldic-Aleksic, Rade Stankic
Professors, Faculty of Economics
Belgrade University, Belgrade, Serbia

There is no unique definition of the term and concept of Business Intelligence (BI) adopted both in the academic community and commercial business circles. However, there are various short and broad definitions that emphasize some specific aspects of this complex concept. In general, the concept of Business Intelligence refers to in-depth analysis of company data for better decision-making. However, more specifically, the concept may be explained as follows: BI is an umbrella term combining architectures, tools, databases, analytical tools, applications, and methodologies for gathering, storing, analyzing, and providing access to data for improving business performance and helping enterprise users make better business and strategic decisions (Turban et al. 2007).

From a historical point of view, the term business intelligence appeared in the late 1950s, but many years later its usage has been widened to cover the sense it is associated with nowadays. The history of BI concept development is closely connected to the evolution of information systems for enterprise decision support. This is the field in which the roots of the BI concept can be identified, i.e., the Management Information System – MIS reporting systems of the 1970s. These systems were characterized by static reporting features without analytical capabilities. The next generation of information systems – the Executive Information System (EIS) that emerged in the early 1980s provided some additional capabilities, such as: ad-hoc reporting, on-demand reporting, forecasting and prediction, trend analysis, drill-down capability, and critical success factors. These capabilities were essential for computerized support to top-level managers and executives and they were all integrated in the BI system. While the widespread usage of the term and the concept had been recorded in the late 1990s, the rapid growth of BI tools and technologies became evident in 2000s, when many sophisticated data analysis tools were being included in BI enterprise information systems.

Considering the broader scene of the main business and technological trends that had an impact on the development of the BI concept, the following are particularly worth noting: globalization, rapid business growth, strong market competition, data explosion and information overload in all spheres of business and ordinary life, user dissatisfaction with fragmented information systems capabilities, client/server architecture, Enterprise Resource Planning (ERP), Data warehouse (DW) technology, artificial intelligence computing, and web-based technologies and applications. Among all of these technologies, one can be distinguished as the biggest technological catalyst for BI rapid development and success – the DW technology. DW is a repository of subject-oriented, consistent, integrated, historic, and non-volatile collection of data, organized in such way to allow the end-user easy access to data in a form acceptable for analytical processing activities. One of the key features of the DW refers to specific data organization. In contrast to classical data organization approach, with data stored in operational systems (legacy, inventory, shipping, or ERP) and organized according to a concrete business process (purchasing, shipping, billing...), data in DW are organized by subject. Special software, called ETL (Extract, Transform and Load) is responsible for extracting data from different sources, cleaning them up and loading them into a data warehouse. Thanks to this feature, DW provides fast retrieval of historical data (allowing users to analyze time series, trends, and historical patterns) and allows users to analyze broader sets of data. In addition to the *data warehouse*, the architecture of a BI system comprises three other components (Turban et al. 2006): *business analytics* as a collection of tools for data query and reporting, online analytical processing, statistical analysis, prediction, data visualization, ▶data mining, text and web mining; *business performance management* (BPM) used for monitoring and analyzing performance through business performance indicators; and a *user interface* that aims to adequate user communication with the BI system (e.g., dashboard, alerts, and notifications).

For analytical purposes, the BI concept assumes a broad usage of statistical software products and machine learning techniques. The combination of statistics and predictive analytics is recognized as one of the crucial trends in business intelligence. So, it is not unusual to observe some of the most famous statistical software products being incorporated in the integrated predictive analytics technology suits or portfolios. The most impressive examples in this sense are the transformation of two statistical software leaders – SPSS ("*Statistical Package for the Social Sciences*") and SAS ("*Statistical Analysis System*") into remarkable providers of predictive analytics software and services. By using this combined technology, organizations can address their vital BI needs, be it reporting, querying, business visualization, statistics, survey analysis, ▶data mining, text and Web analysis, decision optimization, or, very often, a combination of previous capabilities.

The market of BI products and tools has been on the rise in the last few years and shows a strong tendency for further expansion in the future. According to the Business Intelligence Tools Survey (published in October 2008), the most renowned providers of BI products are: Oracle, SAP, SAS Institute, IBM, EFM Software, Information Builders, Microsoft, QlikTech, Microstrategy, and Actuate.

Cross References

►Business Statistics
►Statistical Software: An Overview

References and Further Reading

Inmon WH (2005) Building the data warehouse, 4th edn. Wiley, New York

Turban E, Arinson JE, Liang T, Sharda R (2006) Decision support and business intelligence systems, 8th edn. Pearson Prentice Hall, Upper Saddle River, NJ

Turban E, Ledner D, McLean E, Wetherbe J (2007) Information technology for management: transforming organizations in the digital economy, 6th edn. Wiley, New York

Business Statistics

Mark L. Berenson
Professor
Montclair State University, Montclair, NJ, USA

An Overview and Definitions

Business statistics can be viewed from two perspectives. One focuses on the use of the statistics themselves. The other sees business statistics as a practitioner-based discipline. To the user, business statistics are intended to be helpful information pertaining to the efficacy of either a company (e.g., financial statements and financial ratios at a particular point in time over several time periods), an industry (e.g., a series of data or an index constructed over time), or the overall economy. The information gleaned is often intended to help the user in making decisions regarding planning, monitoring or investing. To the practitioner, however, business statistics is a particular academic branch of study, similar to other applications-based branches of statistics such as agricultural statistics, ►astrostatistics, ►biostatistics, educational statistics, ►medical statistics, psychological statistics and sociological statistics. The operational definition of business statistics adopted herein is the latter.

Business statistics is an applied branch of mathematics that transforms data into useful information for decision-making purposes (Berenson et al. 2009a). The applications transcend the various functional areas of business – accounting (e.g., auditing), finance (e.g., portfolio development, ►risk analysis, forecasting, econometric modeling), information systems (e.g., database management, e-commerce analysis), management (e.g., project management, quality and productivity management) and marketing (e.g., consumer behavior, sales forecasting, market segmentation, conjoint analysis).

Owing to the wide variety of applications within these areas, business statistics is sometimes referred to as industrial statistics or economic statistics. However, the scope of such nomenclature is too narrow and the more encompassing definition of business statistics is preferred. Industrial statistics typically pertains to quality and productivity whereas economic statistics usually pertains to econometric modeling.

The key feature which distinguishes business statistics as a discipline distinct from other subject area applications of statistics is that in business two types of studies can be conducted, enumerative or analytic, and each has its own methodology. To distinguish between these two types of studies, consider a photograph in a frame versus a motion picture film played on a DVD. Whereas an enumerative study is a "snapshot at a single moment in time," an analytic study is a "motion picture taken over time."

Enumerative studies are common to other application-based branches of statistics and various methodologies employed transcend these statistics disciplines. Enumerative studies involve decision making regarding a population and/or its characteristics at a particular point in time. The sampled data collected can be analyzed descriptively and inferentially provided that appropriate probability sampling methods are used in the data collection stage. Inferences drawn through confidence interval estimation and regression modeling would be considered appropriate for that population at that moment in time. For example, a 95% confidence interval estimate based on a survey conducted today might show that adult males are between 7% and 12% more confident than are adult females that there will be improvements in global economic conditions next year, but who would believe that such opinions would remain static over time as events which led to such opinions change? Decisions made based on inferences drawn from enumerative studies hold true only as long as the population frames from which the samples were drawn remain unchanged.

On the other hand, analytic studies are specifically applicable to industrial or business situations (Deming

1986). Analytic studies involve taking some action on a process to improve future performance (Berenson et al. 2009a; Cryer and Miller 1991; Hoerl and Snee 2002). Data from a process are monitored over time and the focus of the study is to understand process behavior in order to make improvements to the process. Data collected typically constitute samples (i.e., *subgroups*) from some ongoing production process or service process obtained at regular intervals of time. The data can be analyzed descriptively but inferential analysis at this stage is usually limited to studying whether there are recognizable trends or other patterns in some critical-to-quality (*CTQ*) characteristic over time. By monitoring and understanding process variation, once a process is deemed "in-control," the objective is to improve the process either by enhancing the critical-to-quality characteristic's average and/or reducing process variation. This is done through inference – designing appropriate experiments based on the principle of ►randomization (Box et al. 1978). The fields of quality and productivity management, CQI/TQM, and Six-Sigma management have evolved from analytic studies.

A parallel to the analytic study described in managing a manufacturing or service process is time-series analysis (Roberts 1988) which, for many years, was used for either decomposing a series of data into its trend, seasonal or cyclical components in order to compare the series with others in the same industry, across industries or against movements in the overall economy or for forecasting the series into the future (Croxton and Cowden 1955). The fields of econometric modeling and forecasting have evolved from time-series analysis.

Evolution of Business Statistics Education

Through the 1950s teaching emphasis was mainly on descriptive statistical methods along with time series decomposition, forecasting, and index number construction (Croxton and Cowden 1955). Numbers-crunching was typically achieved through the use of desktop calculators. However, in the 1960s, as growth in computer technology permitted the storing of large data sets along with more rapid numbers-crunching capability, research using multivariate analysis methods that were developed in the 1930s through 1950s was finally plausible. Statistical software packages such as SPSS, SAS and BMDP, designed for research in the social sciences, psychology and medicine, were also adopted for use in business, particularly in marketing research and in finance. Thus, from the mid 1960s through the mid 1980s, business statistics education mainly focused on probability (Schlaifer 1961), sampling and, in particular, statistical inference.

In the mid 1980s, sparked by world-wide recognition of the achievements of W. Edwards Deming, the need to distinguish between enumerative and analytic studies in a business environment in which collected data are often longitudinal and not a random sample led to the development of annual MSMESB (Making Statistics More Effective in Schools and Business) conferences and resulted in the infusion of TQM into the business statistics curriculum (Tiao et al. 1986). Probability and inference were deemphasized. Following Deming's death in 1993, the TQM approach, which so successfully developed in Japan, was marred by financial failures and corrupt business practices in that country. Unfortunately, MSMESB was unable to sustain its earlier momentum regarding the importance of TQM in the business statistics course curriculum, particularly after TQM evolved into a Six-Sigma approach (see ►Six Sigma) which B-school faculty considered "too much" to handle in a 1-year, let alone one-semester introductory course.

The late 1990s and the first decade of the new millennium have witnessed an even more rapid expansion of information technology capability. Owing to increased computer literacy and the ability to use computer software to solve larger-scale problems, there has been a reduction in numbers-crunching and more emphasis is now placed on analyzing computer-obtained results. In addition, advances in information technology have led to further advances in the data mining of enormous data bases and emphasis on regression model-building and other analytical tools for evaluating such large data sets is emerging in business statistics education.

The Future of Business Statistics Education

Approximately 20% of all undergraduate students in the US who currently take an introductory statistics course are business majors. Thus, as recognized at the very first MSMESB conference (Tiao et al. 1986) in 1986, it is essential that the course be practical and relevant, demonstrate applications to the functional areas of business, use real data, employ computer information technology rather than "numbers crunching" to get results, emphasize analysis and interpretation of results, and enhance critical thinking and problem-solving skills through written articulation of findings (Berenson et al. 2009b).

In addition, attention must be given to proper tabular and graphic presentation that assist in descriptive analysis and decision making. The importance of probability sampling and randomization in making inferences from sample statistics to population parameters in enumerative studies must be emphasized. And the course must provide an understanding of models useful for prediction and for

forecasting as well as an understanding of process management in analytic studies through control chart monitoring and the use of other quality tools (Berenson et al. 2009a).

Owing to the need for B-schools to produce graduates who will be able to make useful contributions to business in an ever dynamic global environment, the field of business statistics will continue to make important contributions by providing the tools and methods along with enhancing the quantitative and critical thinking skills of the future workforce. With continued advances in IT, and continued developments in data mining and methodology for handling and analyzing enormously large sets of data collected over time, emerging functional area disciplines in business such as business analytics and business informatics will likely subsume business statistics along with the fields of operations research, management information systems and supply chain management. Future students preparing for work in such emerging fields will have a tremendous opportunity to learn and contribute to the world at large.

About the Author

Dr. Mark L. Berenson is Professor, Department of Management and Information Systems, Montclair State University, USA, and Professor Emeritus, Statistics, Zicklin School of Business, Baruch College CUNY, USA. He is also a former President of the New York Metropolitan Area Chapter of the American Statistical Association. Over his academic career Professor Berenson has received several awards for excellence in teaching. He has authored and/or co-authored dozens papers and 11 books, including Basic Business Statistics: Concepts and Applications with David M. Levine and Timothy C. Krehbiel (11th edition, Prentice Hall, 2009), Business Statistics: A First Course with David M. Levine and Timothy C. Krehbiel (5th edition, Prentice Hall, 2010), and Statistics for Managers Using Microsoft Excel with David M. Levine, David F. Stephan and Timothy C. Krehbiel (6th edition, Prentice Hall, 2011).

Cross References

►Banking, Statistics in
►Business Forecasting Methods
►Business Intelligence
►Business Surveys
►Data Analysis
►Detection of Turning Points in Business Cycles
►Economic Statistics
►Index Numbers
►Industrial Statistics
►Insurance, Statistics in
►Multivariate Data Analysis: An Overview
►National Account Statistics
►SIPOC and COPIS: Business Flow–Business Optimization Connection in a Six Sigma Context
►Statistics Education
►Statistics: An Overview

References and Further Reading

Berenson ML, Levine DM, Krehbiel TC (2009a) Basic business statistics: concepts and applications, 11th edn. Prentice Hall, Upper Saddle River, NJ

Berenson ML, McKenzie J, Ord JK (2009b) Statistics in business schools: the future? In: Proceedings of the joint statistical meetings, Washington, DC, 2009

Box GEP, Hunter WG, Hunter JS (1978) Statistics for experimenters. Wiley, New York

Croxton FE, Cowden DJ (1955) Applied General Statistics, 2nd edn. Prentice Hall, Englewood Cliffs, NJ

Cryer J, Miller R (1991) Statistics for business: data analysis and modeling. PWS-Kent, Boston

Deming WE (1986) Out of the crisis. MIT Center for Advanced Engineering Study, Cambridge, MA

Hoerl R, Snee RD (2002) Statistical thinking: improving business performance. Duxbury, Pacific Grove, CA

Roberts HV (1988) Data analysis for managers with minitab. Scientific Press, Redwood City, CA

Schlaifer R (1961) Introduction to statistics for business decisions. McGraw-Hill, New York

Tiao G, Roberts HV, Easton G, organizers (1986) First annual conference on making statistics more effective in schools and business (MSMESB). The University of Chicago, June 20–21, 1986

Business Surveys

Ger Snijkers[1,2], Mojca Bavdaž[3]
[1]Professor in Business Survey Methodology at Utrecht University, Utrecht, The Netherlands
[2]Senior Researcher in Business Survey Data Collection Methodology at Statistics, The Netherlands
[3]Assistant Professor at the Faculty of Economics University of Ljubljana, Ljubljana, Slovenia

The label *business survey* is typically attached to surveys that concern organizations involved in business activities such as manufacturing, commerce, finance, and other services. Business surveys may be considered a subgroup of establishment surveys. Establishment surveys refer to any formal organization engaged in any kind of productive activities, which also includes schools, hospitals, prisons, and other types of institutions that always or prevalently lack a lucrative purpose.

Surveying businesses serves several *aims*. Governments use business surveys to collect data for *economic indicators* within the system of national accounts

(e.g., for gross domestic product) and other areas of economic and ▶business statistics. The collected data are quantitative rather than categorical data, and among the quantitative data the continuous type tends to be more frequent than discrete (Cox and Chinnappa 1995). Some examples are revenues, costs, value of imported and exported goods, number of employees, amount of produced goods, and energy consumption. The most important statistical parameters include totals, structures, and averages, for example, total production or total employment, the structure of businesses by size, and average wages. Changes in these statistics may even be more interesting for the conduct of economic policies (Rivière 2002). This is why panel or longitudinal sampling designs and recurring surveys are often used in such business surveys. In addition, academic researchers use business surveys to collect data for *theoretical models* that often describe complex structures and relationships, for example, organizational structures, the relationship between business innovativeness and revenue growth, the relationship between atmosphere at work and working at home, etc. Chambers of commerce and industry, boards of trade, commodity and industrial boards, and similar organizations representing businesses also use business surveys when they are interested in getting data from their members. Finally, there are business research methods (e.g., Blumberg et al. 2008) that are used by businesses to support their *internal decision making* related to business-to-business relations and their market position.

The *nature of participation* in business surveys depends on their aims. In Europe, business surveys conducted by governmental institutions are typically mandatory, which means that sanctions may be used in case of noncompliance. In the US, this also applies to many but not all business surveys and censuses carried out by the Census Bureau. Mandatory business surveys usually achieve a response rate of 70% or more, often after a series of reminders. Voluntary business surveys not conducted by governmental statistical agencies achieve much lower response rates. These may be slightly higher for academic surveys (e.g., Baruch and Holtom (2008) calculated an average around 35%) compared to commercial surveys.

Business surveys have many unique characteristics given specific characteristics of the business population. The business population is volatile (this is particularly true of small enterprises) and difficult to track due to mergers, acquisitions, numerous forms of linkages, and cooperation. Various business units beyond administrative ones may constitute the most appropriate and meaningful *unit of observation* depending on the survey objectives: an enterprise, a local unit, a kind-of-activity unit etc., but they may not be simple to determine in practice. The choice, definition, and identification of the unit of observation are of paramount importance as they may have a considerable impact on estimates, in particular on breakdowns by activity and region. A public business register can support this complex task and help to construct the survey frame, particularly if it builds on standard statistical units, provides common background variables (at least main economic activity, size indicator, and location), uses standard classifications for these variables (see, for instance, the UN Classifications Registry ▶unstats.un.org/unsd/cr/registry/ and the latest version of the International Standard Industrial Classification of all Economic Activities (ISIC)), and is regularly updated. This reduces frame errors like omissions (e.g., start-ups), erroneous inclusions (e.g., bankrupt companies), duplications (e.g., change of the business name), and misclassifications (e.g., change of the main economic activity). This is why the construction, maintaining, and updating of business registers has traditionally been given a lot of attention in business survey methodology (see, for instance, Cox et al. 1995).

Businesses are also heterogeneous in terms of their activity, behavior, and size. Many variables in business surveys have skewed distributions. Larger businesses in the economy (or in an activity or a region) may therefore be selected in the sample at all times and receive special attention in statistics production. Availability of background variables and a larger weight of some businesses make stratified sampling and probability-proportional-to-size typical *sampling methods* for business surveys (see ▶Sample survey methods).

Questionnaires in business surveys often resemble forms with item labels. Item labels substitute survey questions; they often consist of technical terms, are accompanied by detailed instructions, and request data kept in business records, which all make the response task quite burdensome and time-consuming, and call for self-administered modes of data collection. The trend in governmental business surveys is towards mixed-mode designs in which paper questionnaires are complemented with electronic ones that can either be completed on-line or downloaded to be completed off-line (see ▶Internet survey methodology – recent trends and developments). Both paper and electronic questionnaires should be designed in such a way that it serves the response process within businesses (see e.g., Snijkers et al. 2007).

In business surveys, more than one person may be necessary to reach the decision on survey participation and provide survey answers because authority, capacity, and motivation to respond rarely reside in

one single person within the organization. Organizations have several characteristics that influence their *decision on survey participation* such as centralization of decision making, fragmentation of knowledge through specialization and founding of subsidiaries and branch plants, boundary-spanning roles, environmental dependence, etc. (Tomaskovic-Devey et al. 1994; Snijkers et al. 2007). Some factors influencing this decision are under researchers' control while others are not. The former, for instance, include the selection of the mode, the questionnaire design, and the contact strategies (e.g., introducing the survey, motivating respondents to participate, and respondent chasing); the latter include the selection of the actual respondent(s), internal business factors (e.g., policy not to participate in surveys), and factors in external environment such as survey-taking climate, political and economic climate, and legal and regulatory requirements (Willimack et al. 2002; Snijkers 2008). Paying attention to these factors, even taking the factors that are out of control into account in the survey design, may help prevent or reduce non-response in business surveys.

As in other surveys, the measurement process culminates in the *process of responding* to survey questions. Tourangeau (1984) pointed to four *cognitive components* of this process: understanding of the survey question, retrieval of information from memory, judgment of retrieved pieces of information, and communication of the response (see ►Sample survey methods). While the essence of the response process stays the same also in business surveys, many specifics shape this process considerably. Several people are often involved in the process: in addition to the respondent, also referred to as a reporter or an informant that indeed acts as a business representative or spokesperson, there may be other *business participants* such as gatekeepers (e.g., boundary-spanning units, receptionists), response coordinators, data providers, and authorities who also may serve as gatekeepers. When the required data and/or knowledge to extract them are dispersed across the business, response coordinators organize the survey task. Respondents retrieve the required data themselves or have data providers to retrieve data for them. Authorities have an important role in providing a mandate to work on the ►questionnaire, delegating and scheduling the survey response task and in authorizing the survey response.

Another specific of business surveys is heavy reliance on data stored in business records. When the requested data are not (readily) available in business records, the response burden increases. If the survey design is tailored to internal business processes, the response burden is likely to decrease, which affects the occurrence of non-response and measurement errors (Dale and Haraldsen 2007; Snijkers 2008).

Complexity of data collection in business surveys gave rise to several models of response processes. The latest contributions are the *hybrid model* by Willimack and Nichols (2010), the model introducing the idea of *socially distributed cognition* by Lorenc (2006), the response model focusing on motivating people and influencing their response behavior by Snijkers (2008), and the *multidimensional integral business survey response* (*MIBSR*) *model* by Bavdaž (2010b).

Quality assessment of business surveys may be based on quality frameworks developed by statistical organizations (e.g., IMF, Eurostat, Statistics Canada, Australian Bureau of Statistics, etc.), especially if these surveys provide data for official statistics. None of these frameworks can do without the accuracy of statistical data. Accuracy can be assessed using the concept of the total survey error (see ►Total survey error) even though some specifics may apply especially to non-sampling errors (see ►Non-sampling errors in surveys). Specification errors, for instance, may be quite large because business surveys attempt to measure complex economic phenomena. Frame errors depend on the existence and quality of a business register; these errors do not only refer to under coverage and over coverage but also to duplications and misclassifications. When several units within a business constitute units of observation, businesses may not report data for all of them, which results in within-business non-response; when units of observation differ from existing organisational units, businesses may report data for wrong units, which results in measurement errors. Given the specifics and complexity of the response process in business surveys, sources of measurement errors include all business participants and survey staff involved in the response process, the business environment, the survey instrument, and various survey characteristics and procedures beyond the mode of data collection, in particular the recurrence of the response process (Bavdaž 2010a). For many items in business survey questionnaires, it is difficult to distinguish between item non-response and a valid zero value. Non-response and measurement errors may be reduced in the data editing process. Given the usual abundance of auxiliary information taken from previous data reporting in the same survey, other surveys, or administrative sources, data editing has traditionally been given a lot of attention in business surveys. This is why data processing errors may be relatively small.

Since the 1990s, business survey methodology has been increasingly given a systematic approach in order to solve business surveys' unique problems (Cox et al. 1995). Up

to now, three International Conferences on Establishments Surveys (ICES) have been organized in 1993, 2000 and 2007, and the fourth one is coming up (2012). Other initiatives also serve this goal, for example, the International Workshop on Business Data Collection Methodology (www.ssb.no/bdcmethods) and the European Establishment Statistics Workshop (www.enbes.org). In Europe, Eurostat has started a major program on Modernisation of European Enterprise and Trade Statistics (MEETS; Behrens 2009). These initiatives try to provide a platform for business survey methodologists to discuss their problems, promote exchange of ideas and international collaboration, foster research in this field, and come to current best practices.

Cross References

►Business Statistics
►Internet Survey Methodology: Recent Trends and Developments
►Sample Survey Methods

References and Further Reading

Baruch Y, Holtom BC (2008) Survey response rate levels and trends in organizational research. Hum Relat 61(8):1139–1160

Bavdaž M (2010a) Sources of measurement errors in business surveys. J Official Stat 26(1):25–42

Bavdaž M (2010b) The multidimensional integral business survey response model. Surv Methodol 36(1):81–93

Behrens A (2009) Modernisation of European Enterprise and Trade Statistics (MEETS). Paper presented at the international conference on new techniques and technologies for statistics, Brussels, Belgium, 18–20 February 2009 (epp.eurostat.ec.europa.eu/portal/page/portal/research methodology/NTTS 2009)

Blumberg B, Cooper D, Schindler PS (2008) Business research methods: second European edition. McGraw Hill, London

Cox BG, Chinnappa BN (1995) Unique features of business surveys. In: Cox BG et al (eds) Business survey methods. Wiley, New York, pp 1–17

Cox BG, Binder DA, Chinnappa BN, Christianson A, Colledge MJ, Kott PS (eds) (1995) Business survey methods. Wiley, New York

Dale T, Haraldsen G (eds) (2007) Handbook for monitoring and evaluating business survey response burden. Eurostat, Luxembourg

Lorenc B (2006) Two topics in survey methodology: modelling the response process in establishment surveys; inference from non-probability samples using the double samples setup. Doctoral dissertation, Stockholm University

Rivière P (2002) What makes business statistics special? Int Stat Rev 70(1):145–159

Snijkers G (2008) Getting data for business statistics: a response model. In: Proceeding of Q2008; European conference on quality in official statistics, Rome, 8–11 July 2008

Snijkers G, Berkenbosch B, Luppes M (2007) Understanding the decision to participate in a business survey. In: Proceedings of ICES-III (3rd international conference on establishment surveys), Montreal, Canada, 18–21 June 2007, American Statistical Association, Alexandria, pp 1048–1059

Snijkers G, Onat E, Vis-Visschers R (2007) The annual structural business survey: developing and testing an electronic form. In: Proceedings of ICES-III (3rd international conference on establishment surveys), Montreal, Canada, 18–21 June 2007, American Statistical Association, Alexandria, pp 456–463

Tomaskovic-Devey D, Leiter J, Thompson S (1994) Organizational survey nonresponse. Adm Sci Q 39(3):439–457

Tourangeau R (1984) Cognitive science and survey methods. In: Jabine TB, Straf M, Tanur JM, Tourangeau R (eds) Cognitive aspects of survey methodology: building a bridge between disciplines. National Academy Press, Washington, pp 73–100

Willimack D, Nichols E (2010) A hybrid response process model for business surveys. J Official Stat 26(1):3–24

Willimack D, Nichols E, Sudman S (2002) Understanding unit and item nonresponse in business surveys. In: Groves RM, Dillman DA, Eltinge JL, Little RJA (eds) Survey nonresponse. Wiley, New York, pp 213–227

C

Calibration

Christos P. Kitsos
Professor and Head
Technological Educational Institute of Athens, Athens, Greece

Various methods and different (linear or not, simple linear, or multivariate) models have been adopted in industry to address the calibration problem. In practice, most of the models attempt to deal with the simple linear calibration technique, mostly applied in chemical applications, especially when some instruments are to be calibrated (examples include pH meters, NIR instruments, and establishing calibration graphs in chromatography).

The early work of Shukla (1972) put forward the problem on the real statistical dimensions, and even early on it was realized that when a non-linear model describes the phenomenon (Schwartz 1978), a linear approximation is eventually adopted. But even so, in the end we come to a nonlinear function to be estimated as best as possible (Kitsos and Muller 1995). When the variance of the measurement is due to many sources of variability, different techniques are used. Statistical calibration has been reviewed by Osborn (1991), who provides a list of pertinent references; when a robust approach might be appropriate, see Kitsos and Muller (1995). Certainly, to consider the variance constant and to follow a statistical quality control method (see ▶Statistical Quality Control), Hochberg and Marom (1983) might be helpful, but not in all cases. For the multivariate case, see the compact book of Brown (1993), Brereton (2000), and for an application Oman and Wax (1984). Moreover, different methods have been used on the development of the calibration problem like cross-validation (see Clark 1980).

Next we briefly introduce the statistical problem and the optimal design approach is adopted in the sequence to tackle the problem.

Consider the simple regression model with

$$n = E(y|u) = \theta_0 + \theta_1 u_1 \quad u_1 \in U = [-1,1]$$

where U is the design space, which can always be transformed to $[-1,1]$. Moreover, the involved error is assumed to be from the normal distribution with mean zero and variance $\sigma^2 > 0$.

The aim is to estimate the value of $u_1 = u_0$ given $n = C$, i.e.,

$$u_0 = (C - \theta_0)/\theta_1$$

which is a nonlinear function of the involved linear parameters, as we have already emphasized above.

The most well-known competitive estimators of u_0 when y_0 is provided are the so-called "*classical predictor*"

$$C(u_0) = \bar{x} + \frac{S_{xx}}{S_{xy}}(\dot{y}_0 - \bar{y})$$

and the "*inverse predictor*"

$$I(u_0) = \bar{u} + \frac{S_{xy}}{S_{yy}}(\dot{y}_0 - \bar{y})$$

with:

$$S_{\text{tr}} = \sum (t_i - \bar{t})(r_i - \bar{r})$$

were by $\dot{y}_0$ we mean the average of the possible k observations taken at the prediction stage (or experimental condition) and $\bar{y}$ as usually the average of the collected values.

The comparison of $C(u_0)$ and $I(u_0)$ is based on the values of the sample size n and the proportion $|\sigma/\theta_1|$ under the assumption that x_0 belongs to the experimenter area.

One of the merits of $C(u_0)$ is that when the usual normal assumption for the errors is imposed, the classical predictor is the maximum likelihood estimator. Moreover, $C(u_0)$ is a consistent estimator while $I(u_0)$ is inconsistent. The $I(u_0)$ estimation is criticized as it provides a confidence interval that might be the whole real line or, in the best case, two disjoint semi-infinite intervals. When $|\sigma/\theta_1|$ is small the asymptotic mse (mean square error) of $C(u_0)$ is smaller than with the use of $I(u_0)$, when x_0 does not lie in the neighborhood of $\bar{u}$.

The main difficulty is the construction of confidence intervals, as the variance of u_0 does not exist. This provides food for thought for an optimal design approach for the calibration problem. To face these difficulties the

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

optimal experimental approach is adopted (see *Optimum Experimental Designs*, also see Kitsos 2002).

For the one-stage design we might use of the criterion function Φ, either D-optimality for (θ_0, θ_1) or c-optimality for estimating u_0. The D-optimal design is of interest because its effectiveness can be investigated, as measured by the c-optimality criterion. Under c-optimality, thanks to Elfving's theorem, locally optimal two-point design can be constructed geometrically. The criterion that experimenters like to use is

$$\min Var(\hat{u}_0).$$

Different approaches have been adopted for this crucial problem: Bayesian, see Chaloner and Verdinelli (1995), Hunter and Lamboy (1981); structural inference, see Kalotay (1971). There is a criticism that structural inference is eventually Bayesian, but this is beyond the scope of this discussion.

When suitable priors for u_0 are chosen the calibrative density functions come from the non-central Student with mean $\mathrm{Ba}(u_0)$ as

$$\mathrm{Ba}(u_0) = \bar{u} + \frac{S_{yy}}{r}(\dot{y}_0 - \bar{y})$$

where $r = S_{yy} + \sum_j^k (y_{0_j} - \bar{y})^2$. When $k = 1$ the Bayesian estimator coincides with the inverse, namely $\mathrm{Ba}(u_0) = I(u_0)$.

The structural approach forms the simple linear model as a "structural model" and obtains a rather complicated model, which, again, with $k = 1$, coincides with the inverse regression.

The nonlinear calibration has attracted classical and Bayesian approaches, both based on the Taylor expansion of the nonlinear model. Therefore, calibration is based on the linear approach of the nonlinear model.

About the Author

Dr. Christos Kitsos is a Professor and Head, Department of Mathematics, of the Technological Educational Institute of Athens, Greece. His first degree is in mathematics, Athens University, his MA degree from the Math and Stat of University of New Brunswick, Canada, while his PhD is from the Glasgow University, UK. He is a member of ISI (and ISI-Committee of Risk Analysis), IEEE and was elected member of IASC, were he is also a member. He has organized a number of statistical conferences in Greece, and has founded the series of International Conference on Cancer Risk Assessment (ICCRA). He has published 88 papers in journals and proceedings volumes, participated in 80 conferences, and published 12 books in Greek (8 are textbooks), and is a coauthor of 5 international books as editor. Professor Kitsos has been the national representative at EUROSTAT and OECD for educational statistics.

Cross References

►Chemometrics
►Measurement Error Models
►Optimal Regression Design
►Optimum Experimental Design

References and Further Reading

Brereton GR (2000) Introduction to multivariate calibration in analytical chemistry. Analyst 125(11):2125–2154

Brown JP (1993) Measurement, regression and calibration. Oxford Science Publication, Oxford Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. Stat Sci 10:273–304

Clark RM (1980) Calibration, cross-validation and carbon-14. II. J Roy Stat Soc, Ser A 143:177–194

Frank IE, Friedman JH (1993) A Statistical view of some chemometrics regression tools. Technometrics 35:109–148. With discussion

Hochberg Y, Marom I (1983) On improved calibrations of unknowns in a system of quality-controlled assays. Biometrics 39:97–108

Hunter WG, Lamboy WFA (1981) Bayesian analysis of the linear calibration problem. Technometrics 23:323–338

Kalotay AJ (1971) Structural solution to the linear calibration problem. Technometrics 13:761–769

Kanatani K (1992) Statistical analysis of focal-length calibration using vanishing points. IEEE Trans Robotic Autom 8:767–775

Kitsos CP (1992) Quasi-sequential procedures for the calibration problem. In Dodge Y, Whittaker J (eds) COMPSTAT 1992, vol 2. Physica-Verlag, Heidelberg, pp 227–231

Kitsos CP (2002) The simple linear calibration problem as an optimal experimental design. Commun Stat – Theory Meth 31:1167–1177

Kitsos CP, Muller Ch (1995) Robust linear calibration. Statistics 27:93–106

Oman SD, Wax Y (1984) Estimating fetal age by ultrasound measurements: an example of multivariate calibration. Biometrics 40:947–960

Osborne C (1991) Statistical calibration: a review. Int Stat Rev 59:309–336

Schwartz LM (1978) Statistical uncertainties of analyses by calibration of counting measurements. Anal Chem 50:980–985

Shukla GK (1972) On the problem of calibration. Technometrics 14:547–553

Canonical Analysis and Measures of Association

Jacques Dauxois[1], Guy Martial Nkiet[2]
[1]Professor
Institut de Mathématiques de Toulouse, Toulouse, France
[2]Professor
Université des Sciences et Techniques de Masuku, Franceville, Gabon

Introduction

The Bravais–Pearson linear correlation coefficient and the Sarmanov maximal coefficient are well known statistical

tools that permit to measure, respectively, correlation (also called linear dependence) and stochastic dependence of two suitable random variables X_1 and X_2 defined on a probability space $(\Omega, \mathcal{A}, P)$. Since these coefficients just are the first canonical coefficients obtained from linear and nonlinear canonical analysis, respectively, it is relevant to improve them by using all the canonical coefficients. In order to give an unified framework for these notions, we introduce the canonical analysis (CA) of two closed subspaces H_1 and H_2 of a Hilbert space H. Then, a class of measures of association that admits the aforementioned coefficients as particular cases can be constructed.

Canonical Analysis

Let H be a separable real Hilbert space with inner product and related norm denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ respectively, and H_1 and H_2 be two closed subspaces of H. Then we have the following definition that comes from Dauxois and Pousse (1975).

Definition 1 *The canonical analysis (CA) of H_1 and H_2 is any triple*

$$\left(\{\rho_i\}_{i\in I_0}, \{f\}_{i\in I_1}, \{g\}_{i\in I_2}\right),$$

with $I_\ell \subset \mathbb{N}^*$ for $\ell \in \{0,1,2\}$, that satisfies:

1. *The system $\{f\}_{i\in I_1}$ (resp. $\{g\}_{i\in I_2}$) is an orthonormal basis of H_1 (resp. H_2)*
2. *$\rho_1 = \langle f_1, g_1\rangle = \sup\{\langle f, g\rangle; (f,g) \in H_1 \times H_2,\ \|f\| = \|g\| = 1\}$*
3. *For any $i \in I_0$ such that $i \geq 2$, one has:*

$$\rho_i = \langle f_i, g_i\rangle = \sup\left\{\langle f, g\rangle; (f,g) \in F_i^{\perp} \times G_i^{\perp}, \|f\| = \|g\| = 1\right\}$$

where $F_i = span\{f_1, \cdots, f_{i-1}\}$ and $G_i = span\{g_1, \cdots, g_{i-1}\}$.

Conditions for existence of canonical analysis have been investigated in the aforementioned work. More precisely, denoting by Π_E the orthogonal projector onto the closed subspace E of H, a sufficient condition is the compacity of $T_1 = \Pi_{H_1}\Pi_{H_2|H_1}$ or, equivalently, that of $T_2 = \Pi_{H_2}\Pi_{H_1|H_2}$. In this case, we say that we have a compact CA, and the following proposition holds:

Proposition 1 *Consider a compact CA $\left(\{\rho_i\}_{i\in I_0}, \{f\}_{i\in I_1}, \{g\}_{i\in I_2}\right)$, of H_1 and H_2, where the ρ_i's are arranged in nonincreasing order. Then:*

1. *$\{\rho_i^2\}_{i\in I_0}$ is the noincreasing sequence of eigenvalues of T_1 and T_2 and, for any $i \in I_0$, one has $0 \leq \rho_i \leq 1$.*
2. *$\{f\}_{i\in I_1}$ (resp. $\{g\}_{i\in I_2}$) is an orthonormal basis of H_1 (resp. H_2) such that, for any $i \in I_0$, f_i (resp. g_i) is an eigenvector of T_1 (resp. T_2) associated with ρ_i^2.*
3. *$\forall (i,j) \in (I_0)^2$, $\langle f_i, g_j\rangle = \delta_{ij}\rho_i$, $\Pi_{H_1} f_i = \rho_i g_i$, $\Pi_{H_2} g_i = \rho_i f_i$.*
4. *$\{f\}_{i\in I_1 - I_0}$ (resp. $\{g\}_{i\in I_2 - I_0}$) is an orthonormal basis of $\ker(T_1) = H_1 \cap H_2^{\perp}$ (resp. $\ker(T_2) = H_2 \cap H_1^{\perp}$).*

Remark 1

1. The ρ_i's are termed the *canonical coefficients*. They permit to study the relative postions of each of the preceding subspace with respect to the other. For instance, the nullity of all these coefficients is equivalent to the orthogonality of H_1 and H_2, and if one of these subspaces is included into the other these coefficients are all equal to 1. Note that, in this later case there does not exist a compact CA when the considered subspaces are infinite-dimensional ones. Nevertheless, it is possible to find a triple having the same properties than a compact CA. Such a triple can be given by $\left(\mathbb{1}, (f_i)_{i\in\mathbb{N}^*}, (g_i)_{i\in\mathbb{N}^*}\right)$, where $\mathbb{1}$ is the numerical sequence with all terms equal to 1, $(f_i)_{i\in\mathbb{N}^*}$ is a orthonormal basis of H_1 and $(g_i)_{i\in\mathbb{N}^*}$ is the previous system possibly completed with an orthonormal basis of $\ker T_2 = H_2 \cap H_1^{\perp}$ so as to obtain an orthonormal basis of H_2.
2. From the previous notion of CA it is possible to define a canonical analysis of two subspaces H_1 and H_2 relatively to a third one H_3. It is just the CA of the subspaces $H_{1\cdot 3} := (H_1 \oplus H_3) \cap H_3^{\perp}$ and $H_{2\cdot 3} := (H_2 \oplus H_3) \cap H_3^{\perp}$. This CA leads to interesting properties given in Dauxois et al. (2004a), and is useful in statistics for studying conditional independence between random vectors (see, e.g., Dauxois et al. [2004b]).
3. When $X_1 = \left(X_1^1, \cdots, X_1^{p_1}\right)^T$ and $X_2 = \left(X_2^1, \cdots, X_2^{p_2}\right)^T$ are two random vectors such that any X_i^j belongs to $L^2(P)$, their *Linear Canonical Analysis* (LCA) is the CA of H_1 and H_2 where $H_i = \text{span}\left(X_i^1, \cdots, X_i^{p_i}\right)$. The spectral analysis of T_1 is equivalent to that of $V_1^{-1}V_{12}V_2^{-1}V_{21}$, where V_i (resp. V_{12}; resp. V_{21}) denotes the covariance (resp. cross-covariance) operator of X_i (resp. X_1 and X_2; resp. X_2 and X_1). So, it is just the CA of random vectors introduced by Hotelling (1936). The first canonical coefficient is the Bravais–Pearson linear correlation coefficient.
4. When X_1 and X_2 are arbitrary random variables, their *Nonlinear Canonical Analysis* (NLCA) is the CA of H_1 and H_2 where H_i is the closed subspace of $L^2(P)$ consisting in random variables of the form $\varphi(X_i)$, where φ is a measurable function valued into $\mathbb{R}$. In this case, the first canonical coefficient just is the Sarmanov maximal coefficient.

Measures of Association

Let $\mathcal{C}$ be the set of pairs (H_1, H_2) of closed subspaces of a Hilbert space, having a compact CA or being infinite-dimensional and such that $H_1 \subset H_2$ or $H_2 \subset H_1$. We consider an equivalence relation $\simeq$ defined on $\mathcal{C}$, such that $(H_1, H_2) \simeq (E_1, E_2)$ if there exists a pair (I_1, I_2) of isometries satisfying: $I_1(H_1) = E_1$, $I_2(H_2) = E_2$ and $\forall (x,y) \in H_1 \times H_2$, $< I_1(x), I_2(y) >_E = < x, y >_H$, where

H (resp. E) denotes the separable real Hilbert space which contains H_1 and H_2 (resp. E_1 and E_2). We also consider a preordering relation $\preceq$ on $\mathcal{C}$, such that $(H_1, H_2) \preceq (E_1, E_2)$ if there exists a pair $\left(E_1', E_2'\right)$ of closed subspaces satisfying: $E_1' \subset E_1$, $E_2' \subset E_2$ and $(H_1, H_2) \simeq \left(E_1', E_2'\right)$.

Definition 2 *A measure of association r between Hilbertian subspaces is any map from a subset $\mathcal{C}_r$ of $\mathcal{C}$ into $[0,1]$ such that the following conditions are satisfied:*

$$
\begin{aligned}
r(H_1, H_2) &= r(H_2, H_1);\\
H_1 \perp H_2 &\Leftrightarrow r(H_1, H_2) = 0;\\
H_1 \subset H_2 \quad \text{or} \quad H_1 \subset H_2 &\Rightarrow r(H_1, H_2) = 1;\\
(H_1, H_2) \simeq (E_1, E_2) &\Rightarrow r(H_1, H_2) = r(E_1, E_2);\\
(H_1, H_2) \preceq (E_1, E_2) &\Rightarrow r(H_1, H_2) \leq r(E_1, E_2).
\end{aligned}
$$

Remark 2

1. When $X_1 = \left(X_1^1, \cdots, X_1^{p_1}\right)^T$ and $X_2 = \left(X_2^1, \cdots, X_2^{p_2}\right)^T$ are two random vectors such that any X_i^j belongs to $L^2(P)$, we obtain a measure of linear dependence between X_1 and X_2 by putting $r(X_1, X_2) := r(H_1, H_2)$ with $H_i = \text{span}\left(X_i^1, \cdots, X_i^{p_i}\right)$. Indeed, from second axiom given above, $r(X_1, X_2) = 0$ if and only if X_1 and X_2 are uncorrelated, that is $V_{12} = 0$. From the third one, it is seen that if there exists a linear map A such that $X_1 = AX_2$ then $r(X_1, X_2) = 1$.
2. When X_1 and X_2 are arbitrary random variables, considering $H_i = \{\varphi(X_i) / \mathbb{E}(\varphi(X_i)^2) < +\infty\}$, a measure of stochastic dependence of X_1 and X_2 is obtained by putting $r(X_1, X_2) := r(H_1, H_2)$. In this case, the above axioms are closed to the conditions proposed by Rényi (1959) for good measures of dependence. In particular, the second axiom gives the equivalence between the independence of X_1 and X_2 and the nullity of $r(X_1, X_2)$, and from the third axiom it is seen that for any one to one and bimeasurable functions f and g, one has $r(f(X_1), g(X_2)) = r(X_1, X_2)$.

A class of measures of association can be built by using symmetric non decreasing functions of canonical coefficients. In what follows, $\mathcal{P}(\mathbb{N}^*)$ denotes the set of permutations of $\mathbb{N}^*$. For $\sigma \in \mathcal{P}(\mathbb{N}^*)$ and $x = (x_n)_n \in c_0$, we put $x_\sigma = \left(x_{\sigma(n)}\right)_n$ and $|x| = (|x_n|)_n$. We denote by c_0 the space of numerical sequences $x = (x_n)_n$ such that $\lim_{n\to\infty} x_n = 0$.

Definition 3 *A symmetric nondecreasing function (sndf) is a map Φ from a subset c_Φ of c_0 to $\mathbb{R}_+$ satisfying:*

1. *For all $x \in c_\Phi$ and $\sigma \in \mathcal{P}(\mathbb{N}^*)$, one has $x_\sigma \in c_\Phi$ and $\Phi(|x_\sigma|) = \Phi(x)$.*
2. *For all $(x, y) \in (c_\Phi)^2$, if $\forall n, |x_n| \leq |y_n|$, then $\Phi(x) \leq \Phi(y)$.*
3. *There exists a nondecreasing function f_Φ from $\mathbb{R}$ to $\mathbb{R}$ such that : $f_\Phi(0) = 0$; $\forall u \in \mathbb{R}$, $(u, 0, \cdots) \in c_\Phi$ and $\Phi(u, 0, \cdots) = f_\Phi(|u|)$.*

We denote by Ψ the map from $\mathcal{C}$ to $c_0 \cup \{\mathbb{1}\}$ such that $\Psi(H_1, H_2)$ is the noincreasing sequence of canonical coefficients of H_1 and H_2. Then we have:

Proposition 2 *Let Φ be a sndf with definition domain c_Φ, and such that $\Phi(\mathbb{1}) = 1$. Then, the map $r_\Phi = \Phi \circ \Psi$ is a measure of association defined on the subset $\mathcal{C}_\Phi = \{(H_1, H_2) \in \mathcal{C};\ \Psi(H_1, H_2) \in c_\Phi \cup \{\mathbb{1}\}\}$.*

This proposition means that a measure of association between two subspaces is obtained as a function of the related nonincreasing sequence of canonical coefficients through a sndf. Some examples of such measures are:

$$
\begin{aligned}
r_1(H_1, H_2) &= 1 - \exp\left(-\sum_{i=1}^{+\infty} \rho_i^2,\ r_{2,p}(H_1, H_2)\right)\\
&= \sqrt{\frac{\sum_{i=1}^{+\infty} \rho_i^{2p}}{1 + \sum_{i=1}^{+\infty} \rho_i^{2p}}}\ (p \in \mathbb{N}^*),\ r_3(H_1, H_2)\\
&= \max_{i\geq 1} |\rho_i| = \rho_1.
\end{aligned}
$$

On the one hand, this class of measures of association contains all the measures built by using LCA of random vectors (see Cramer and Nicewander (1979), Lin (1987), Dauxois and Nkiet (1997b)). On the other hand, when H_1 and H_2 are the subspaces considered in the second assertion of Remark 2, r_3 just is the Sarmanov maximal coefficient. In this case, estimation of the coefficients from NLCA and, therefore, the related measures of associations can be obtained from approximation based on step functions or B-spline functions, and from sampling. Using this approach, a class of independence tests that admits the chi-squared test of independence as particular case, have been be proposed (see Dauxois and Nkiet (1998)).

Cross References

►Canonical Correlation Analysis
►Multivariate Data Analysis: An Overview

References and Further Reading

Cramer EM, Nicewander WA (1979) Some symmetric invariant measures of multivariate association. Psychometrika 41:347–352

Dauxois J, Nkiet GM (1997a) Canonical analysis of Euclidean subspaces and its applications. Linear Algebra Appl 264:355–388

Dauxois J, Nkiet GM (1997b) Testing for the lack of a linear relationship. Statistics 30:1–23

Dauxois J, Nkiet GM (1998) Nonlinear canonical analysis and independence tests. Ann Stat 26:1254–1278

Dauxois J, Pousse A (1975) Une extension de l'analyse canonique. Quelques Applications. Ann Inst Henri Poincaré XI:355–379

Dauxois J, Nkiet GM, Romain Y (2004a) Canonical analysis relative to a closed subspace. Linear Algebra Appl 388:119–145

Dauxois J, Nkiet GM, Romain Y (2004b) Linear relative canonical analysis, asymptotic study and some applications. Ann Inst Stat Math 56:279–304

Hotelling H (1936) Relations between two sets of variables. Biometrika 28:321–377

Lin PE (1987) Measures of association between vectors. Commun Stat Theory Meth 16:321–338

Rényi A (1959) On measures of dependence. Acta Math Acad Sci Hung 10:57–71

Canonical Correlation Analysis

Tenko Raykov
Professor of Measurement and Quantitative Methods
Michigan State University, East Lansing, MI, USA

Introduction

Canonical correlation analysis (CCA) is one of the most general multivariate statistical analysis methods (see ►Multivariate Statistical Analysis). To introduce CCA, consider two sets of variables, denoted A and B for ease of reference (e.g., Raykov and Marcoulides 2008). Let A consist of p members collected in the vector $\underline{\mathbf{x}}$, and let B consist of q members placed in the vector $\underline{\mathbf{y}}$ $(p > 1, q > 1)$. In an application setting, the variables in either set may or may not be considered response variables (dependent or outcome measures) or alternatively independent variables (predictors, explanatory variables). As an example, A may consist of variables that have to do with socioeconomic status (e.g., income, education, job prestige, etc.), while B may comprise cognitive functioning related variables (e.g., verbal ability, spatial ability, intelligence, etc.).

Consider the correlation matrix R of all variables in A and B taken together, which has $(p + q) \cdot (p + q - 1)/2$ non-duplicated (non-redundant) correlations. Obviously, even for relatively small p and q, there are many non-duplicated elements of R. CCA deals with reducing this potentially quite large number of correlations to a more manageable group of interrelationship indices that represent the ways in which variables in A covary with variables in B, i.e., the interrelationships among these two sets of variables. More specifically, the purpose of CCA is to obtain a small number of derived variables (measures) from those in A on the one hand, and from those in B on the other, which show high correlations across the two sets (e.g., Johnson and Wichern 2002). That is, a main goal of CCA is to "summarize" the correlations between variables in set A and those in set B into a much smaller number of corresponding linear combinations of them, which in a sense are representative of those correlations. With this feature, CCA can be used as a method for (1) examining independence of two sets of variables (viz. A and B), (2) data reduction, and (3) preliminary analyses for a series of subsequent statistical applications.

Achieving this goal is made feasible through the following steps (cf. Raykov and Marcoulides 2008). First, a linear combination Z_1 of the variables $\underline{\mathbf{x}}$ in A is sought, as is a linear combination W_1 of the variables $\underline{\mathbf{y}}$ in B, such that their correlation $\rho_{1,1} = \mathrm{Corr}(Z_1, W_1)$ is the highest possible across all choices of combination weights for W_1 and Z_1 (see next section for further details). Call Z_1 and W_1 the *first pair of canonical variates*, and $\rho_{1,1}$ the *first canonical correlation*. In the next step, another linear combination Z_2 of variables in A is found and a linear combination W_2 of variables in B, with the following property: their correlation $\rho_{2,2} = \mathrm{Corr}(Z_2, W_2)$ is the highest possible under the assumption of Z_2 and W_2 being uncorrelated with the variables in the first combination pair, Z_1 and W_1. Z_2 and W_2 are referred to as the *second pair of canonical variates*, and $\rho_{2,2}$ as the *second canonical correlation*. This process can be continued until t pairs of canonical variates are obtained, where $t = \min(p, q)$ being the smaller of p and q. While in many applications t may be fairly large, it is oftentimes the case that only up to the first two or three pairs of canonical variates are really informative (see following section). If all canonical correlations are then uniformly weak and close to zero, A and B can be considered largely (linearly) unrelated. Otherwise, one could claim that there is some (linear) interrelationship between variables in A with those in B. Individual scores on the canonical variates can next be computed and used as values on new variables in subsequent analyses. These scores may be attractive then, since they capture the essence of the cross-set variable interrelationships.

Procedure

To begin a CCA, two linear combinations $Z_1 = \underline{\mathbf{a}}_1' \underline{\mathbf{x}}$ and $W_1 = \underline{\mathbf{b}}_1' \underline{\mathbf{y}}$ are correspondingly sought from the variables

in A and in B, such that $\rho_{1,1} = \mathrm{Corr}(Z_1, W_1)$ is at its maximal possible value across all possible choices of $\underline{\mathbf{a}}_1$ and $\underline{\mathbf{b}}_1$. Consider the covariance matrix S of the entire set of $p + q$ variables in A and B:

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

where S_{11} is the covariance matrix of the p variables in A, S_{22} that of the q variables in B, S_{21} that of the q variables in B with the p in A, and S_{12} denotes the covariance matrix of the p variables in A with the q measures in B. It can be shown (e.g., Johnson and Wichern 2002) that this maximum correlation $\rho_{1,1}$ will be achieved if the following holds:

1. $\underline{\mathbf{a}}_1$ is taken as the (generalized) eigenvector pertaining to the largest solution ρ^2 of the equation $\left|S_{12}S_{22}^{-1}S_{21} - \rho^2 S_{11}\right| = 0$, where $|.|$ denotes determinant, that is, $\underline{\mathbf{a}}_1$ fulfils the equation $\left(S_{12}S_{22}^{-1}S_{21} - \rho^2 S_{11}\right)\underline{\mathbf{a}}_1 = \underline{0}$, with ρ^2 being the largest solution of the former equation.
2. $\underline{\mathbf{b}}_1$ is the (generalized) eigenvector pertaining to the largest root of the equation $\left|S_{21}S_{11}^{-1}S_{12} - \pi^2 S_{22}\right| = 0$, that is, $\underline{\mathbf{b}}_1$ fulfils the equation $\left(S_{21}S_{11}^{-1}S_{12} - \pi^2 S_{22}\right)\underline{\mathbf{b}}_1 = \underline{0}$, with the largest π^2 satisfying the former equation.

The solutions of the two involved determinantal equations are identical, that is, $\rho^2 = \pi^2$, and the positive square root of the largest of them equals $\rho_{(1)} = \pi_{(1)} = \rho_{1,1} = \mathrm{Corr}(Z_1, W_1)$, the maximal possible correlation between a linear combination of variables in A with a linear combination of those in B. Then $Z_1 = \underline{\mathbf{a}}_1' \underline{\mathbf{x}}$ and $W_1 = \underline{\mathbf{b}}_1' \underline{\mathbf{y}}$ represent the first canonical variate pair, with this maximal correlation, $\mathrm{Corr}(Z_1, W_1)$, being the first canonical correlation.

As a next step, the second canonical variate pair is furnished as a linear combination of the variables in A, using the eigenvector pertaining to the second largest solution of $\left|S_{12}S_{22}^{-1}S_{21} - \rho^2 S_{11}\right| = 0$ on the one hand, and a linear combination of the B variables using the second largest solution of $\left|S_{21}S_{11}^{-1}S_{12} - \pi^2 S_{22}\right| = 0$ on the other hand; then their correlation is the second canonical correlation. One continues in the same manner until $t = \min(p, q)$ canonical variate pairs are obtained; the corresponding canonical correlations are calculated as their interrelationship indices (correlations). From the construction of the canonical variates follows that they are uncorrelated with one another:

$$\begin{aligned} \mathrm{Cov}(Z_i, Z_j) &= \mathrm{Cov}(W_i, W_j) = \mathrm{Cov}(Z_i, W_j) \\ &= 0 \ (\text{for all } i \neq j;\ i, j = 1, \ldots, t). \end{aligned}$$

Interpretation

Even though there are $t = \min(p, q)$ canonical variate pairs and canonical correlations, oftentimes in applications not all are important for understanding the relationships among variables in A and B. Statistical tests are available which help evaluate the importance of canonical variate pairs and aid a researcher in finding out how many pairs could be retained for further analysis. The tests assume multivariate normality and examine hypotheses of canonical correlations being 0 in a given population. The first test evaluates the null hypothesis that all canonical correlations are 0. If this hypothesis is rejected, at least the first canonical variate pair is of relevance when trying to understand the interrelationship between the variables in A and B; more specifically, at least the first canonical correlation is not zero in the population. Then the second test examines the null hypothesis that apart from the first canonical correlation, all remaining ones are 0; and so on. If the first tested hypothesis is not rejected, it can be concluded that A and B are (linearly) unrelated to one another.

After completing these tests, and in case at least the first canonical correlation is significant, the next question may well be how to interpret the canonical variates. To this end, one can use the correlations of each canonical variate with variables within its pertinent set. That is, when trying to interpret Z_1, one can look at its correlations with the variables in A. Similarly, when trying to interpret W_1, one can examine its correlations with the variables B; and so on for the subsequent canonical variate pairs and their members. The principle to follow thereby, is to interpret each canonical variate as representing the common features of initial variables correlated highly with that variate. Furthermore, for a given canonical correlation $\rho_i = \pi_i$, its square ρ_i^2 can be interpreted as a squared multiple correlation coefficient for the regression relating the ith canonical variate for any of the sets A or B, with the variables of the other set (B or A, respectively; $i = 1, \ldots, t$). With this in mind, ρ_i^2 can be viewed as proportion shared variance between A and B, as captured by the ith canonical variate pair ($i = 1, \ldots, t$); the square of the first canonical correlation is interpretable as a measure of "set overlap."

Similarly to principal components and factors, canonical variates can be used to obtain individual subject scores on them. They can be used in subsequent analyses, e.g., as scores on explanatory variables. Like principal components, the units of a canonical variate may not be meaningful. It is stressed that canonical variates are not latent variables, but instead share the same observed status as manifest (recorded) variables, since they are linear combinations of the latter.

Relationship to Discriminant Function Analysis

It can be shown (Tatsuoka 1971) that with $k > 2$ groups discriminant function analysis (DFA) is identical to CCA using additionally defined variables $D_1, D_2, \ldots, D_{k-1}$ as comprising set A, while the original explanatory (predictor) variables, say $\underline{\mathbf{x}} = (x_1, x_2, \ldots, x_p)'$, are treated as set B. These ▶dummy variables $D_1, D_2, \ldots, D_{k-1}$ are defined in exactly the same way they would be for purposes of regression analysis with categorical predictors. If one then performs a CCA with these sets A and B, the results will be identical to those obtained with a DFA on the original variables $\underline{\mathbf{x}}$. Specifically, the first canonical variate for B will equal the first discriminant function; the second canonical variate for B will equal the second discriminant function, etc. The test for significance of the canonical correlations is then a test for significance of discriminant functions, and the number of significant such functions and of canonical correlations is the same. Further, each consecutive eigenvalue for the discriminant criterion, v_i, is related to a corresponding generalized eigenvalue (determinantal equation root) ρ_i : $v_i = \frac{\rho_i^2}{1-\rho_i^2}$ ($i = 1, 2, \ldots, r$; Johnson and Wichern 2002). Testing the significance of discriminant functions is thus equivalent to testing significance of canonical correlations.

Generality of Canonical Correlation Analysis

CCA is a very general multivariate statistical method that unifies a number of analytic approaches. The canonical correlation concept generalizes the notion of bivariate correlation that is a special case of the former for $p = q = 1$ variables. The multiple correlation coefficient of main relevance in regression analysis is also a special case of canonical correlation, which is obtained when the set A consists of $p = 1$ variable – the response measure – and the set B consists of q variables that are the predictors (explanatory variables) in the pertinent regression model. The multiple correlation coefficient is then identical to the first canonical correlation. Third, since various uni- and multivariate ANOVA designs can be obtained as appropriate special cases of regression analysis, these designs and corresponding ANOVAs can be seen as special cases of canonical correlation analysis. Also, as indicated in the preceding section, discriminant function analysis is a special case of CCA as well. (Since DFA is a "reverse" MANOVA – e.g., Raykov and Marcoulides 2008 – one can alternatively see the latter also as a special case of CCA.) Hence, canonical correlation analysis is a very general multivariate analysis method, which subsumes a number of others that are widely used in statistical applications.

About the Author

Tenko Raykov is a Professor of Measurement and Quantitative Methods at Michigan State University. He received his Ph.D. in Mathematical Psychology from Humbolt University in Berlin. He is an editorial board member of the *Structural Equation Modeling*, *Multivariate Behavioral Research*, *Psychological Methods* and the *British Journal of Mathematical and Statistical Psychology*. He is a coauthor (with G.A. Marcoulides) of the text *A First Course in Structural Equation Modeling* (Lawrence Erlbaum Associates, 2006), *An Introduction to Applied Multivariate Analysis* (Routledge 2008), and *Introhhduction to Psychometric Theory* (Routledge 2010).

Cross References

- ▶Analysis of Multivariate Agricultural Data
- ▶Canonical Analysis and Measures of Association
- ▶Discriminant Analysis: An Overview
- ▶Eigenvalue, Eigenvector and Eigenspace
- ▶Multivariate Data Analysis: An Overview
- ▶Multivariate Statistical Analysis

References and Further Reading

Johnson RA, Wichern DW (2002) Applied multivariate statistical analysis. Prentice Hall, Upper Saddle River

Raykov T, Marcoulides GA (2008) An introduction to applied multivariate analysis. Taylor & Francis, New York

Tatsuoka MM (1971) Multivariate analysis: techniques for educational and psychological research. Wiley, New York

Careers in Statistics

Daniel R. Jeske[1], Janet Myhre[2]
[1]Professor and Chair
University of California-Riverside, Riverside, CA, USA
[2]MARC and Professor Emeritus
Claremont McKenna College, Claremont, CA, USA

Statistics has changed over the last decades from being a discipline that primarily studied ways to characterize randomness and variation to a discipline that emphasizes the importance of data in the explanation of phenomenon and in problem solving. While statisticians routinely use mathematics and computer programming languages as key

tools in their work, they usually also function as an important data-driven decision maker within their application domain. Consequently, a statistician must have a genuine curiosity about the subject domain they work within, and furthermore, must have strong collaborative and communication skills in order successfully interact with the many colleagues they will encounter and rely on for information.

As the world becomes more quantitative through the data revolution, more professions and businesses depend on data and on the understanding and analyses of these data. Data are not simply numbers. Data contain information that needs to be understood and interpreted. As a result, statisticians are much more than bean counters or number crunchers. They possess skills to find needles in haystacks and to separate noise from signal. They are able to translate a problem or question into a framework that enables data collection and data analysis to provide meaningful insights that can lead to practical conclusions.

Loosely speaking there is a spectrum of statisticians that ranges from very applied on one end to very theoretical on the other end. Applied statisticians skillfully select and implement an appropriate statistical methodology to solve a problem. They are a statistician who has a problem and is looking for a solution. Theoretical statisticians are interested in trying to expand the toolkit of applied statisticians by generalizing or creating new methods that are capable of solving new problems or solving existing problems more efficiently. They are statisticians who might get motivated by a problem someone else encountered in practice, but who then abstract the problem as much as possible so that their solution has as broad an impact as possible. Most statisticians are not planted firmly on either end of this spectrum, but instead find themselves moving around and adapting to the particular challenge they are facing.

Another way to loosely categorize statisticians is in terms of industrial (or government) versus academic statisticians. Academic statisticians are primarily involved with innovative research and the teaching of statistics classes. Aside from Statistics departments, there are many alternative departments for academic statisticians including Mathematics, Economics, Business, Sociology and Psychology. Research goals for an academic statistician vary with their interests, and also depend on their emphasis toward either applied or theoretical research. In addition, the University at which they work can emphasize either a teaching or research mission that further dictates the quantity and type of research they engage in. In any case, it is a primary responsibility of an academic statistician to publish papers in leading statistics journals to advance the field. Teaching responsibilities can include introductory Statistics for undergraduate non-majors, core statistical theory and methods classes for Statistics majors and in many cases advanced graduate-level Statistics classes for students pursuing an MS and/or PhD degree in Statistics.

Industrial statisticians are frequently focused on problems that have some bearing on the company's business. In some large companies there may be a fundamental research group that operates more like an academic environment, but in recent years the number and size of these groups are diminishing as corporations are more squarely focused on their bottom lines. Industrial statisticians are expected to assimilate the company culture and add value to the projects they work on that goes well beyond the contributions that their statistical skills alone enable. They might, for example, become project managers and even technical managers where their organizational, motivational, and leadership skills become important assets to the company.

Many statisticians engage in statistical consulting, either as their primary vocation or as a part-time endeavor. Academic statisticians, for example, often have opportunities to lend their data analysis and quantitative problem solving skills to government and industry clients, and can contribute to litigation cases as an expert consultant or even an expert witness. Consultants must have exceptionally strong communication skills to be able to translate the interpretation of their findings into the language of the client. In the same way, they have to be able to elicit information from their clients that will ensure the efficacy of their data analyses. Industrial statisticians often function as internal consultants to the company they work for. This is particularly true in large companies where there can be a group of statisticians that serve as a shared central resource for the entire company.

The following alphabetical list is meant to provide an appreciation of the diversity of fields where statisticians are gainfully employed: Agriculture, Archaeology, Astronomy, Biology, Chemistry, Computer Science, Demography, Economics, Ecology, Education, Engineering, Epidemiology, Finance, Forestry, Genetics, Health Sciences, Insurance, Law, Manufacturing, Medicine, National Defense, Pharmacology, Physics, Psychology, Public Health, Safety, Sociology, Sports, Telecommunications, and Zoology. To be more specific, consider the following brief descriptions of work and employment of statisticians in the following fields:

Medicine

Florence Nightingale was not only a historic figure because of what she brought to the profession of nursing, but she was also a pioneering figure in the use of statistics. Statistical work in medicine involves designing studies

and analyzing their data to determine if new (or existing) drugs, medical procedures and medical devices are safe and effective. Statisticians find careers at pharmaceutical companies, medical research centers and governmental agencies concerned with drugs, public health and medicine.

Ecology

Research laboratories, commercial firms and government environmental agencies employ statisticians to evaluate the environmental impact of air, water and soil pollutants. Statisticians also work with government lawyers to analyze the impact (false positive or false negative) of proposed federal or state pollution tests and regulations.

Market Research

Statisticians analyze consumer demand for products and services, analyze the effectiveness of various types of advertising, and analyze the economic risk of satisfying consumer demand for products and services.

Manufacturing

The success of manufacturing industries such as aerospace, electronics, automobile, chemical or other product producing industries depends, at least in part, on the efficiency of production and the quality and reliability of their products. Statistical techniques and models are used for predicting inventory needs, improving production flow, quality control, reliability prediction and improvement, and development of product warranty plans. The Deming Prize, named after the prolific statistician W. Edwards Deming, was established in 1950 and is annually awarded to companies that make major advances in quality improvement. The Malcolm Baldridge National Quality Award, named after Malcolm Baldridge who served as the United States Secretary of Commerce under President Regan, was established in 1988 and is annually awarded to U.S. organizations for performance excellence.

Actuarial Sciences

Actuarial statisticians use Mathematics and Statistics to assess the risk of insurance and financial portfolios. Statistical methods are used, for example, to determine a wide variety of appropriate insurance premiums (e.g., homeowner, life, automobile, flood, etc.) and to manage investment and pension funds.

Safety

Statisticians are employed by many businesses and agencies to model safety concerns and to estimate and predict the probability of occurrence of these safety concerns. Nuclear power plants, national defense agencies and airlines are just a few of the businesses that statistically analyze safety risks.

Telecommunications

The reliability of voice and data networks is paramount to a telecommunication company's revenue and their brand name image. Statisticians work collaboratively with engineers to model alternative design architectures and choose the most cost-effective design that minimizes customer-perceived downtime. Statisticians working in telecommunication companies frequently shift into new technology areas to keep up with the vastly changing landscape of high-tech companies.

The authors have found that their careers in statistics involve work that is usually very interesting, often involves new ideas and learning experiences, and can definitely bring value to problem solving.

For more information on careers in statistics consult www.amstat.org or e-mail asainfo@amstat.org.

About the Authors

Dr. Daniel R. Jeske is a Professor and Chair, Department of Statistics, University of California – Riverside, CA, and is the first Director of the Statistical Consulting Collaboratory at UCR. He has published over 45 journal articles and over 35 refereed conference papers. Prior to his academic career, he was a Distinguished Member of Technical Staff and a Technical Manager at AT&T Bell Laboratories. Concurrent with that, he was a part-time Visiting Lecturer in the Department of Statistics at Rutgers University. Currently, he is an Associate Editor for The *American Statistician*.

Dr. Janet Myhre is President and Founder of Mathematical Analysis Research Corporation of Claremont, California. She is Professor Emeritus, Honorary Alumna, and Founder of the Reed Institute of Applied Statistics at Claremont McKenna College. She is a Fellow of the American Statistical Association and and has served as an Associate Editor of *Technometrics* and as Chair of the Committee on Careers in Statistics of the American Statistical Association.

Cross References

►Online Statistics Education
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Role of Statistics in Advancing Quantitative Education
►Statistics Education

Case-Control Studies

Alastair Scott, Chris Wild
Professors
The University of Auckland, Auckland, New Zealand

Introduction

The basic aim of a case-control study is to investigate the association between a disease (or some other condition of interest) and potential risk factors by drawing separate samples of "cases" (people with the disease, say) and "controls" (people at risk of developing the disease). Let Y denote a binary response variable which can take values $Y = 1$, corresponding to a case, or $Y = 0$, corresponding to a control, and let $\boldsymbol{x}$ be a vector of explanatory variables or covariates. Our aim is to fit a binary regression model to explain the probabilistic behavior of Y as a function of the observed values of the explanatory variables recorded in $\boldsymbol{x}$. We focus particularly on the logistic regression model (see ▶Logistic Regression),

$$\begin{aligned}\text{logit}\{\text{pr}(Y \mid \boldsymbol{x};\boldsymbol{\beta})\} &= \log\left\{\frac{\text{pr}(Y=1 \mid \boldsymbol{x};\boldsymbol{\beta})}{\text{pr}(Y=0 \mid \boldsymbol{x};\boldsymbol{\beta})}\right\} \\ &= \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}_1, \qquad (1)\end{aligned}$$

since this makes the analysis of case-control data particularly simple and is the model of choice in most applications.

In principle, the most straightforward way of obtaining data from which to build regression models for $\text{pr}(Y \mid \boldsymbol{x})$ would be to use a prospective sampling design. Here covariate information is ascertained for a cohort of individuals who are then tracked through time until the end of the study when whether they have contracted the disease ($Y = 1$) or not ($Y = 0$) is recorded. With prospective sampling designs, observation proceeds from covariates (explanatory variables) to response, corresponding to the logic underlying the modelling. With case-control sampling, the order is reversed, with data collection proceeding from response to covariates. The parameter $\boldsymbol{\beta}_1$ in Model (1) can still be estimated, however. Consider the simplest situation of a single binary covariate taking values $x = 0$ or $x = 1$. Using Bayes Theorem, Cornfield (1951) showed that the prospective odds ratio, $\frac{\text{pr}(Y=1|x=1)}{\text{pr}(Y=0|x=1)} / \frac{\text{pr}(Y=1|x=0)}{\text{pr}(Y=0|x=0)}$, can be expressed as $\frac{\text{pr}(x=1|Y=1)}{\text{pr}(x=0|Y=1)} / \frac{\text{pr}(x=1|Y=0)}{\text{pr}(x=0|Y=0)}$, which only involves quantities that can be estimated directly from case-control data. Cornfield also pointed out that the relative risk, $\text{pr}(Y = 1 \mid x = 1)/\text{pr}(Y = 1 \mid x = 0)$, which is usually of more interest, is approximated well by the odds ratio if the disease is rare. If the overall probability of a case can be estimated from other sources, then this can be combined with the relative risk to give estimates of the absolute risk of disease for exposed ($x = 1$) and non-exposed ($x = 0$) groups. All this extends immediately to general $\boldsymbol{\beta}_1$, all of whose components represent individual log odds ratios.

Types of Case-Control Studies

There are two broad types of case-control study, population-based and matched, corresponding to two different ways of controlling for confounding variables. In the simplest form of population-based sampling sampling, random samples are drawn independently from the case- and control-subpopulations of a real, finite target population or cohort. Covariate information, $\boldsymbol{x}$, is then ascertained for sampled individuals. Fitting logistic model (1) is particularly simple here. Following earlier work for discrete covariates, Prentice and Pyke (1979) showed that we can get valid inferences about $\boldsymbol{\beta}_1$ by running the case-control data through a standard logistic regression program designed for prospective data. The intercept β_0, which is needed if we want to estimate the absolute risk for given values of the covariates, is completely confounded with the relative sampling rates of cases and controls but can be recovered using additional information such as the finite population totals of cases and controls.

Prentice and Pyke extended this to stratified case-control sampling, where the target population is first split into strata on the basis of variables known for the whole population and separate case-control samples are drawn from each stratum. Again we get valid inferences about all the other coefficients by running the data through a prospective logistic regression program, **provided** that we introduce a separate intercept for each stratum into our model. Otherwise standard logistic programs need to be modified slightly to produce valid inferences (Scott andh Wild 1997).

In designing a population-based study, it is important to make sure that the controls really are drawn from the same population, using the same protocols, as the cases. Increasingly, controls are selected using modern sample survey techniques, involving multi-stage sampling and varying selection probabilities, to help ensure this. The modifications needed to handle these complications are surveyed in Scott and Wild (2009).

In a matched case-control study, each case is individually matched with one or more controls. This could be regarded as an limiting case of a stratified study with the

strata so finely defined that each stratum includes only a single case. If we introduce an extra intercept for each matched set, then we can no longer use a simple logistic program since the plethora of parameters will lead to inconsistent parameter estimates. Instead we need to carry out a conditional analysis. More specifically, suppose that there are M controls in the *jth* matched set and model (1) is replaced by $\text{logit}\{\text{pr}(Y \mid \boldsymbol{x};\boldsymbol{\beta})\} = \beta_{0j} + \boldsymbol{x}^T\boldsymbol{\beta}_1$ for these observations. Then the conditional probability that the covariates $\boldsymbol{x}_{j0}$ are those of the case and $(\boldsymbol{x}_{j1},\ldots,\boldsymbol{x}_{jM})$ are those of the M controls, given the set of $M+1$ covariates can be expressed in the form $exp\left(\boldsymbol{x}_{j0}^T\boldsymbol{\beta}_1\right) / \sum_{m=0}^{M+1} exp\left(\boldsymbol{x}_{jm}^T\boldsymbol{\beta}_1\right)$, which does not involve the intercept terms. Inferences about $\boldsymbol{\beta}_1$ can then be made from the conditional likelihood obtained when we combine these terms over all matched sets. With pair matching ($M = 1$), this likelihood is identical to a simple logistic regression on the difference between the paired covariates.

More sophisticated designs, including incidence density sampling, nested case-control studies and case-cohort studies, that can handle complications such as time-varying covariates and ►survival data are discussed in other chapters in this volume.

Discussion

Case-control sampling is a cost-reduction device. If we could afford to collect data on the whole finite population or cohort, then we would do so. There many practical difficulties that need to be overcome to run a successful case-control study; a good account of these is given in Breslow (2005). Despite this, the case-control study in its various forms is one of the most common designs in health research. In fact, Breslow and Day (1980) described such studies as "perhaps the dominant form of analytical research in epidemiology" and since that time the rate of appearance of papers reporting on case-control studies has gone up by a factor of more than 20. These designs are also used in other fields, sometimes under other names. In econometrics, for example, the descriptor "choice-based" is used (see Manski and McFadden (1981)).

There are several reasons for the popularity of case-control studies. The first is the simplicity of the logistic analysis outlined above. The other two reasons concern efficiency: time efficiency and statistical efficiency. The former comes from being able to use historical information immediately rather than having to follow individuals through time and then wait to observe an outcome as in a prospective study. The first chapter of Breslow and Day (1980) has a good discussion of the attendant risks. The gain in statistical efficiency can be huge. For example, suppose that we have a condition that affects only 1 individual in 20 on average and we wish to investigate the effect of an exposure that affects 50% of people. In this situation a case-control study with equal numbers of cases and controls has the same power for detecting a small increase in risk as a prospective study with approximately five times as many subjects. If the condition affects only one individual in 100 then the prospective study would need 25 times as many subjects!

About the Authors

Professor Scott is Past President of the New Zealand Statistical Society (1989–1990). He is of one of New Zealand's foremost mathematical scientists. He was founding head of the University of Auckland's Department of Statistics (and previously Head of the Department of Mathematics and Statistics). He is an elected member of the International Statistical Institute, and one of 12 honorary life members of the New Zealand Statistical Association. His 1981 paper with JNK Rao, published in the *Journal of American Statistical Association*, was selected as one of 19 landmark papers in the history of survey sampling for the 2001 centenary volume of the International Association of Survey Statisticians.

Professor Wild is, with Professor Scott, the only statistician in New Zealand that has been elected a Fellow of the American Statistical Association. He is a Past President of the International Association for Statistics Education (2003–2005). He is currently Editor of the *International Statistical Review*.

Cross References

►Medical Statistics
►Statistical Methods in Epidemiology

References and Further Reading

Breslow NE (1996) Statistics in epidemiology: the case-control study. J Am Stat Assoc 91:14–28
Breslow NE (2005) Case-control studies. In: Aherns W, Pigeot I (eds) Handbook of epidemiology. Springer, New York, pp 287–319
Breslow NE, Day NE (1980) The analysis of case-control studies. International Agency for Research on Cancer, Lyon
Cornfield J (1951) A method of estimating comparitive rates from clinical data. J Natl Cancer Inst 11:1269–1275
Manski CF, McFadden D (eds) (1981) Structural analysis of discrete data with econometric applications. Wiley, New York. Models and case-control studies. Biometrika 66:403–411
Scott AJ, Wild CJ (1997) Fitting regression models to case-control data by maximum likelihood. Biometrika 84:57–71
Scott AJ, Wild CJ (2009) Population-based case control studies. In: Pfefferman D, Rao CR (eds) Ch 38 in Handbook of statistics 29: sample surveys. Elsevier, Amsterdam, pp 1009–1031

Categorical Data Analysis

Alan Agresti[1], Maria Kateri[2]
[1]Distinguished Professor Emeritus
University of Florida, Gainesville, FL, USA
[2]Associate Professor
University of Ioannina, Ionnina, Greece

Introduction

A categorical variable consists of a set of non-overlapping categories. Categorical data are counts for those categories. The measurement scale is *ordinal* if the categories exhibit a natural ordering, such as opinion variables with categories from "strongly disagree" to "strongly agree." The measurement scale is *nominal* if there is no ordering. The types of possible analysis depend on the measurement scale.

When the subjects measured are cross-classified on two or more categorical variables, the table of counts for the various combinations of categories is a *contingency table*. The information in a contingency table can be summarized and further analyzed through appropriate *measures of association and models*. A standard reference on association measures is Goodman and Kruskal (1979).

Most studies distinguish between one or more *response variables* and a set of *explanatory variables*. When the main focus is on the association and interaction structure among a set of response variables, such as whether two variables are conditionally independent given values for the other variables, *log-linear models* are useful. More commonly, research questions focus on effects of explanatory variables on a categorical response variable. *Logistic regression models* (see ▶Logistic Regression) are then of particular interest. For *binary* (success-failure) response variables, they describe the *logit,* which is $\log[P(Y=1)/P(Y=2)]$, using

$$\log[P(Y=1)/P(Y=2)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

where Y is the binary response variable and $x_1, \ldots, x_p$ the set of the explanatory variables. For a nominal response Y with J categories, the model simultaneously describes

$$\log[P(Y=1)/P(Y=J)],$$
$$\log[P(Y=2)/P(Y=J)], \ldots, \log[P(Y=J-1)/P(Y=J)].$$

For ordinal responses, a popular model uses explanatory variables to predict a logit defined in terms of a cumulative probability (McCullagh 1980),

$$\log[P(Y \le j)/P(Y > j)],\ j = 1, 2, \ldots, J-1.$$

For categorical data, the binomial (see ▶Binomial Distribution) and multinomial distributions (see ▶Multinomial Distribution) play the central role that the normal does for quantitative data. Models for categorical data assuming the binomial or multinomial were unified with standard regression and ▶analysis of variance (*ANOVA*) models for quantitative data assuming normality through the introduction by Nelder and Wedderburn (1972) of the *generalized linear model* (*GLM*, see ▶Generalized Linear Models). This very wide class of models can incorporate data assumed to come from any of a variety of standard distributions (such as the normal, binomial, and Poisson). The GLM relates a function of the mean (such as the log or logit of the mean) to explanatory variables with a linear predictor.

Contingency Tables

Two categorical variables are *independent* if the probability of response in any particular category of one variable is the same for each category of the other variable. The most well-known result on two-way contingency tables is the test of the null hypothesis of independence, introduced by Karl Pearson in 1900. If X and Y are two categorical variables with I and J categories respectively, then their cross-classification leads to a $I \times J$ table of observed frequencies $\mathbf{n} = (n_{ij})$. Under this hypothesis, the expected cell frequencies equal $m_{ij} = n\pi_{i\cdot}\pi_{\cdot j}, i = 1, \ldots, I, j = 1, \ldots, J$, where n is the total sample size $\left(n = \sum_{i,j} n_{ij}\right)$ and $\pi_{i\cdot}(\pi_{\cdot j})$ is the ith row (jth column) marginal of the underlying probabilities matrix $\boldsymbol{\pi} = (\pi_{ij})$. Then the corresponding maximum likelihood (ML) estimates equal $\hat{m}_{ij} = np_{i\cdot}p_{\cdot j} = \frac{n_{i\cdot}n_{\cdot j}}{n}$, where p_{ij} denotes the sample proportion in cell (i,j). The hypothesis of independence is tested through Pearson's chi-squared statistic

$$\chi^2 = \frac{\sum_{i,j}(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}. \qquad (1)$$

The P-value is the right-tail probability above the observed χ^2 value. The distribution of χ^2 under the null hypothesis is approximated by a $\chi^2_{(I-1)(J-1)}$, provided that the individual expected cell frequencies are not too small. When a contingency table has ordered row or column categories (ordinal variables), specialized methods can take advantage of that ordering.

More generally, models can be formulated that are more complex than independence, and expected frequencies m_{ij} can be estimated under the constraint that the model holds. If $\hat{m}_{ij}$ are the corresponding maximum likelihood estimates, then, to test the hypothesis that the model holds, we can use the Pearson statistic (1) or the statistic that results from the standard statistical approach of

conducting a *likelihood-ratio test,* which has test statistic

$$G^2 = 2\sum_{i,j} n_{ij} \ln\left(\frac{n_{ij}}{\hat{m}_{ij}}\right). \quad (2)$$

Independence between the classification variables X and Y (i.e., $m_{ij} = n\pi_{i\cdot}\pi_{\cdot j}$, for all i and j) can be expressed in terms of a log-linear model as

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y, \quad i = 1, \ldots, I,\ j = 1, \ldots, J.$$

The more general model that allows association between the variables is

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad i = 1, \ldots, I,\ j = 1, \ldots, J. \quad (3)$$

Log-linear models describe the way the categorical variables and their association influence the count in each cell of the contingency table. They can be considered as a discrete analogue of *ANOVA*. The two-factor interaction terms relate to odds ratios describing the association.

Associations can be modeled through simpler *association models*. The simplest such model, the *linear-by-linear association model,* is relevant when both classification variables are ordinal. It replaces the interaction term λ_{ij}^{XY} by the product $\phi\mu_i\nu_j$, where μ_i and ν_j are known scores assigned to the row and column categories respectively. This model is

$$\log(m_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \phi\mu_i\nu_j, \quad i = 1, \ldots, I,\ j = 1, \ldots, J. \quad (4)$$

More general models treat one or both sets of scores as parameters.

The special case of square $I \times I$ contingency tables with the same categories for the rows and the columns occurs with matched-pairs data. For example, such tables occur in the study of *rater agreement* and in the analysis of social mobility. A condition of particular interest for such data is *marginal homogeneity,* that $\pi_{i\cdot} = \pi_{\cdot i}, i = 1, \ldots, I$. For the 2×2 case of binary matched pairs, the test comparing the margins using the chi-squared statistic $(n_{12} - n_{21})^2/(n_{12} + n_{21})$ is called *McNemar's test.*

The models for two-way tables extend to higher dimensions. The various models available vary in terms of the complexity of the association and interaction structure.

Inference and Software

Standard statistical packages, such as *SAS*, *R*, and *SPSS*, are well suited for analyzing categorical data, mainly using maximum likelihood for inference. For *SAS*, a variety of codes are presented and discussed in the Appendix of Agresti (2002), and see also Stokes et al. (2000). For *R*, see the on-line manual of Thompson (2008). Bayesian analysis of categorical data can be carried out through WinBUGS (http://wlww.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml).

The standard reference on log-linear models is Bishop et al. (1975). For logistic regression, Hosmer and Lemeshow (2000) is popular. A more comprehensive book dealing with categorical data analysis using various types of models and analyses is Agresti (2002), with Agresti (2010) focusing on ordinal data.

About the Author

Professor Agresti is recipient of the first Herman Callaert Leadership Award in Statistical Education and Dissemination, Hasselt University, Diepenbeek, Belgium (2004), and Statistician of the Year award, Chicago chapter of American Statistical Association (2002–2003). He received an honorary doctorate from De Montfort University in the U.K. in 1999. He has presented invited lectures and short courses for universities and companies in about 30 countries. Professor Agresti is author or coauthor of five textbooks, including the internationally respected text "*Categorical Data Analysis*." Professor Kateri won the Myrto Lefkopoulou award for her Ph.D. thesis in Greece and has since published extensively on methods for categorical data.

Cross References

►Algebraic Statistics
►Association Measures for Nominal Categorical Variables
►Chi-Square Test: Analysis of Contingency Tables
►Data Analysis
►Exact Inference for Categorical Data
►Generalized Linear Models
►Logistic Regression
►Variation for Categorical Variables

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York

Agresti A (2010) Analysis of ordinal categorical data, 2nd edn. Wiley, New York

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge

Goodman LA, Kruskal WH (1979) Measures of association for cross classifications. Springer, New York

Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York

McCullagh P (1980) Regression models for ordinal data (with discussion). J R Stat Soc B 42:109–142

Nelder J, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc A 135:370–384

Stokes ME, Davis CS, Koch GG (2000) Categorical data analysis using the SAS system, 2nd edn. SAS Institute, Cary

Thompson LA (2008) R (and S-PLUS) manual to accompany Agresti's Categorical data analysis (2002), 2nd edn. https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf

Causal Diagrams

Sander Greenland[1], Judea Pearl[2]
[1]Professor
University of California-Los Angeles, Los Angeles, CA, USA
[2]Professor, Director of Cognitive Systems Laboratory
University of California-Los Angeles, Los Angeles, CA, USA

From their inception, causal systems models (more commonly known as structural-equations models) have been accompanied by graphical representations or path diagrams that provide compact summaries of qualitative assumptions made by the models. These diagrams can be reinterpreted as probability models, enabling use of graph theory in probabilistic inference, and allowing easy deduction of independence conditions implied by the assumptions. They can also be used as a formal tool for causal inference, such as predicting the effects of external interventions. Given that the diagram is correct, one can see whether the causal effects of interest (target effects, or causal estimands) can be estimated from available data, or what additional observations are needed to validly estimate those effects. One can also see how to represent the effects as familiar standardized effect measures. The present article gives an overview of: (1) components of causal graph theory; (2) probability interpretations of graphical models; and (3) methodologic implications of the causal and probability structures encoded in the graph, such as sources of bias and the data needed for their control.

Introduction

From their inception in the early twentieth century, causal models (more commonly known as structural-equations models) were accompanied by graphical representations or path diagrams that provided compact summaries of qualitative assumptions made by the models. Figure 1 provides a graph that would correspond to any system of five equations encoding these assumptions:

1. independence of A and B
2. direct dependence of C on A and B
3. direct dependence of E on A and C
4. direct dependence of F on C and
5. direct dependence of D on B, C, and E

The interpretation of "direct dependence" was kept rather informal and usually conveyed by causal intuition, for example, that the entire influence of A on F is "mediated" by C.

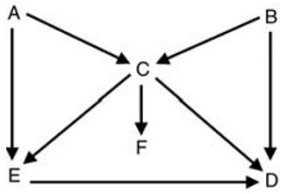
By the 1980s it was recognized that these diagrams could be reinterpreted formally as probability models, enabling use of graph theory in probabilistic inference, and allowing easy deduction of independence conditions implied by the assumptions (Pearl 1988). By the 1990s it was further recognized that these diagrams could also be used as tools for guiding causal and counterfactual inference (Pearl 1995, 2000; Pearl and Robins 1995; Spirtes et al. 2001) and for illustrating sources of bias and their remedy in empirical research (Greenland et al. 1999; Greenland 2000, 2003; Robins 2001; Greenland and Brumback 2002; Cole and Hernán 2002; Hernán et al. 2002; Jewell 2004; Pearl 2009; Glymour and Greenland 2008). Given that the graph is correct, one can see whether the causal effects of interest (target effects, or causal estimands) can be estimated from available data, or what additional observations are needed to validly estimate those effects. One can also see how to represent the effects as familiar standardized effect measures.

The present article gives an overview of: (1) components of causal graph theory; (2) probability interpretations of graphical models; and (3) methodologic implications of the causal and probability structures encoded in the graph, such as sources of bias and the data needed for their control. *See* ▶Causation and Causal Inference for discussion of definitions of causation and statistical models for causal inference.

Graphical Models and Causal Diagrams

Basics of Graph Theory

As befitting a well developed mathematical topic, graph theory has an extensive terminology that, once mastered, provides access to a number of elegant results which may be used to model any system of relations. The term *dependence* in a graph, usually represented by connectivity, may refer to mathematical, causal, or statistical dependencies. The connectives joining variables in the graph are called *arcs, edge,* or *links,* and the variables are also called *nodes* or *vertices*. Two variables connected by an arc are *adjacent* or *neighbors* and arcs that meet at a variable are also adjacent. If the arc is an arrow, the tail (starting) variable is the *parent* and the head (ending) variable is the *child*. In causal diagrams, an arrow represents a "direct effect" of the parent on the child, although this effect is direct only relative



Causal Diagrams. Fig. 1 $E \leftarrow C \rightarrow D$ is open, $E \rightarrow A \rightarrow C \leftarrow B \rightarrow D$ is closed

to a certain level of abstraction, in that the graph omits any variables that might mediate the effect.

A variable that has no parent (such as A and B in Fig. 1) is *exogenous* or *external*, or a *root* or *source* node, and is determined only by forces outside of the graph; otherwise it is *endogenous* or *internal*. A variable with no children (such as D in Fig. 1) is a *sink* or *terminal node*. The set of all parents of a variable X (all variables at the tail of an arrow pointing into X) is denoted $\text{pa}[X]$; in Fig. 1, $\text{pa}[D] = \{B, C, E\}$.

A *path* or *chain* is a sequence of adjacent arcs. A *directed path* is a path traced out entirely along arrows tail-to-head. If there is a directed path from X to Y, X is an *ancestor* of Y and Y is a *descendant* of X. In causal diagrams, directed paths represent causal pathways from the starting variable to the ending variable; a variable is thus often called a cause of its descendants and an effect of its ancestors. In a *directed* graph the only arcs are arrows, and in an *acyclic* graph there is no feedback loop (directed path from a variable back to itself). Therefore, a *directed acyclic graph* or DAG is a graph with only arrows for edges and no feedback loops (i.e., no variable is its own ancestor or its own descendant).

A variable *intercepts* a path if it is in the path (but not at the ends); similarly, a set of variables S intercepts a path if it contains any variable intercepting the path. Variables that intercept directed paths are *intermediates* or *mediators* on the pathway. A variable is a *collider* on the path if the path enters and leaves the variable via arrowheads (a term suggested by the collision of the arrows at the variable). Note that being a collider is relative to a path; for example in Fig. 1, C is a collider on the path $A \rightarrow C \leftarrow B \rightarrow D$ and a noncollider on the path $A \rightarrow C \rightarrow D$. Nonetheless, it is common to refer to a variable as a collider if it is a collider along any path (i.e., if it has more than one parent). A path is *open* or *unblocked* at noncolliders and *closed* or *blocked* at colliders; hence a path with no collider (like $E \leftarrow C \leftarrow B \rightarrow D$) is *open* or *active*, while a path with a collider (like $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$) is closed or inactive.

Some authors use a bidirectional arc (two-headed arrow, $\leftrightarrow$) to represent the assumption that two variables share ancestors that are not shown in the graph; $A \leftrightarrow B$ then means that there is an unspecified variable U with directed paths to both A and B (e.g., $A \leftarrow U \rightarrow B$).

Interpretations of Graphs

Depending on assumptions used in its construction, graphical relations may be given three distinct levels of interpretation: probabilistic, causal, and functional. We now briefly describe these levels, providing further details in later sections.

The probabilistic interpretation requires the weakest set of assumptions. It treats the diagram as a carrier of conditional independencies constraints on the joint distribution of the variables in the graph. To serve in this capacity, the parents $\text{pa}[X]$ of each variable X in the diagram are chosen so as to render X independent of all its nondescendants, given $\text{pa}[X]$. When this condition holds, we say that the the diagram is *compatible* with the joint distribution. In Fig. 1, for example, variable E is assumed to be independent of its nondescendants $\{B, D, F\}$ given its parents $\text{pa}[E] = \{A, C\}$. We will see that compatibility implies many additional independencies (e.g., E and F are independent given C) that could be read from the diagram by tracing its paths. In real-life problems, compatibility arises if each parent–child family $\{X, \text{pa}[X]\}$ represents a stochastic process by which nature determines the probability of the child X as a function of the parents $\text{pa}[X]$, independently of values previously assigned to variables other than the parents.

To use diagrams for causal inference, we must assume that the direction of the arrows correspond to the structure of the causal processes generating the data. More specifically, the graph becomes a *causal diagram* if it encodes the assumption that for each parent–child family, the conditional probability $\Pr(x|\text{pa}[X])$ would remain the same regardless of whether interventions take place on variables not involving $\{X, \text{pa}[X]\}$, even if they are ancestors or descendants of X. In Fig. 1, for example, the conditional probability $P(C|A, B)$ is assumed to remain invariant under manipulation of the consequences of C, i.e., E, F or D. A causal DAG represents a complete causal structure, in that all sources of causal dependence are explained by causal links; in particular, it is assumed that all common (shared) causes of variables in the graph are also in the graph, so that all exogenous variables (root nodes) are causally independent (although they may be unobserved).

If we assume further that the arrows represent functional relationships, namely processes by which nature assigns a definite value to each internal node, the diagram can then be used to process counterfactual information and display independencies among potential outcomes

(including counterfactual variables) (Pearl 1995, Chap. 7). We will describe such *structural diagrams* and potential outcomes below.

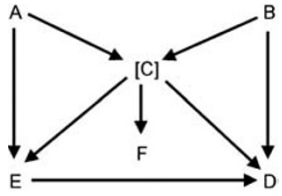
Control: Manipulation Versus Conditioning

The word "control" is used throughout science, but with a variety of meanings that are important to distinguish. In experimental research, to control a variable C usually means to manipulate or set its value. In observational studies, however, to control C (or more precisely, to control for C) more often means to condition on C, usually by stratifying on C or by entering C in a regression model. The two processes are very different physically and have very different representations and implications (Pearl 1995; Greenland et al. 1999).

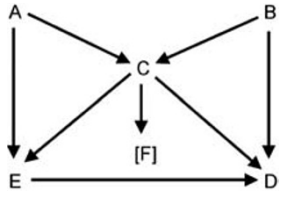
If a variable X is influenced by a researcher, a realistic causal diagram would need an ancestor R of X to represent this influence. In the classical experimental case in which the researcher alone determines X, R and X would be identical. In human trials, however, R more often represents just an *intention* to treat (with the assigned level of X), leaving X to be influenced by other factors that affect compliance with the assigned treatment R. In either case, R might be affected by other variables in the graph. For example, if the researcher uses age to determine assignments (an age-biased allocation), age would be a parent of R. Ordinarily however R would be exogenous, as when R represents a randomized allocation.

In contrast, by definition in an observational study there is no such variable R representing the researcher's influence on X. Conditioning is often used as a substitute for experimental control, in the hopes that with sufficient conditioning, X will be independent of uncontrolled influences. Conditioning on a variable C closes open paths that pass through C. However, if C is a collider, conditioning on C opens paths that were blocked by C or by an ancestral collider A. In particular, conditioning on a variable may open a path even if it is not on the path, as with F in Figs. 1 and 3.

To illustrate conditioning in a graph, we will redraw the graph to surround conditioned variables with square brackets (conditioned variables are often circled instead). We may now graphically determine the status of paths after conditioning by regarding the path open at colliders that are bracketed or have bracketed descendants, open at unbracketed noncolliders, and closed elsewhere. Figure 2 shows Fig. 1 after conditioning on C, from which we see that the $E-D$ paths $E \leftarrow C \leftarrow B \rightarrow D$ and $E \leftarrow A \rightarrow C \rightarrow D$ have been blocked, but the path $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$ has been opened. Were we to condition on F but not C, no open path would be blocked, but the path $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$ would again be opened.



Causal Diagrams. Fig. 2 Conditional on C, $E \leftarrow C \rightarrow D$ is closed but $E \rightarrow A \rightarrow C \leftarrow B \rightarrow D$ is open



Causal Diagrams. Fig. 3 Conditional on F, $E \leftarrow C \rightarrow D$ and $E \rightarrow A \rightarrow C \leftarrow B \rightarrow D$ are both open

The opening of paths at conditioned colliders reflect the fact that we should expect two unconditionally independent causes A and B become dependent if we condition on their consequences, which in Fig. 1 are C and F. To illustrate, suppose A and B are binary indicators (i.e., equal to 1 or 0), marginally independent, and $C = A + B$. Then among persons with $C = 1$, some will have $A = 1$, $B = 0$ and some will have $A = 0$, $B = 1$ (because other combinations produce $C \neq 1$). Thus when $C = 1$, A and B will exhibit perfect negative dependence: $A = 1 - B$ for all persons with $C = 1$.

The distinction between manipulation and conditioning is brought to the fore when considering the notion of "holding a variable constant." Conditioning on a variable X means that we choose to narrow the scope of discussion to those situations only where X attains a given value, regardless of how that value is attained. Manipulating X means that we physically intervene and set X to a given value, say $X = x$. The difference can be profound. For example, in cancer screening, conditioning on the absence of lighters and matches in the home lowers dramatically the probability of finding lung cancer, because restricting our attention to those who do not have these tools for smoking is tantamount to examining nonsmokers. In contrast, removing lighters and matches from people's homes during the screening will not lower the probability of finding lung cancer, since any lung cancers present will be unaffected by this act. Likewise, conditional on a low barometer reading we will have a lower probability of rain than

unconditionally, but setting the barometer to a low reading (e.g., by pushing its needle down) will have no effect on the weather.

Graphical Representation of Manipulation

One way of representing manipulation in the graph is to simulate the act of setting X to a constant, or the immediate implications of that act. If prior to intervention the probability of X is influenced by its parents via $P(x|\text{pa}[X])$, such influence no longer exists under an intervention that is made without reference to the parents or other variables. In that case, physically setting X at x dislodges X from the influence of its parents and subjects it to a new influence that keeps its value at $X = x$ regardless of the values taken by other variables. This can be represented by cutting out all arrows pointing to X and thus creating a new graph, in which X is an exogenous (root) node, while keeping the rest of the graph (with its associated conditional probabilities) intact. For example, setting C to a constant in Fig. 1, will render E and D independent, because all $E - D$ paths will be blocked by such intervention, including $E \leftarrow A \rightarrow C \leftarrow B \rightarrow D$, even though the latter path would be opened by conditioning on C. On the other hand, manipulating F but not C would leave all $E - D$ paths intact, and the $E - D$ association will therefore remain unaltered.

Assuming the graph is correct, graphical representation of interventions by deleting arrows enables us to compute post-intervention distributions from pre-intervention distributions (Pearl 1995, 2001, 2009; Spirtes et al. 2001; Lauritzen 2001) for a wide variety of interventions, including those that have side effects or that are conditioned upon other variables in the graph (Pearl 1995, pp. 105, 113). Nonetheless, "holding X constant" does not always correspond to a physically feasible manipulation, not even conceptually. Consider systolic blood pressure (SBP) as a cause of stroke (Y). It is easy to "hold SBP constant" in the sense of conditioning on each of its observed values. But what does it mean to "hold SBP constant" in the manipulative sense? There is only one condition under which SBP is constant: Death, when it stays put at zero. Otherwise, SBP is fluctuating constantly in some strictly positive range in response to posture, activity, and so on. Furthermore, no one knows how to influence SBP except by interventions R which have side effects on stroke (directed paths from R to Y that do not pass through SBP). Yet these side effects vary dramatically with intervention (e.g., there are vast differences between exercise versus medication side effects).

On the other hand, consider the problem of estimating the causal effect of SBP on the rate of blood flow in a given blood vessel. At this physiological level of discussion we can talk about the effect on blood flow of "changing SBP from level s to level s'," without specifying any mechanism for executing that change. We know from basic physics that the blood flow in a vessel depends on blood pressure, vessel diameter, blood viscosity, and so on; and we can ask what the blood flow would be if the blood pressure were to change from s to s' while the other factors remained at their ambient values. Comparing the results from conditioning on SBP $= s$ versus conditioning on SBP $= s'$ would not give us the desired answer because these conditioning events would entail different distributions for the causes (ancestors) of SBP, some of which might also affect those determinants of flow which we wish held constant when comparing.

We may thus conclude that there are contexts in which it makes no practical sense to speak of "holding X constant" via manipulation. In these contexts, manipulation of a given variable X can only be represented realistically by an additional node R representing an actual intervention, which may have side effects other than those intended or desired. On the other hand, such an R node will be redundant if X itself is amenable to direct manipulation. For such an X, manipulation can be represented by removing the arrows ending in X which correspond to effects overridden by the manipulation (Pearl 1995, 2000, 2009; Spirtes et al. 2001; Lauritzen 2001). When X is completely randomized or held constant physically, this corresponds to removing all arrows into X.

The phrase "holding X constant" may also be meaningful when X is not directly manipulable. In these cases, we may still be able to estimate a causal effect of X if we can find an instrumental variable Z (a variable that is associated with X but not with any uncontrolled confounding variable U, and Z has no effect on Y except through X). Although the operational meaning of these effects is not immediately apparent when direct manipulation of X free of side effects is not conceivable, estimation of these effects can help judge proposed interventions that affect Y via effects on X.

Separation

The intuition of closing and opening paths by conditioning is captured by the concept of "separation" which will be defined next. We say that a path is *blocked* by a set S if the path contains either an arrow-emitting node that is in S, or a collider that is outside S and has no descendant in S.

Two variables (or sets of variables) in the graph are *d-separated* (or just separated) by a set S if, after conditioning on S, there is no open path between them. Thus S d-separates X from Y if S blocks all paths from X to Y. In Fig. 1, $\{A, C\}$ d-separates E from B, but $\{C\}$ does

not (because conditioning on C alone results in Fig. 2, in which E and B are connected via the open path A). In a causal DAG, pa$[X]$ d-separates X from every variable that is not affected by X (i.e., not a descendant of X). This feature of DAGs is sometimes called the "Markov condition," expressed by saying the parents of a variable "screen off" the variable from everything but its effects. Thus in Fig. 1 pa$[E] = \{A, C\}$, which d-separates E from B but not from D.

In a probability graph, d-separation of X and Y by S implies that X and Y are independent given S in any distribution compatible with graph. In a causal diagram, d-separation of X and Y by S implies additionally that manipulation of X will not alter the distribution of Y if the variables in S are held constant physically (assuming this can be done). More generally, the distribution of Y will remain unaltered by manipulation of X if we can hold constant physically a set S that intercepts all directed paths from X to Y, even if S does not d-separate X and Y. This is so because only descendants of X can be affected by manipulation of X. In sharp contrast, conditioning on X may change the probabilities of X's ancestors; hence the stronger condition of d-separation by S is required to insure that conditioning on X does not alter the distribution of Y given S.

Statistical Interpretations and Applications

Earlier we defined the notion of compatibility between a joint probability distribution for the variables in a graph and the graph itself. It can be shown that compatibility is logically equivalent to requiring that two sets of variables are independent given S whenever S separates them in the graph. Moreover these conditional independencies constitute the *only* testable implications of a causal model specified by the diagram (Pearl 1988, p. 120). Thus, given compatibility, two sets of variables will be independent in the distribution if there is no open path between them in the graph.

Many special results follow for distributions compatible with a DAG. For example, if in a DAG, X is not an ancestor of any variable in a set T, then T and X will be independent given pa$[X]$. A distribution compatible with a DAG thus can be reduced to a product of factors $\Pr(x|\text{pa}[X])$, with one factor for each variable X in the DAG; this is sometimes called the "Markov factorization" for the DAG. When X is a treatment, this condition implies the probability of treatment is fully determined by the parents of X, pa$[X]$. Algorithms are available for constructing DAGs that are compatible with a given distribution (Pearl 1988, pp. 119–121).

Some of the most important constraints imposed by a graphical model on a compatible distribution correspond to the independencies implied by absence of open paths; e.g., absence of an open path from A to B in Fig. 1 constrains A and B to be marginally independent (i.e., independent if no stratification is done). Nonetheless, the converse does not hold; i.e., presence of an open path allows but does not imply dependency. Independence may arise through cancellation of dependencies; as a consequence even adjacent variables may be marginally independent; e.g., in Fig. 1, A and E could be marginally independent if the dependencies through paths $A \rightarrow E$ and $A \rightarrow C \rightarrow E$ cancelled each other. The assumption of faithfulness, discussed below, is designed to exclude such possibilities.

Bias and Confounding

Usually, the usage of terms like "bias," "confounding" and related concepts refer to dependencies that reflect more than just the effect under study. To capture these notions in a causal graph, we say that an open path between X and Y is a *biasing path* if it is not a directed path. The association of X with Y is then *unbiased* for the effect of X on Y if the only open paths from X to Y are the directed paths. Similarly, the dependence of Y on X is *unbiased given* S if, after conditioning on S, the open paths between X and Y are exactly (only and all) the directed paths in the starting graph. In such a case we say S is sufficient to block bias in the $X - Y$ dependence, and is minimally sufficient if no proper subset of S is sufficient.

Informally, confounding is a source of bias arising from causes of Y that are associated with but not affected by X (*see* ►Confounding). Thus we say an open nondirected path from X to Y is a *confounding path* if it ends with an arrow into Y. Variables that intercept confounding paths between X and Y are *confounders*. If a confounding path is present, we say *confounding* is present and that the dependence of Y on X is *confounded*. If no confounding path is present we say the dependence is *unconfounded*, in which case the only open paths from X to Y through a parent of Y are directed paths. Similarly, the dependence of Y on X is *unconfounded given* S if, after conditioning on S, the only open paths between X and Y through a parent of Y are directed paths.

An unconfounded dependency may still be biased due to nondirected open paths that do not end in an arrow into Y. These paths can be created when one conditions on a descendant of both X and Y, or a descendant of a variable intercepting a directed path from X to Y (Pearl 2000, p. 339). The resulting bias is called *Berksonian bias*, after its discoverer Joseph Berkson (Rothman et al. 2008). Most epidemiologists call this type of bias "selection bias" (Rothman et al. 2008) while computer scientists refer to

it as "explaining away" (Pearl 1988). Nonetheless, some writers (especially in econometrics) use "selection bias" to refer to confounding, while others call any bias created by conditioning "selection bias".

Consider a set of variables S that contains no effect (descendant) of X or Y. S is *sufficient* to block confounding if the dependence of Y on X is unconfounded given S. "No confounding" thus corresponds to sufficiency of the empty set. A sufficient S is called *minimally sufficient* to block confounding if no proper subset of S is sufficient. The initial exclusion from S of descendants of X or Y in these definitions arises first, because conditioning on X-descendants can easily block directed (causal) paths that are part of the effect of interest, and second, because conditioning on X or Y descendants can unblock paths that are not part of the $X - Y$ effect, and thus create new bias.

These considerations lead to a graphical criterion called the *back-door criterion* which identifies sets S that are sufficient to block bias in the $X - Y$ dependence (Pearl 1995, 2000). A *back-door path* from X to Y is a path that begins with a parent of X (i.e., leaves X from a "back door") and ends at Y. A set S then satisfies the back-door criterion with respect to X and Y if S contains no descendant of X and there are no open back-door paths from X to Y after conditioning on S.

In a unconditional DAG, the following properties hold (Pearl 1995, 2000; Spirtes et al. 2001; Glymour and Greenland 2008):

1. All biasing paths are back-door paths.
2. The dependence of Y on X is unbiased whenever there are no open back-door paths from X to Y.
3. If X is exogenous, the dependence of any Y on X is unbiased.
4. All confounders are ancestors of either x or of y.
5. A back-door path is open if and only if it contains a common ancestor of X and Y.
6. If S satisfies the back-door criterion, then S is sufficient to block $X - Y$ confounding.

These conditions do not extend to conditional DAGs like Fig. 2. Also, although pa$[X]$ always satisfies the back-door criterion and hence is sufficient in a DAG, it may be far from minimal sufficient. For example, there is no confounding and hence no need for conditioning whenever X separates pa$[X]$ from Y (i.e., whenever the only open paths from pa$[X]$ to Y are through X).

As a final caution, we note that the biases dealt with by the above concepts are only confounding and selection biases. To describe biases due to measurement error and model-form misspecification, further nodes representing mismeasured or misspecifed variables must be introduced (Glymour and Greenland 2008).

Estimation of Causal Effects

Suppose now we are interested in the effect of X on Y in a causal DAG, and we assume a probability model compatible with the DAG. Then, given a sufficient set S, the only source of association between X and Y within strata of S will be the directed paths from X to Y. Hence the *net effect* of $X = x_1$ vs. $X = x_0$ on Y when $S = s$ is defined as $\Pr(y|x_1, s) - \Pr(y|x_0, s)$, the difference in risks of $Y = y$ at $X = x_1$ and $X = x_0$. Alternatively one may use another effect measure such as the risk ratio $\Pr(y|x_1, s)/\Pr(y|x_0, s)$. A *standardized effect* is a difference or ratio of weighted averages of these stratum-specific $\Pr(y|x, s)$ over S, using a common weighting distribution. The latter definition can be generalized to include intermediate variables in S by allowing the weighting distribution to causally depend on X. Furthermore, given a set Z of intermediates along all directed paths from X to Y and identification of the $X - Z$ and $Z - Y$ effects, one can produce formulas for the $X - Y$ effect as a function of the $X - Z$ and $Z - Y$ effects ("front-door adjustment" (Pearl 1995, 2000)).

The above form of standardized effect is identical to the forms derived under other types of causal models, such as potential-outcome models (*see* ▶Causation and Causal Inference). In those models, the outcome Y of each unit is replaced by a vector of outcomes Y containing components Y_x, where Y_x represents the outcome when $X = x$ is the treatment given. When S is sufficient, some authors (Pearl 2000) go so far as to identify the $\Pr(y|x, s)$ with the distribution of potential outcomes Y_x given S, thereby creating a *structural model* for the potential outcomes. If the graph is based on functional rather than probabilistic relationships between parents and children, this identification can also model unit-based counterfactuals $Y_x(u)$ for any pair (X, Y), where u is a unit index or a vector of exogeneous variables characterizing the units.

There have been objections to this identification on the grounds that not all variables in the graph can be manipulated, and that potential-outcome models do not apply to nonmanipulable variables. The objection loses force when X is an intervention variable, however. In that case, sufficiency of a set S implies that the marginal potential-outcome distribution $\Pr(Y_x = y)$ equals $\sum_s \Pr(y|x, s)\Pr(s)$, which is the risk of $Y = y$ given $X = x$ standardized to the S distribution.

In fact, sufficiency of S implies the stronger condition of *strong ignorability* given S, which says that X and the vector Y of potential outcomes are independent given S. In particular, strong ignorability given S follows if S

satisfies the back-door criterion, or if X is randomized given S. Nonetheless, for the equation $\Pr(Y_x = y) = \sum_s \Pr(y|x,s)\Pr(s)$ it suffices that X be independent of each component potential outcome Y_x given S, a condition sometimes called weak ignorability given S.

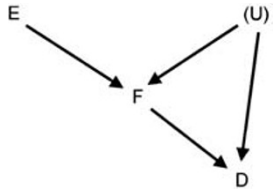
Identification of Effects and Biases

To check sufficiency and identify minimally sufficient sets of variables given a graph of the causal structure, one need only see whether the open paths from X to Y after conditioning are exactly the directed paths from X to Y in the starting graph. Mental effort may then be shifted to evaluating the reasonableness of the causal independencies encoded by the graph, some of which are reflected in conditional independence relations. This property of graphical analysis facilitates the articulation of necessary background knowledge for estimating effects, and eases teaching of algebraically difficult identification conditions.

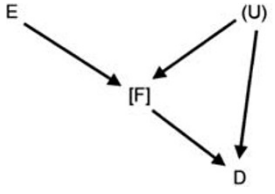
As an example, spurious sample associations may arise if each variable affects selection into the study, even if those selection effects are independent. This phenomenon is a special case of the collider-stratification effect illustrated earlier. Its presence is easily seen by starting with a DAG that includes a selection indicator $F = 1$ for those selected, 0 otherwise, as well as the study variables, then noting that we are always forced to examine associations within the $F = 1$ stratum (i.e., by definition, our observations stratify on selection). Thus, if selection (F) is affected by multiple causal pathways, we should expect selection to create or alter associations among the variables.

Figure 4 displays a situation common in randomized trials, in which the net effect of E on D is unconfounded, despite the presence of an uncontrolled cause U of D. Unfortunately, a common practice in health and social sciences is to stratify on (or otherwise adjust for) an intermediate variable F between a cause E and effect D, and then claim that the estimated (F-residual) association represents that portion of the effect of E on D not mediated through F. In Fig. 4 this would be a claim that, upon stratifying on the collider F, the $E - D$ association represents the direct effect of E on D. Figure 5 however shows the graph conditional on F, in which we see that there is now an open path from E to D through U, and hence the residual $E - D$ association is confounded for the direct effect of E on D.

The $E - D$ confounding by U in Fig. 5 can be seen as arising from the confounding of the $F - D$ association by U in Fig. 4. In a similar fashion, conditioning on C in Fig. 1 opens the confounding path through A, C, and B as seen in Fig. 2; this path can be seen as arising from the confounding of the $C - E$ association by A and the $C - D$ association by B in Fig. 1. In both examples, further stratification on either A or B blocks the created path and thus removes the new confounding.



Causal Diagrams. Fig. 4 $E \to F \to D$ is open, $E \to F \leftarrow U \to D$ is closed



Causal Diagrams. Fig. 5 Conditional on F, $E \to F \to D$ is closed but $E \to F \leftarrow U \to D$ is open

Bias from conditioning on a collider or its descendant has been called "collider bias" (Greenland 2003; Glymour and Greenland 2008). Starting from a DAG, there are two distinct forms of this bias: Confounding induced in the conditional graph (Figs. 2, 3, and 5), and Berksonian bias from conditioning on an effect of X and Y. Both biases can in principle be removed by further conditioning on certain variables along the biasing paths from X to Y in the conditional graph. Nonetheless, the starting DAG will always display ancestors of X or Y that, if known, could be used remove confounding; in contrast, no variable need appear or even exist that could be used to remove Berksonian bias.

Figure 4 also provides a schematic for estimating the $F - D$ effect, as in randomized trials in which E represents assignment to or encouragement toward treatment F. In this case E acts as an *instrumental variable* (or instrument), a variable associated with F such that every open path from E to D includes an arrow pointing into F (Pearl 2000; Greenland 2000; Glymour and Greenland 2008). Although the $F - D$ effect is not generally estimable, using the instrument E one can put bounds on confounding of the $F - D$ association, or use additional assumptions that render the effect of F on D estimable.

Questions of Discovery

While deriving statistical implications of graphical models is uncontroversial, algorithms that claim to discover causal (graphical) structures from observational data have been

subject to strong criticism (Freedman and Humphreys 1999; Robins and Wasserman 1999). A key assumption in certain "discovery" algorithms is a converse of compatibility called *faithfulness* Spirtes et al. 2001. A compatible distribution is *faithful* to the graph (or *stable* Pearl (2000)) if for all X, Y, and S, X and Y are independent given S **only** when S separates X and Y (i.e., the distribution contains no independencies other than those implied by graphical separation). Faithfulness implies that minimal sufficient sets in the graph will also be minimal for consistent estimation of effects. Nonetheless, there are real examples of near cancellation (e.g., when confounding obscures a real effect), which make faithfulness questionable as a routine assumption. Fortunately, faithfulness is not needed for the uses of graphical models discussed here.

Whether or not one assumes faithfulness, the generality of graphical models is purchased with limitations on their informativeness. Causal diagrams show whether the effects can be estimated from the given information, and can be extended to indicate effect direction when that is monotone VanderWeele and Robins 2010;. Nonetheless, the nonparametric nature of the graphs implies that parametric concepts like effect-measure modification (heterogeneity of arbitrary effect measures) cannot be displayed by the basic graphical theory. Similarly, the graphs may imply that several distinct conditionings are minimal sufficient (e.g., both $\{A, C\}$ and $\{B, C\}$ are sufficient for the *ED* effect in Fig. 1), but offer no further guidance on which to use. Open paths may suggest the presence of an association, but that association may be negligible even if nonzero. Because association transmitted by an open path may become attenuated as the length of the path increases, there is often good reason to expect certain phenomena (such as the conditional $E - D$ confounding shown in Figs. 2, 3 and 5) to be small in practical terms.

Further Readings

Full technical details of causal diagrams and their relation to causal inference can be found in the books by Pearl (2000) and Spirtes et al. (2001). A compact survey is given in Pearl (2009). Less technical reviews geared toward health scientists include Greenland et al. (2002), Greenland and Brumback (2008), and Glymour and Greenland (1999).

About the Authors

For Dr. Greenland's biography *see* the entry ►Confounding and Confounder Control.

Dr. Pearl is Professor of Computer Science at the University of California in Los Angeles, and Director of UCLA's Cognitive Systems Laboratory. He is one of the pioneers of Bayesian networks and the probabilistic approach to artificial intelligence. He is a member National Academy of Engineering (1995) and Corresponding Member, Spanish Academy of Engineering (2002). Professor Pearl was awarded the Lakatos Award, London School of Economics and Political Science for "an outstanding contribution to the philosophy of science" (2001); the ACM Allen Newell Award for "groundbreaking contributions that have changed the scientific world beyond computer science and engineering. Dr. Pearl made seminal contributions to the field of artificial intelligence that extend to philosophy, psychology, medicine, statistics, econometrics, epidemiology and social science."; the Benjamin Franklin Medal in Computers and Cognitive Science for "creating the first general algorithms for computing and reasoning with uncertain evidence, allowing computers to uncover associations and causal connections hidden within millions of observations. His work has had a profound impact on artificial intelligence and statistics, and on the application of these fields to a wide range of problems in science and engineering" (2008). He has written three fundamental books in artificial intelligence: *Heuristics: Intelligent Search Strategies for Computer Problem Solving* (Addison-Wesley, 1984), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan-Kaufmann, 1988) and *Causality: Models, Reasoning, and Inference* (Cambridge University Press, 2000). Profeshhhhsor Pearl holds two Honorary Doctorates.

Cross References

►Causation and Causal Inference
►Confounding and Confounder Control
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Rubin Causal Model
►Structural Equation Models

References and Further Reading

Cole S, Hernán MA (2002) Fallibility in estimating direct effects. Int J Epidemiol 31:163–165

Freedman DA, Humphreys P (1999) Are there algorithms that discover causal structure? Synthese 121:29–54

Glymour MM, Greenland S (2008) Causal diagrams. Ch. 12. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott, Philadelphia

Greenland S (2000) An introduction to instrumental variables for epidemiologists. Int J Epidemiol 29:722–729 (Erratum: 2000, 29, 1102)

Greenland S (2003) Quantifying biases in causal models: classical confounding versus collider-stratification bias. Epidemiology 14:300–306

Greenland S, Brumback BA (2002) An overview of relations among causal modelling methods. Int J Epidemiol 31:1030–1037

Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10:37–48
Hernán MA, Hernandez-Diaz S, Werler MM, Mitchell AA (2002) Causal knowledge as a prerequisite for confounding evaluation. Am J Epidemiol 155:176–184
Jewell NP (2004) Statistics for epidemiology. Chapman and Hall/CRC Press, Boca Raton, Sect. 8.3
Lauritzen SL (2001) Causal inference from graphical models. In: Cox DR, Kluppelberg C (eds) Complex stochastic systems. Chapman and Hall/CRC Press, Boca Raton, pp 63–107
Pearl J (1988) Probabilistic reasoning in intelligent systems. Morgan Kaufmann, San Mateo
Pearl J (1995) Causal diagrams for empirical research (with discussion). Biometrika 82:669–710
Pearl J (2000) Causality. Cambridge University Press, New York. 2nd edition, 2009
Pearl J (2009) Causal inference in statistics: an overview. Statist Surv 3:96–146
Pearl J, Robins JM (1995) Probabilistic evaluation of sequential plans from causal models with hidden variables. In: Proceedings of the eleventh conference annual conference on uncertainty in artificial intelligence (UAI-95). Morgan Kaufmann, San Francisco, pp 444–453
Robins JM (2001) Data, design, and background knowledge in etiologic inference. Epidemiology 12:313–320
Robins JM, Wasserman L (1999) On the impossibility of inferring causation from association without background knowledge. In: Glymour C, Cooper G (eds) Computation, causation, and discovery. AAAI Press/The MIT Press, Menlo Park/Cambridge, pp 305–321
Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edn. Lippincott, Philadelphia
Spirtes P, Glymour C, Scheines R (2001) Causation, prediction, and search, 2nd edn. MIT Press, Cambridge
VanderWeele, TJ, Robins JM (2010) Signed directed acyclic graphs for causal inference. J R Stat Soc Ser B 72(1):111–127

Causation and Causal Inference

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

In the health sciences, definitions of cause and effect have not been tightly bound with methods for studying causation. Indeed,many approaches to causal inference require no definition, leaving users to imagine causality however they prefer. As Sir Austin Bradford Hill said in his famous article on causation: "I have no wish…to embark upon a philosophical discussion of the meaning of 'causation'" (Hill 1965). Without a formal definition of causation, an association is distinguished as causal only by having been identified as such based on external and largely subject-matter considerations, such as those Hill put forth.

Nonetheless, beneath most treatments of causation in the health sciences, one may discern a class of definitions built around the ideas of counterfactuals or potential outcomes. These ideas have a very long history and form the foundation of most current statistical methods for causal inference. Thus, the present article will begin with these definitions and the methods they entail. It will then turn methods that explicitly presume no definition of causation but rather begin with an idea of what a causal association should look like (perhaps derived from subject-matter judgments, including consideration of possible counterfactuals), and employ statistical methods to estimate those associations.

Counterfactuals and Potential Outcomes

Skeptical that induction in general and causal inference in particular could be given a sound logical basis, David Hume nonetheless captured the foundation of the potential-outcome approach when he wrote

> We may define a cause to be an object, followed by another, … where, if the first object had not been, the second had never existed.
>
> (Hume 1748, p. 115)

A key aspect of this view of causation is its *counterfactual* element: It refers to how a certain outcome event (the "second object," or effect) would not have occurred if, *contrary to fact*, an earlier event (the "first object," or cause) had not occurred. In this regard, it is no different from conventional statistics, which refers to samples that might have occurred, but did not. This counterfactual view of causation was adopted by numerous philosophers and scientists after Hume (e.g., Mill 1843; Fisher 1918; Cox 1958; Simon and Rescher 1966; MacMahon and Pugh 1967; Stalnaker 1968; Lewis 1973).

The development of this view into a statistical theory with methods for causal inference is recounted by Rubin (1990), Greenland et al. (1999), Greenland (2004), and Pearl (2009). The earliest such theories were developed in the 1920s by Fisher, Neyman, and others for the analysis of randomized experiments and are today widely recognized under the heading of *potential-outcome models* of causation (also known in engineering as *destructive-testing* models). Suppose we wish to study the effect of an intervention variable X on a subsequent outcome variable Y defined on an observational unit or a population; for example, X could be the daily dose regimen for a drug in a clinical trial, and Y could be survival time. Given X has potential values $x_1, \ldots, x_J$ (e.g., drug doses), we suppose that there is a list of *potential outcomes* $\mathbf{y} = (y(x_1), \ldots, y(x_J))'$ such that if $X = x_j$ then $Y = y(x_j)$. The list $\mathbf{y}$ thus exhibits

the correspondence between treatments, interventions, or actions (the X values) and outcomes or responses (the Y values) for the unit, and so is sometimes called a *response schedule* (Berk 2004). A simpler and common notation has $\mathbf{y} = (y_1, \ldots, y_J)'$, with Y_j denoting the random variable "outcome when treated with $X = x_j$."

Under this model, assignment of a unit to a treatment level x_j is a choice of which potential outcome $y(x_j)$ from the list $\mathbf{y}$ to attempt to observe. It is ordinarily assumed that the assignments made for other units do not affect the outcomes of another unit, although there are extensions of the model to include between-unit interactions, as in contagious outcomes (Halloran and Struchiner 1995). Regardless of the X assignment, the remaining potential outcomes are treated as existing pre-treatment covariates on which data are missing (Rubin 1978, 1991). Because at most one of the J potential outcomes is observed per unit, the remaining potential outcomes can be viewed as missing data, and causal inference can thus be seen as a special case of inference with missing data.

To say that intervention x_i causally affects Y relative to intervention x_j means that $y(x_i) \neq y(x_j)$, i.e., X "matters" for Y for the unit. The *sharp* (or strong) null hypothesis is that $y(x)$ is constant over x within units. This hypothesis states that changing X would not affect the Y of any unit, i.e., $y(x_i) = y(x_j)$ for every unit and every x_i and x_j; it forms the basis of exact ►permutation tests such as ►Fisher's exact test (Greenland 1991). The effect of intervention x_i relative to x_j on a unit may be measured by the difference in potential outcomes $y(x_i) - y(x_j)$. If the outcome is strictly positive (like life expectancy or mortality risk), it could instead be measured by the ratio $y(x_i)/y(x_j)$.

Because we never observe two potential outcomes on a unit, we can only estimate population averages of the potential outcomes. We do this by observing average outcomes in differently exposed groups and substituting those observations for the average potential outcomes in the group of interest – a perilous process whenever the observed exposure groups are atypical of the population of interest with respect to other risk factors for the outcome (Maldonado and Greenland 2002) (see Confounding and Confounder Control).

A more subtle problem is that only for difference measures will the population effect (the difference of average potential outcomes) equal the population average effect (the average difference of potential outcomes). Hence the average of the differences $y(x_i) - y(x_j)$ in the population is often called the *average causal effect* (ACE) (Angrist et al. 1996). For some popular measures of effect, such as rate ratios and odds ratios, the population effect may not even equal any average of individual effects (Greenland 1987, 1996; Greenland et al. 1999).

The theory extends to probabilistic outcomes by replacing the $y(x_j)$ by probability functions $p_j(y)$ (Greenland 1987; Robins 1988; Greenland et al. 1999). The theory also extends to continuous X by allowing the potential-outcome list $\mathbf{y}$ to contain the potential outcome $y(x)$ or $p_x(y)$ for every possible value x of X. Both extensions are embodied in Pearl's notation for intervention effects, in which $p_x(y)$ becomes $P(Y = y|\text{set}[X = x])$ or $P(Y = y|\text{do}[X = x])$ (Pearl 1995, 2009). Finally, the theory extends to complex longitudinal data structures by allowing the treatments to be different event histories or processes (Robins 1987, 1997).

From Randomized to Observational Inference

Potential outcomes were developed part of a design-based strategy for causal inference in which ►randomization provided the foundation for inference. Indeed, before the 1980s, the model was often referred to as "the randomization model," even though the causal concepts within it do not hinge on randomization (e.g., Wilk 1955; Copas 1973). It thus seems that the early strong linkage of potential outcomes to randomized designs deflected consideration of the model for observational research. In the 1960s, however, a number of philosophers used counterfactuals to build general foundations for causal analysis (e.g., Simon and Rescher 1966; Stalnaker 1968; Lewis 1973). Similar informal ideas can be found among epidemiologists of the era (e.g., MacMahon and Pugh 1967), and conceptual models subsuming counterfactuals began to appear shortly thereafter (e.g., Miettinen 1972; Rothman 1976; Hamilton 1979).

The didactic value of these models was quickly apparent in the clarification they brought to ideas of strength of effect, synergy, and antagonism (MacMahon and Pugh 1967; Rothman 1976; see also Rothman et al. 2008, Chaps. 2 and 5). Most importantly, the models make clear distinctions between causal and statistical relations: Causal relations refer to relations of treatments to potential outcomes *within* treated units, whereas statistical relations refer associations of treatments with actual outcomes *across* units (Rothman et al. 2008, Chap. 4). Consequently, the models have aided in distinguishing confounding from collapsibility (Greenland and Robins 1986; Greenland et al. 1999), synergy from statistical interaction (Greenland and Poole 1988), and causation probabilities from attributable fractions (Greenland et al. 1999; Greenland and Robins 2000).

The conceptual clarification also stimulated development of statistical methods for observational studies. Rubin (1974, 1978) and his colleagues extended statistical machinery based on potential outcomes from

the experimental setting to observational data analysis, leading, for example, to propensity-scoring and inverse-probability-of-treatment methods for confounder adjustment (Rosenbaum 2002; Hirano et al. 2003), as well as new insights into analysis of trials with noncompliance (Angrist et al. 1996) and separation of direct and indirect effects (Robins and Greenland 1992, 1994; Frangakis and Rubin 2002; Kaufman et al. 2004). In many cases, such insights have led to methodologic refinements and better-informed choices among existing methods. In the longitudinal-data setting, potential-outcome modeling has led to entirely new methodologies for analysis of time-varying covariates and outcomes, including g-estimation and marginal structural modeling (Robins 1987, 1998; Robins et al. 1992, 1999, 2000).

A serious caution arises, however, when it is not clear that the counterfactual values for X (treatments other than the actual one) represent physical possibilities or even unambiguous states of nature. A classic example is gender (biological sex). Although people speak freely of gender (male vs. female) as cause of heart disease, given a particular man, it is not clear what it would mean for that man to have been a woman instead. Do we mean that the man cross-dressed and lived with a female identity his entire life? Or that he received a sex-change operation after birth? Or that the zygote from which he developed had its male chromosome replaced by a female chromosome?

Potential-outcome models bring to light such ambiguities in everyday causal language but do not resolve them (Greenland 2005a; Hernán 2005). Some authors appear to insist that use of the models be restricted to situations in which ambiguities are resolved, so that X must represent an intervention variable, i.e., a precise choice among treatment actions or decisions (Holland 1986). Many applications do not meet this restriction, however, and some go so far as to confuse outcomes (Y) with treatments (X), which can lead to nonsense results. Examples include estimates of mortality after "cause removal," e.g., removal of all lung-cancer deaths. Sensible interpretation of any effect estimate requires asking what intervention on a unit could have given the unit a value of X (here, lung-cancer death) other than the one that was observed, and what the side effects that intervention would have. One cannot remove all lung-cancer deaths by smoking cessation. A treatment with a 100% cure rate might do so but need not guarantee the same subsequent lifespan as if the cancer never occurred. If such questions cannot be given at least a speculative answer, the estimates of the impact of cause removal cannot be expected to provide valid information for intervention and policy purposes (Greenland 2005a).

More sweeping criticisms of potential-outcome models are given by Dawid (2000), for example, that the distribution of the full potential-outcome vector $\mathbf{Y}$ (i.e., the joint distribution of the $Y(x_1), \ldots, Y(x_J)$) cannot be nonparametrically identified by randomized experiments. Nonetheless, as the discussants point out, the practical implication of these criticisms are not clear, because the marginal distributions of the separate potential outcomes $Y(x_j)$ are nonparametrically identifiable, and known mechanisms of action may lead to identification of their joint distribution as well.

Canonical Inference

Before the extension of potential outcomes to observational inference, the only systematic approach to causal inference in epidemiology was the informal comparison of observations to characteristics expected of causal relations. Perhaps, the most widely cited of such approach is based on Hill's considerations (Hill 1965), which are discussed critically in numerous sources (e.g., Koepsell and Weiss 2003; Phillips and Goodman 2004; Rothman et al. 2008, Chap. 2) as well as by Hill himself.

The canonical approach usually leaves terms like "cause" and "effect" as undefined concepts around which the self-evident canons are built, much like axioms are built around concepts like "set" and "is an element of" in mathematics. Only proper temporal sequence (cause must precede effect) is a necessary condition for a cause–effect relation to hold. The remaining considerations are more akin to diagnostic symptoms or signs of causation – that is, they are properties an association is assumed more likely to exhibit if it is causal than if it is not (Hill 1965; Susser 1988, 1991). Furthermore, some of these properties (like specificity and dose response) apply only under specific causal models (Weiss 1981, 2002).

Thus, the canonical approach makes causal inference more closely resemble clinical judgment than experimental science, although experimental evidence is listed among the considerations (Hill 1965; Rothman et al. 2008, Chap. 2; Susser 1991). Some of the considerations (such as temporal sequence, association, dose response or predicted gradient, and specificity) are empirical signs and thus subject to conventional statistical analysis. Others (such as plausibility) refer to prior belief, and thus (as with disease symptoms) require elicitation, the same process used to construct priors for Bayesian analysis.

The canonical approach is widely accepted in health sciences, subject to many variations in detail. Nonetheless, it has been criticized for its incompleteness and informality, and the consequent poor fit it affords to the deductive

or mathematical approaches familiar to classic science and statistics (Rothman et al. 2008, Chap. 2). Although there have been some interesting attempts to reinforce or reinterpret certain canons as empirical predictions of causal hypotheses (e.g., Susser 1988; Weed 1986; Weiss 1981, 2002; Rosenbaum 2002), there is no generally accepted mapping of the entire canonical approach into a coherent statistical methodology; one simply uses standard statistical techniques to test whether empirical canons are violated. For example, if the causal hypothesis linking X to Y predicts a strictly increasing trend in Y with X, a test of this statistical prediction may serve as a statistical criterion for determining whether the hypothesis fails the dose-response canon. Such usage falls squarely in the falsificationist/frequentist tradition of the twentieth-century statistics, but leaves unanswered most of the policy questions that drive causal research; this gap led to the development of methodologic modeling or *bias analysis*.

Bias Analysis

In the second half of the twentieth-century, a more rigorous approach to observational studies emerged in the wake of major policy controversies such as those concerning cigarette smoking and lung cancer (e.g., Cornfield et al. 1959). This approach begins with the idea that, conditional on some sufficient set of confounders Z, there is a population association or relation between X and Y that is the target of inference. In other words, the Z-stratified associations are presumed to accurately reflect the effect of X on Y in that population stratum, however "effect" may be defined. Estimates of this presumably causal association are then the effect estimates.

Observational and analytic shortcomings bias or distort these estimates: Units may be selected for observation in a nonrandom fashion; stratifying on additional unmeasured covariates U may be essential for the X-Y association to approximate a causal effect; inappropriate covariates may be entered into the analysis; components of X or Y or Z may not be adequately measured; and so on. In methodologic modeling or *bias analysis*, one models these shortcomings. In effect, one attempts to model the design and execution of the study, including features (such as selection biases and measurement errors) beyond investigator control. The process is thus a natural extension to observational studies of the design-based paradigm in experimental and survey statistics. For further details, see BIAS MODELING or the overviews by Greenland (2005b, 2009).

Structural Equations and Causal Diagrams

Paralleling the development of potential-outcome models, an entirely different approach causal analysis arose in observational research in economics and related fields. Like methodologic modeling, this *structural-equations* approach does not begin with a formal definition of cause and effect, but instead develops models to reflect assumed causal associations, from which empirical (and hence testable) associations may be derived. Like most of statistics before the 1980s, structural-equations methods were largely limited to normal linear models to derive statistical inferences. Because these models bear no resemblance to typical epidemiologic data, this limitation may in part explain the near absence of structural equations from epidemiology, despite their ubiquity in social-science methodology. From their inception, however, causal system models have been accompanied by graphical representations or path diagrams that provided compact summaries of qualitative assumptions made by the structural model; see ►Causal Diagrams for a review.

Conclusion

Different approaches to causal inference represent separate historical streams rather than distinct methodologies, and can be blended in various ways. The result of any modeling exercise is simply one more input to informal judgments about causal relations, which may be guided by canonical considerations. Insights and innovations in any approach can thus benefit the entire process of causal inference, especially when that process is seen as part of a larger context. Other traditions or approaches (some perhaps yet to be imagined) may contribute to the process. It thus seems safe to say that no one approach or blend is a complete solution to the problem of causal inference, and that the topic remains one rich with open problems and opportunities for innovation.

Acknowledgments

Some of the above material is adapted from Greenland (2004).

About the Author

For biography *see* the entry ►Confounding and Confounder Control.

Cross References

►Bias Analysis
►Causal Diagrams
►Complier-Average Causal Effect (CACE) Estimation

►Confounding and Confounder Control
►Event History Analysis
►Forecasting Principles
►Interaction
►Misuse of Statistics
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Rubin Causal Model
►Simpson's Paradox
►Spurious Correlation
►Structural Equation Models

References and Further Reading

Angrist J, Imbens G, Rubin DB (1996) Identification of causal effects using instrumental variables (with discussion). J Am Stat Assoc 91:444–472

Berk R (2003) Regression analysis: a constructive critique. Sage Publications, Thousand Oaks

Copas JG (1973) Randomization models for matched and unmatched 2 × 2 tables. Biometrika 60:267–276

Cornfield J, Haenszel W, Hammond WC, Lilienfeld AM, Shimkin MB, Wynder EL (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. J Nat Cancer Inst 22:173–203

Cox DR (1958) The planning of experiments. Wiley, New York

Dawid P (2000) Causal inference without counterfactuals (with discussion). J Am Stat Assoc 95:407–448

Fisher RA (1918) The causes of human variability. Eugenics Rev 10:213–220

Frangakis C, Rubin DB (2002) Principal stratification in causal inference. Biometrics 58:21–29

Greenland S (1987) Interpretation and choice of effect measures in epidemiologic analysis. Am J Epidemiol 125:761–768

Greenland S (1991) On the logical justification of conditional tests for two-by-two contingency tables. Am Stat 45:248–251

Greenland S (1996) Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology 7:498–501

Greenland S (1998) Induction versus Popper: substance versus semantics. Int J Epidemiol 27:543–548

Greenland S (1999) The relation of the probability of causation to the relative risk and the doubling dose: a methodologic error that has become a social problem. Am J Publ Health 89: 1166–1169

Greenland S (2000) An introduction to instrumental variables for epidemiologists. Int J Epidemiol 29:722–729 (Erratum: Int J Epidemiol 29:1102)

Greenland S (2003) Quantifying biases in causal models: classical confounding versus collider-stratification bias. Epidemiology 14:300–306

Greenland S (2004) An overview of methods for causal inference from observational studies. In: Gelman A, Meng XL (eds) Applied Bayesian modeling and causal inference from an incomplete-data perspective. Wiley, New York, pp 3–13

Greenland S (2005a) Epidemiologic measures and policy formulation: lessons from potential outcomes (with discussion). Emerg Themes in Epidemiol (online journal), http://www.ete-online.com/content/2/1/5

Greenland S (2005b) Multiple-bias modeling for observational studies (with discussion). J Roy Stat Soc, ser A 168:267–308

Greenland S (2009) Relaxation penalties and priors for plausible modeling of nonidentified bias sources. Stat Sci 24:195–210

Greenland S, Brumback BA (2002) An overview of relations among causal modelling methods. Int J Epidemiol 31:1030–1037

Greenland S, Poole C (1988) Invariants and noninvariants in the concept of interdependent effects. Scand J Work Environ Health 14:125–129

Greenland S, Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15:413–419

Greenland S, Robins JM (2000) Epidemiology, justice, and the probability of causation. Jurimetrics 40:321–340

Greenland S, Robins JM, Pearl J (1999) Confounding and collapsibility in causal inference. Stat Sci 14:29–46

Halloran ME, Struchiner CJ (1995) Causal inference for infectious diseases. Epidemiology 6:142–151

Hamilton MA (1979) Choosing a parameter for 2 × 2 table or 2 × 2 × 2 table analysis. Am J Epidemiol 109:362–379

Hernán MA (2005) Hypothetical interventions to define causal effects — afterthought or prerequisite? Am J Epidemiol 162: 618–620

Hill AB (1965) The environment and disease: association or causation? Proc Roy Soc Med 58:295–300

Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71:1161–1189

Hume D (1748) An enquiry concerning human understanding. 1988 reprint by Open Court Press, LaSalle

Jewell NP (2004) Statistics for epidemiology. Chapman & Hall/CRC Press, Boca Raton

Kaufman JS, MacLehose R, Kaufman S (2004) A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. Epidemiol Perspect Innov (online journal) 1:4

Lewis DK (1973) Causation. J Philos 70:556–567

MacMahon B, Pugh TF (1967) Causes and entities of disease. In: Clark DW, MacMahon B (eds) Preventive medicine. Little Brown, Boston, pp 11–18

Maldonado G, Greenland S (2002) Estimating causal effects (with discussion). Int J Epidemiol 31:421–438

Miettinen OS (1972) Standardization of risk ratios. Am J Epidemiol 96:383–388

Mill JS (1843) A system of logic, ratiocinative and inductive. 1956 reprint by Longman & Greens, London

Morrison AS (1985) Screening in chronic disease. Oxford, New York

Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–710

Pearl J (2009) Causality, 2nd edn. Cambridge University Press, New York

Phillips CV, Goodman K (2004) The missed lessons of Sir Austin Bradford Hill. Epidemiologic Perspectives Innovations (online journal), http://www.epi-perspectives.com/content/1/1/3

Robins JM (1987) A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. J Chronic Dis 40 (suppl 2):139s–161s

Robins JM (1988) Confidence intervals for causal parameters. Stat Med 7:773–785

Robins JM (1989) The control of confounding by intermediate variables. Stat Med 8:679–701

Robins JM (1997) Causal inference from complex longitudinal data. In: Berkane M (ed) Latent variable modeling and applications to

causality. Lecture notes in statistics (120). Springer, New York, pp 69–117
Robins JM (1998) Structural nested failure time models. In: Armitage P, Colton T (eds) The encyclopedia of biostatistics. Wiley, New York, pp 4372–4389
Robins JM (1999) Marginal structural models versus structural nested models as tools for causal inference. In: Halloran ME, Berry DA (eds) Statistical models in epidemiology. Springer, New York, pp 95–134
Robins JM (2001) Data, design, and background knowledge in etiologic inference. Epidemiology 12:313–320
Robins JM, Greenland S (1989) The probability of causation under a stochastic model for individual risks. Biometrics 46: 1125–1138
Robins JM, Greenland S (1992) Identifiability and exchangeability for direct and indirect effects. Epidemiology 3:143–155
Robins JM, Greenland S (1994) Adjusting for differential rates of PCP prophylaxis in high-versus low-dose AZT treatment arms in an AIDS randomized trial. J Am Stat Assoc 89: 737–749
Robins JM, Blevins D, Ritter G, Wulfson M (1992) G-estimation of the effect of prophylaxis therapy for Pnemocystis carinii pneumonia on the survival of AIDS patients. Epidemiology 3:319–336 (Errata: Epidemiology 4:189)
Robins JM, Greenland S, Hu FC (1999) Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). J Am Stat Assoc 94: 687–712
Robins JM, Hernán MA, Brumback BA (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11:550–560
Rosenbaum P (2002) Observational studies, 2nd edn. Springer, New York
Rothman KJ (1976) Causes. Am J Epidemiol 104:587–592
Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology, 3rd edn. Lippincott, Philadelphia
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educat Psychol 66: 688–701
Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6:34–58
Rubin DB (1990) Comment: Neyman (1923) and causal inference in experiments and observational studies. Stat Sci 5: 472–480
Rubin DB (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics 47:1213–1234
Simon HA, Rescher N (1966) Cause and counterfactual. Philos Sci 33:323–340
Stalnaker RC (1968) A theory of conditionals. In: Rescher N (ed) Studies in logical theory. Blackwell, Oxford
Stone (1993) The assumptions on which causal inference rest. J Roy Stat Soc, ser B 55:455–466
Susser M (1988) Falsification, verification and causal inference in epidemiology: reconsideration in light of Sir Karl Popper's philosophy. In: Rothman KJ (ed) Causal inference. Epidemiology Resources, Boston, pp 33–57
Susser M (1991) What is a cause and how do we know one? A grammar for pragmatic epidemiology. Am J Epidemiol 133: 635–648
Weed DL (1986) On the logic of causal inference. Am J Epidemiol 123:965–979
Weiss NS (1981) Inferring causal relationships: elaboration of the criterion of "dose-response." Am J Epidemiol 113:487–490
Weiss NS (2002) Can "specificity" of an association be rehabilitated as a basis for supporting a causal hypothesis? Epidemiology 13:6–8
Welch BL (1937) On the z-test in randomized blocks and Latin squares. Biometrika 29:21–52
Wilk MB (1955) The randomization analysis of a generalized randomized block design. Biometrika 42:70–79

Censoring Methodology

Ng Hon Keung Tony
Associate Professor
Southern Methodist University, Dallas, TX, USA

Basic Concepts on Censored Data

In industrial and clinical experiments, there are many situations in which units (or subjects) are lost or removed from experimentation before the event of interest occurs. The experimenter may not always obtain complete information on the time to the event of interest for all experimental units or subjects. Data obtained from such experiments are called *censored data*. Censoring is one of the distinguishing features of lifetime data. Censoring can be either unintentional due to accidental breakage or an individual under study drops out or intentional in which the removal of units or subjects is pre-planned, or both. Censoring restricts our ability to observe the time-to-event and it is a source of difficulty in statistical analysis.

Censoring can occur at either end (single censoring) or at both ends (double censoring). If the event of interest is only known to be occured before a certain time, it is called *left censoring*. The term "left censored" implies that the event of interest is to the left of the observed time point. The most common case of censoring is *right censoring*, in which the exact time to the event of interest is not observed and it is only known to be occured after a certain time. Different types of right censoring schemes are discussed in the subsequent section. For *interval censoring*, the event of interest is only known to be occurred in a given time interval. This type of data frequently comes from experiments where the items under test are not constantly monitored, for example, the patients in a clinical trial have periodic follow-up and events of interest occur in between two consecutive follow-ups. Note that left censoring is a special case of interval censoring where the starting time for the interval is zero.

For life-testing experiments where the event of interest is the failure of the item on test, two common reasons for pre-planned censoring are saving the total time on test and reducing the cost associated with the experiment because failure implies unit's destruction which can be costly. When budget and/or facility constraints are in place, suitable censoring scheme can be used to control the time spent and the cost of the experiment. Nevertheless, censored data usually will reduce the efficient of statistical inference compare to complete data. Therefore, it is desirable to develop censoring scheme which can balance between (i) total time spent for the experiment; (ii) number of units used in the experiment; and (iii) the efficient of statistical inference based on the results of the experiment.

Different Types of Censoring Schemes

Suppose n units are placed on a life-testing experiment. Further, suppose $X_1, X_2, \cdots, X_n$ denote the lifetimes of these n units taken from a population with lifetime distribution function $F(x;\theta)$ and density function $f(x;\theta)$, where θ is an unknown parameter(s) of interest. Let $X_{1:n} \le \cdots \le X_{n:n}$ denote the corresponding ordered lifetimes observed from the life-test. Some commonly used censoring schemes are discussed in the following.

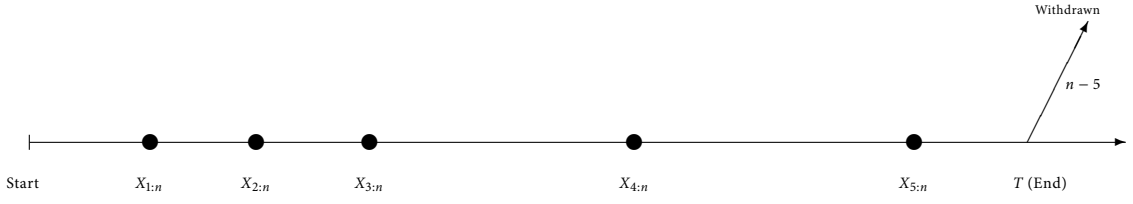
Type-I Censoring

Suppose it is planned that the life-testing experiment will be terminated at a pre-fixed time T. Then, only the failures until time T will be observed. The data obtained from such a restrained life-test will be referred to as a *Type-I censored sample*. It is also called time-censoring since the experimental time is fixed. Note that the number of failures observed here is random and, in fact, has a *Binomial*$(n, F(T;\theta))$ distribution. Figure 1 shows a schematic representation of a Type-I censored life-test with $m = 7$. Inferential procedures based on Type-I censored samples have been discussed extensively in the literature; see, for example, Cohen (1991) and Balakrishnan and Cohen (1991).

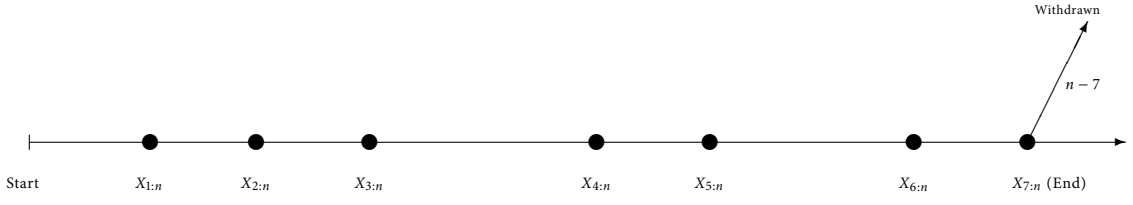
Type-I censoring scheme has the advantage that the experimental time is controlled to be at most T while it has the disadvantage that the effective sample size can turn out to be a very small number (even equal to zero) so that usual statistical inference procedures will not be applicable or they will have low efficiency.

Type-II Censoring

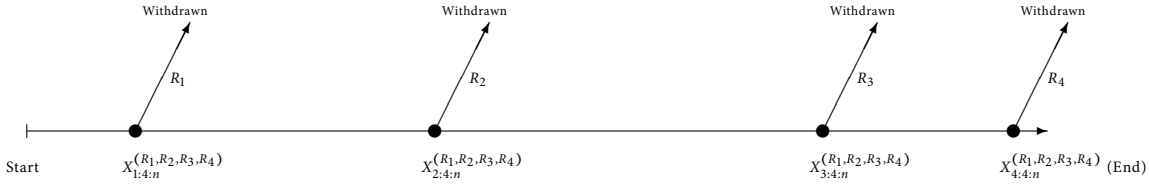
Suppose it is planned that the life-testing experiment will be terminated as soon as the mth (where m is pre-fixed) failure is observed. Then, only the first m failures out of n units under test will be observed. The data obtained from such a restrained life-test will be referred to as a *Type-II censored sample*. In contrast to Type-I censoring, the number of failures observed is fixed (viz., m) while the duration of the experiment is random (viz., $X_{m:n}$). Figure 2 shows a schematic representation of a Type-II censored life-test with m = 7. Inferential procedures based on Type-II censored samples have been discussed extensively in the



Censoring Methodology. Fig. 1 Schematic representation of a Type-I censored life-test



Censoring Methodology. Fig. 2 Schematic representation of a Type-II censored life-test



Censoring Methodology. Fig. 3 Schematic representation of a progressively Type-II censored life-test

literature; see, for example, Nelson (1982), Cohen (1991), and Balakrishnan and Cohen (1991).

Type-II censoring scheme has the advantage that the number of observed failures is fixed to be m which ensure reasonable information is available for statistical inference. However, it has the disadvantage that the experimental time is random and it can be large.

Progressive Censoring

Both the conventional Type-I and Type-II censoring schemes do not have the flexibility of allowing removal of units at points other than the terminal point of the experiment. This restricts our ability to observe extreme failures which may lead to inefficient statistical inference if we are interested in the behavior of the upper tail of the lifetime distribution. For this reason, a more general censoring scheme called *progressive censoring* has been introduced. The censored life-testing experiments described above can be extended to situations wherein censoring occurs in multiple stages. Data arising from such life-tests are referred to as *progressively censored data*. Naturally, progressive censoring can be introduced in both Type-I and Type-II forms.

For example, a progressive Type-II censored life-testing experiment will be carried out in the following manner. Prior to the experiment, a number $m < n$ is determined and the censoring scheme $(R_1, R_2, \ldots, R_m)$ with $R_j > 0$ and $\sum_{j=1}^{m} R_j + m = n$ is specified. During the experiment, j-th failure is observed and immediately after the failure, R_j functioning items are removed from the test. We denote the m completely observed (ordered) lifetimes by $X_{j:m:n}^{(R_1,R_2,\ldots,R_m)}$, $j = 1, 2, \ldots, m$, which are the observed progressively Type-II right censored sample. Figure 3 shows a schematic representation of a progressively Type-II censored life-test with $m = 4$. Notice that the conventional Type-II censoring scheme is a special case of a progressive Type-II censoring scheme when $R_i = 0$, for $i = 1, \ldots, m-1$ and $R_m = n-m$. Similarly, progressive Type-I censoring scheme can be introduced in a similar manner. Inferential procedures based on progressively Type-II censored samples have been discussed in the literature; see, for example, Balakrishnan and Aggarwala (2000) and Balakrishnan (2007) for excellent reviews on the literatures on this topic.

Hybrid Censoring

As mentioned previously, both Type-I and Type-II censoring schemes have some shortcomings. To keep away from these shortcomings, hybrid censoring schemes combining Type-I and Type-II censoring schemes have been proposed. Specifically, if the experiment is terminated at $T^* = \min\{X_{m:n}, T\}$, where m and T are pre-fixed prior to the experiment, then the censoring scheme is called *Type-I hybrid censoring scheme*; if the experiment is terminated at $T^* = \max\{X_{m:n}, T\}$, then the censoring scheme is called *Type-II hybrid censoring scheme*. We can see that both Type-I and Type-II hybrid censoring schemes try to balance between the advantages and disadvantages of conventional Type-I and Type-II censoring schemes. Hybrid censoring schemes has been studied extensively in the literature, one may refer Epstein (1954), Draper and Guttman (1987), Gupta and Kundu (1998), and Childs et al. (2003, 2008), Kundu (2007) for details. In recent years, the idea of hybrid censoring has been generalized to progressive censoring, for discussions on different types of hybrid progressive censoring schemes, see, for example, Kundu and Joarder (2006), Banerjee and Kundu (2008), Ng et al. (2009) and Lin et al. (2009).

About the Author

H. K. T. Ng is an Associate Professor in the Department of Statistical Science at Southern Methodist University. He received his Ph.D. degree in Mathematics (2002) from McMaster University, Hamilton, Canada. He is an elected member of International Statistical Institute (ISI) and an elected senior member of Institute of Electrical and Electronics Engineers (IEEE). He is currently an Associate Editor of Communications in Statistics.

Cross References

►Astrostatistics
►Event History Analysis

►Nonparametric Estimation Based on Incomplete Observations
►Ordered Statistical Data: Recent Developments
►Step-Stress Accelerated Life Tests
►Survival Data

References and Further Reading

Balakrishnan N (2007) Progressive censoring methodology: an appraisal (with discussions). Test 16:211–296
Balakrishnan N, Aggarwala R (2000) Progressive censoring: theory, methods and applications. Birkhäuser, Boston
Balakrishnan N, Cohen AC (1991) Order statistics and inference: estimation methods. Academic, San Diego
Banerjee A, Kundu D (2008) Inference based on Type-II hybrid censored data from a Weibull distribution. IEEE Trans Reliab 57:369–378
Childs A, Chandrasekar B, Balakrishnan N, Kundu D (2003) Exact likelihood inference based on Type-I and Type-II hybrid censored samples from the exponential distribution. Ann Inst Stat Math 55:319–330
Childs A, Chandrasekar B, Balakrishnan N (2008) Exact likelihood inference for an exponential parameter under progressive hybrid censoring schemes. In: Vonta F, Nikulin M, Limnios N, Huber-Carol C (eds) Statistical models and methods for biomedical and technical systems. Birkhäuser, Boston, pp 323–334
Cohen AC (1991) Truncated and censored samples: theory and applications. Marcel Dekker, New York
Draper N, Guttman I (1987) Bayesian analysis of hybrid life tests with exponential failure times. Ann Inst Stat Math 39:219–225
Epstein B (1954) Truncated life tests in the exponential case. Ann Math Stat 25:555–564
Gupta RD, Kundu D (1998) Hybrid censoring schemes with exponential failure distribution. Commun Stat Theory Meth 27:3065–3083
Kundu D (2007) On hybrid censoring Weibull distribution. J Stat Plan Infer 137:2127–2142
Kundu D, Joarder A (2006) Analysis of Type-II progressively hybrid censored data. Comput Stat Data Anal 50:2509–2528
Nelson W (1982) Applied life data analysis. Wiley, New York
Ng HKT, Kundu D, Chan PS (2009). Statistical analysis of exponential lifetimes under an adaptive Type-II progressive censoring scheme, Naval Research Logistics 56:687–698

Census

Margo J. Anderson
Professor of History and Urban Studies
University of Wisconsin–Milwaukee, Milwaukee, WI, USA

Introduction

A census usually refers to a complete count by a national government of the population, with the population further defined by demographic, social or economic characteristics, for example, age, sex, ethnic background, marital status, and income. National governments also conduct other types of censuses, particularly of economic activity. An economic census collects information on the number and characteristics of farms, factories, mines, or businesses.

Most countries of the world conduct population censuses at regular intervals. By comparing the results of successive censuses, analysts can see whether the population is growing, stable, or declining, both in the country as a whole and in particular geographic regions. They can also identify general trends in the characteristics of the population. Because censuses aim to count the entire population of a country, they are very expensive and elaborate administrative operations and thus are conducted relatively infrequently. The United States and the United Kingdom, for example, conduct a population census every 10 years (a *decennial* census), and Canada conducts one every 5 years (a *quinquennial* census). Economic censuses are generally conducted on a different schedule from the population census.

Censuses of population usually try to count everyone in the country as of a fixed date, often known as Census Day. Generally, governments collect the information by sending a ►questionnaire in the mail or a census taker to every household or residential address in the country. The recipients are instructed to complete the questionnaire and send it back to the government, which processes the answers. Trained interviewers visit households that do not respond to the questionnaire and individuals without mail service, such as the homeless or those living in remote areas.

History

Censuses have been taken since ancient times by emperors and kings trying to assess the size and strength of their realms. These early censuses were conducted sporadically, generally to levy taxes or for military conscription. Clay tablet fragments from ancient Babylon indicate that a census was taken there as early as 3800 BCE to estimate forthcoming tax revenues. The ancient Chinese, Hebrews, Egyptians, and Greeks also conducted censuses. However, enumerations did not take place at regular intervals until the Romans began to count of the population in the Republic and later the empire. Among the Romans the census was usually a count of the male population and assessment of property value. It was used mainly for drafting men into military service and for taxing property.

After the fall of the Roman Empire in the fifth century CE, census taking disappeared for several hundred years in the West. The small feudal communities of the Middle

Ages had neither the mechanisms nor the need for censuses. However, in 1086 William the Conqueror ordered the compilation of the census-like Domesday Book, a record of English landowners and their holdings. From the data given in this survey, which was made to determine revenues due to the king, historians have reconstructed the social and economic conditions of the times.

The modern census dates from the seventeenth century, when European powers wanted to determine the success of their overseas colonies. Thus the British crown and the British Board of Trade ordered repeated counts of the colonial American population in the seventeenth and eighteenth centuries, starting in the 1620s in Virginia. The first true census in modern times was taken in New France, France's North American empire, beginning in 1665. The rise of democratic governments resulted in a new feature of the census process: The 1790 census of the United States was the first to have its Constitution require a census and periodic reapportionment of its House of Representatives on the basis of the decennial census results. Sweden began to conduct censuses in the mid-eighteenth century, and England and Wales instituted a regular decennial census in 1801. During the nineteenth century and the first half of the twentieth century, the practice of census taking spread throughout the world. India conducted its first national census in 1871, under British rule. China's first modern census, in 1953, counted 583 million people.

The United Nations encourages all countries to conduct a population count through a census or population registration system. It also promotes adoption of uniform standards and census procedures. The United Nations Statistical Office compiles reports on worldwide population.

Uses of Census Information

Governments use census information in almost all aspects of public policy. In some countries, the population census is used to determine the number of representatives each area within the country is legally entitled to elect to the national legislature. The Constitution of the United States, for example, provides that seats in the House of Representatives should be apportioned to the states according to the number of their inhabitants. Each decade, Congress uses the population count to determine how many seats each state should have in the House and in the electoral college, the body that nominally elects the president and vice president of the United States. This process is known as *reapportionment.* States frequently use population census figures as a basis for allocating delegates to the state legislatures and for redrawing district boundaries for seats in the House, in state legislatures, and in local legislative districts. In Canada, census population data are similarly used to apportion seats among the provinces and territories in the House of Commons and to draw electoral districts.

Governments at all levels – such as cities, counties, provinces, and states – find population census information of great value in planning public services because the census tells how many people of each age live in different areas. These governments use census data to determine how many children an educational system must serve, to allocate funds for public buildings such as schools and libraries, and to plan public transportation systems. They can also determine the best locations for new roads, bridges, police departments, fire departments, and services for the elderly.

Besides governments, many others use census data. Private businesses analyze population and economic census data to determine where to locate new factories, shopping malls, or banks; to decide where to advertise particular products; or to compare their own production or sales against the rest of their industry. Community organizations use census information to develop social service programs and child-care centers. Censuses make a huge variety of general statistical information about society available to researchers, journalists, educators, and the general public.

Conducting a Census

Most nations create a permanent national statistical agency to take the census. In the United States, the Bureau of the Census (Census Bureau), an agency of the Department of Commerce, conducts the national population census and most economic censuses. In Canada, the Census Division of Statistics Canada is responsible for taking censuses.

Conducting a census involves four major stages. First, the census agency plans for the census and determines what information it will collect. Next, it collects the information by mailing questionnaires and conducting personal interviews. Then the agency processes and analyzes the data. Finally, the agency publishes the results to make them available to the public and other government agencies.

Planning the Census

Census agencies must begin planning for a census years in advance. One of the most important tasks is to determine what questions will appear on the census questionnaire. Census agencies usually undertake a lengthy public review process to determine the questions to be asked. They conduct public meetings, consider letters and requests from the general public, and consult with other government agencies and special advisory committees. In the United States, census questions must be approved by Congress and

the Office of Management and Budget. In Canada, questions must be approved by the governor-general on the recommendations of the Cabinet.

The questions included on census forms vary from nation to nation depending on the country's particular political and social history and current conditions. Most censuses request basic demographic information, such as the person's name, age, sex, educational background, occupation, and marital status. Many censuses also include questions about a person's race, ethnic or national origin, and religion. Further questions may ask the person's place of birth; relationship to the head of the household; citizenship status; the individual's or the family's income; the type of dwelling the household occupies; and the language spoken in the household.

Questions that are routine in one nation may be seen as quite controversial in another, depending on the history of the country. The United States census does not ask about religious affiliation because such a question is considered a violation of the First Amendment right to freedom of religion or an invasion of privacy. Other nations, such as India, do collect such information. Questions on the number of children born to a woman were quite controversial in China in recent years because of government efforts to limit families to having only one child. In the United States, asking a question on income was considered controversial in 1940 when it was first asked. It is no longer considered as objectionable. Questions change in response to public debate about the state of society. For example, Americans wanted to know which households had radios in 1930, and the census introduced questions on housing quality in 1940. Canadians have recently begun to ask census questions on disability status and on the unpaid work done in the home.

Besides determining the content of the census, census agencies must make many other preparations. Staffing is a major concern for census agencies because censuses in most countries require a huge number of temporary workers to collect and process data. Consequently, census agencies must begin recruiting and training workers months or years in advance. For example, the U.S. Census Bureau had to fill 850,000 temporary, short-term positions to conduct the 2000 census. In order to hire and retain enough staff, it had to recruit nearly three million job applicants. The majority of temporary workers are hired to go door-to-door to interview households that do not respond to the census questionnaire. In some countries, government employees at a local level, such as schoolteachers, are asked to help conduct the count.

Prior to any census, a census agency must develop an accurate list of addresses and maps to ensure that everyone is counted. The U.S. Census Bureau obtains addresses primarily from the United States Postal Service and from previous census address lists. It also works closely with state, local, and tribal governments to compile accurate lists. Finally, census agencies often conduct an extensive marketing campaign before Census Day to remind the general population about the importance of responding to the census. This campaign may involve paid advertising, distributing materials by direct mail, promotional events, and encouraging media coverage of the census.

Collecting the Information

Until relatively recently, population censuses were taken exclusively through personal interviews. The government sent *enumerators* (interviewers) to each household in the country. The enumerators asked the head of the household questions about each member of the household and entered the person's responses on the census questionnaire. The enumerator then returned the responses to the government. Today, many censuses are conducted primarily through *self-enumeration,* which means that people complete their own census questionnaire. Self-enumeration reduces the cost of a census to the government because fewer enumerators are needed to conduct interviews. In addition, the procedure provides greater privacy to the public and generally improves the accuracy of responses, because household members can take more time to think over the questions and consult their personal records.

Nevertheless, census operations still require hiring very large numbers of temporary enumerators to conduct address canvassing in advance of a mail census and to retrieve forms from non responding households and check on vacant units. Other nations continue to conduct censuses partially or totally through direct enumeration. Some, such as Turkey, require people to stay home on Census Day to await the census taker.

Census agencies make a special effort to count people who may not receive a questionnaire by mail or who have no permanent address. For example, the U.S. Census Bureau sends census takers to interview people at homeless shelters, soup kitchens, mobile food vans, campgrounds, fairs, and carnivals. It consults with experts to find migrant and seasonal farmworkers. Finally, the agency distributes census questionnaires to people living in group quarters, such as college dormitories, nursing homes, hospitals, prisons and jails, halfway houses, youth hostels, convents and monasteries, and women's shelters.

The level of detail on the complete count census varies by country, particularly after the development of probability survey techniques in the 1940s. In the United States, for example, until the 2010 census, most households received a "short form," a brief set of questions on basic

characteristics such as name, age, sex, racial or ethnic background, marital status, and relationship to the household head. But from the mid-twentieth century until 2000, a smaller sample of households received the "long form," with many additional detailed questions. These included questions about the individual's educational background, income, occupation, language knowledge, veteran status, and disability status as well as housing-related questions about the value of the individual's home, the number of rooms and bedrooms in it, and the year the structure was built. These "long form" questions have been collected in the American Community Survey since the early 2000s, and thus are no longer asked on the U.S. Census in 2010.

Processing and Analysis of Data

For most of the 19th century in the United States and Canada, census data were tabulated and compiled by hand, without the aid of machines. Manual processing was very slow, and some figures were obsolete by the time they were published. The invention of mechanical tabulating devices in the late nineteenth century made processing of the data much faster and improved the accuracy of the results. For example, in 2010, the U.S. Census Bureau will scan the data from 100 + million paper questionnaires, and capture the responses using optical character recognition software. Once in electronic form, the data can be analyzed and turned into statistics. Unreadable or ambiguous responses are checked by census clerks and manually keyed into the computer.

Publication of Results

U.S. and Canadian censuses publish only general statistical information and keep individual responses confidential. By law, the U.S. Census Bureau and Statistics Canada are prohibited from releasing individual responses to any other government agency or to any individual or business. Census workers in both countries must swear under oath that they will keep individual responses confidential. Employees who violate this policy face a monetary fine and possible prison term. If an individual's personal data were not kept confidential, people might refuse to participate in the census for fear that their personal information would be made public or used by the government to track their activities. In the United States, individual census responses are stored at the National Archives. After 72 years, the original forms are declassified and opened to the public. These original responses are frequently used by people researching the history of their families or constructing genealogies. In Canada, census responses from 1906 and later are stored at Statistics Canada. Microfilmed records of census responses from 1911 and earlier are stored at the National Archives of Canada; these are the only individual census responses currently available for public use.

Until the 1980s, census agencies published their results in large volumes of numeric tables – sometimes numbering in the hundreds of volumes. Today, the majority of census data is distributed electronically, both in tabulated form, and through anonymized public use microdata samples.

Problems in Census Taking and Issues for the Future

Censuses provide important information about the population of a country. But they can become embroiled in political or social controversy simply by reporting information. Complaints about the census generally involve concerns about the accuracy of the count, the propriety of particular questions, and the uses to which the data are put.

All censuses contain errors of various kinds. Some people and addresses are missed. People may misunderstand a question or fail to answer all the questions. Census officials have developed elaborate procedures to catch and correct errors as the data are collected, but some errors remain. For example, the 1990 U.S. census missed 8.4 million people and mistakenly counted 4.4 million people, according to Census Bureau estimates. The latter figure included people counted more than once, fictitious people listed on forms, and fabrications by enumerators. Such errors undermine the credibility of the census as a mechanism for allocating seats in legislative bodies and government funds.

In recent years, developments in statistical analysis have made it possible to measure the accuracy of censuses. Census results may be compared with population information from other sources, such as the records of births, deaths, and marriages in vital statistics. Census officials can also determine the level of accuracy of the count by conducting a second, sample count called a *post-enumeration survey* or *post-censal survey*. In this technique, census staff knock on the door of each housing unit in selected blocks around the country, regardless of whether the housing unit was on the master address list. The staff member determines whether the household was counted in the census. By comparing the results from this survey with the census records, census officials can estimate how many people from each geographic region were missed in the original census count. Some nations, such as Canada and Australia, have begun to adjust the census results for omissions and other errors.

Concerns about the confidentiality of the census represent another source of data error. Censuses require public understanding, support, and cooperation to be successful. Concerns about government interference with private life

can prevent people from cooperating with what is essentially a voluntary counting process. People may be suspicious of giving information to a government agency or may object that particular census questions invade their privacy. When public trust is lacking, people may not participate. In some nations, past census controversies have led to the elimination of the national census. During World War II (1939–1945), for example, the German Nazi forces occupying The Netherlands used the country's census records and population registration data to identify Jews for detention, removal, and extermination. This use ultimately undermined the legitimacy of the census after World War II. In The Netherlands, the legacy of the Nazi era was one of the major justifications to end census taking. The Netherlands took its last regular census in 1971 and now collects population information through other mechanisms.

Many nations are currently exploring alternatives to or major modernizations of the traditional population census. France, for example, has recently implemented a continuous measurement population counting system. The United States is exploring the use of administrative records and electronic methods of data collection to replace the mail enumeration in 2020.

About the Author

Dr. Margo J. Anderson is Professor of History and Urban Studies, University of Wisconsin–Milwaukee. She specializes in the social history of the United States in the nineteenth and twentieth centuries. She was Chair, History Department (1992–1995). She was a member of the National Academy of Sciences' Panel on Census Requirements for the Year 2000 and Beyond. Dr. Anderson was Vice President (2005), and President, Social Science History Association (2006). She is a Fellow, American Statistical Association (ASA) (1998), and was Chair, Social Statistics Section of ASA (1998). Currently, she is ASA Chair, Committee on Committees. She has authored and coauthored numerous papers and several books including, *The American Census: A Social History* (New Haven: Yale University Press, 1988), and *Who Counts? The Politics of Census-Taking in Contemporary America* (with Stephen E. Fienberg, Russell Sage, 1999, revised and updated 2001), named as one of Choice Magazine's Outstanding Academic Books of 2000. She was Editor-in-Chief of the *Encyclopedia of the U.S. Census* (Washington, D.C.: CQ Press, 2000). Professor Anderson is widely regarded as the leading scholar on the history of the U.S. census, and has made distinguished contributions to research in American social science.

Cross References

►African Population Censuses
►Demography
►Economic Statistics
►Federal Statistics in the United States, Some Challenges
►Population Projections
►Sample Survey Methods
►Simple Random Sample
►Small Area Estimation
►Statistical Publications, History of

References and Further Reading

Anderson M (1988) The American census: a social history. Yale University Press, New Haven

Anderson M, Fienberg SE (2001) Who counts? The politics of census taking in contemporary America, rev edn. Russell Sage Foundation, New York

Desrosieres A. La Politique des Grands Nombres: Histoire de la Raison Statistique. Edition La Découverte, Paris (1998, The politics of large numbers: a history of statistical reasoning. Harvard University Press, Cambridge)

Minnesota Population Center (2010) Integrated public use microdata series. http://ipums.org.

U.K. Office of National Statistics (2001) 200 Years of the Census. http://www.statistics.gov.uk/census2001/bicentenary/pdfs/200years.pdf

U.N. Statistics Division (2010) 2010 World Population and Housing Census Programme. http://unstats.un.org/unsd/demographic/sources/census/2010_PHC/more.htm

Ventresca M (1996) When states count: institutional and political dynamics in modern census establishment, 1800–1993. PhD dissertation, Stanford University, Stanford

Worton DA (1997) The Dominion bureau of statistics: a history of Canada's central statistics office and its antecedents: 1841–1972. McGill-Queens University Press, Kingston

Central Limit Theorems

Julio M. Singer
Professor, Head
Universidade de São Paulo, São Paulo, Brazil

Introduction

One of the objectives of statistical inference is to draw conclusions about some parameter, like the mean or the variance of a (possibly conceptual) population of interest based

on the information obtained in a sample conveniently selected therefrom. For practical purposes, estimates of these parameters must be coupled with statistical properties and except in the most simple cases, exact properties are difficult to obtain and one must rely on approximations. It is quite natural to expect estimators to be consistent, but it is even more important that their (usually mathematically complex) exact sampling distribution be adequately approximated by a simpler one, such as the normal or the χ^2 distribution, for which tables or computational algorithms are available. Here we are not concerned with the convergence of the actual sequence of statistics $\{T_n\}$ to some constant or random variable T as $n \to \infty$, but with the convergence of the corresponding distribution functions $\{G_n\}$ to some specific distribution function F. This is known as weak convergence and for simplicity, we write $T_n \xrightarrow{D} F$. Although this is the weakest mode of stochastic convergence, it is very important for statistical applications, since the related limiting distribution function F may generally be employed in the construction of approximate confidence intervals for and significance tests about the parameters of interest. In this context, central limit theorems (CLT) are used to show that statistics expressed as sums of the underlying random variables, conveniently standardized, are asymptotically normally distributed, i.e., converge weakly to the normal distribution. They may be proved under different assumptions regarding the original distributions.

The simplest CLT states that the (sampling) distribution of the sample mean of independent and identically distributed (*i.i.d*) random variables with finite second moments may be approximated by a normal distribution. Although the limiting distribution is continuous, the underlying distribution may even be discrete. CLT are also available for independent, but not identically distributed (e.g., with different means and variances) underlying random variables, provided some (relatively mild) assumptions hold for their moments. The Liapounov CLT and the Lindeberg-Feller CLT are useful examples. Further extensions cover cases of dependent random underlying variables; in particular, the Hájek-Šidak CLT is extremely useful in regression analysis, where as the sample size increases, the response variables form a triangular array in which for each row (i.e., for given n), they are independent but this is not true among rows (i.e., for different values of n). Extensions to cover cases where the underlying random variables have more complex (e.g., martingale-type) dependence structures are also available. When dealing with partial sum or empirical distributional processes, we must go beyond the finite-dimensional case and assume some *compactness* conditions to obtain suitable results, wherein the so-called *weak invariance principles* play an important role.

Different Versions of the Central Limit Theorem

We now present (without proofs) the most commonly used versions of the CLT. Details and a list of related references may be obtained in Sen et al. (2010).

Theorem 1 *(Classical CLT) Let* $\{X_k, k \geq 1\}$ *be a sequence of i.i.d. random variables such that*

1. $\mathbb{E}(X_k) = \mu$.
2. $\mathbb{V}ar(X_k) = \sigma^2 < \infty$.

Also, let $Z_n = (T_n - n\mu)/(\sigma\sqrt{n})$ *where* $T_n = \sum_{k=1}^n X_k$. *Then,* $Z_n \xrightarrow{D} \mathcal{N}(0,1)$.

In practice, this result implies that for large n, the distribution of the sample mean $\overline{X}_n = T_n/n$ may be approximated by a normal distribution with mean μ and variance σ^2/n. An interesting special case occurs when the underlying variables X_k have Bernoulli distributions with probability of success π. Here the expected value and the variance of X_k are π and $\pi(1-\pi)$, respectively. It follows that the large-sample distribution of the sample proportion, $p_n = T_n/n$ may be approximated by a $\mathcal{N}[\pi, \pi(1-\pi)/n]$ distribution. This result is known as the *De Moivre–Laplace CLT*.

An extension of Theorem 1 to cover the case of sums of independent, but not identically distributed random variables requires additional assumptions on the moments of the underling distributions. In this direction, we consider the following result.

Theorem 2 *(Liapounov CLT) Let* $\{X_k, k \geq 1\}$ *be a sequence of independent random variables such that*

1. $\mathbb{E}(X_k) = \mu_k$.
2. $\nu_{2+\delta}^{(k)} = \mathbb{E}(|X_k - \mu_k|^{2+\delta}) < \infty, \quad k \geq 1$ *for some* $0 < \delta \leq 1$.

Also let $T_n = \sum_{k=1}^n X_k$, $\mathbb{V}ar(X_k) = \sigma_k^2$, $\tau_n^2 = \sum_{k=1}^n \sigma_k^2$, $Z_n = (T_n - \sum_{k=1}^n \mu_k)/\tau_n$ *and* $\rho_n = \tau_n^{-(2+\delta)} \sum_{k=1}^n \nu_{2+\delta}^{(k)}$. *Then, if* $\lim_{n\to\infty} \rho_n = 0$, *it follows that* $Z_n \xrightarrow{D} \mathcal{N}(0,1)$.

This as well as other versions of the CLT may also be extended to the multivariate case by referring to the *Cramér-Wold Theorem*, which essentially states that the asymptotic distribution of the multivariate statistic under

investigation may be obtained by showing that every linear combination of its components follows an asymptotic normal distribution. Given a sequence $\{\boldsymbol{X}_n, n \geq 1\}$ of random vectors in $\mathbb{R}^p$, with mean vectors $\boldsymbol{\mu}_n$ and covariance matrices $\boldsymbol{\Sigma}_n$, $n \geq 1$, to show that $n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}_i) \xrightarrow{D} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\boldsymbol{\Sigma}_i$, one generally proceeds according to the following strategy:

1. Use one of the univariate CLT to show that for every fixed $\boldsymbol{\lambda} \in \mathbb{R}^p$, $n^{-1/2}\sum_{i=1}^{n}\boldsymbol{\lambda}'(\boldsymbol{X}_i - \boldsymbol{\mu}_i) \xrightarrow{D} \mathcal{N}(0, \gamma^2)$ with $\gamma^2 = \lim_{n\to\infty} n^{-1}\boldsymbol{\lambda}'\left(\sum_{i=1}^{n}\boldsymbol{\Sigma}_i\right)\boldsymbol{\lambda}$.
2. Use the Cramér-Wold Theorem to complete the proof.

As an example we have:

Theorem 3 *(Multivariate version of the Liapounov CLT) Let $\{\boldsymbol{X}_n, n \geq 1\}$ be a sequence of random vectors in $\mathbb{R}^p$ with mean vectors $\boldsymbol{\mu}_n$ and finite covariance matrices $\boldsymbol{\Sigma}_n, n \geq 1$, such that $\max_{1\leq i\leq n}\max_{1\leq j\leq p}\mathbb{E}(|X_{ij} - \mu_{ij}|^{2+\delta}) < \infty$ for some $0 < \delta < 1$, and $\boldsymbol{\Sigma} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\boldsymbol{\Sigma}_i$ exists. Then $n^{-1/2}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}_i) \xrightarrow{D} \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$.*

In the original formulation, Liapounov used $\delta = 1$, but even the existence of $v_{2+\delta}^{(k)}$, $0 < \delta \leq 1$ is not a necessary condition, as we may see from the following theorem.

Theorem 4 *(Lindeberg-Feller CLT) Let $\{X_k, k \geq 1\}$ be a sequence of independent random variables satisfying*

1. $\mathbb{E}(X_k) = \mu_k$.
2. $\mathbb{V}ar(X_k) = \sigma_k^2 < \infty$.

Also, let $T_n = \sum_{k=1}^{n}X_k$, $\tau_n^2 = \sum_{k=1}^{n}\sigma_k^2$ and $Z_n = \sum_{k=1}^{n}Y_{nk}$ where $Y_{nk} = (X_k - \mu_k)/\tau_n$ and consider the following additional conditions:

1. *Uniform asymptotic negligibility (UAN):* $\max_{1\leq k\leq n}\left(\sigma_k^2/\tau_n^2\right) \to 0$ *as* $n \to \infty$.
2. *Asymptotic normality:* $Z_n \xrightarrow{D} \mathcal{N}(0,1)$.
3. *Lindeberg-Feller (uniform integrability):*

$$\forall \varepsilon > 0, \frac{1}{\tau_n^2}\sum_{k=1}^{n}\mathbb{E}\left[(X_k - \mu_k)^2 I(|X_k - \mu_k| > \varepsilon\tau_n)\right] \to 0 \text{ as } n \to \infty,$$

where $I(A)$ denotes the indicator function.

Then, (A) and (B) hold simultaneously if and only if (C) holds.

Condition (A) implies that the random variables Y_{nk} are *infinitesimal*, i.e., that $\max_{1\leq k\leq n}P(|Y_{nk}| > \varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$, or, in other words, that the random variables Y_{nk}, $1 \leq k \leq n$, are uniformly in k, asymptotically in n, negligible.

When the underlying random variables under consideration are bounded, i.e., when $P(a \leq X_k \leq b) = 1$ for some finite scalars $a < b$, it follows that a necessary and sufficient condition for $Z_n \xrightarrow{D} \mathcal{N}(0,1)$ is that $\tau_n \to \infty$ as $n \to \infty$.

Up to this point we have devoted attention to the weak convergence of sequences of statistics $\{T_n, n \geq 1\}$ constructed from independent underlying random variables $X_1, X_2, \ldots$. We consider now some extensions of the CLT where such restriction may be relaxed. The first of such extensions holds for sequences of (possibly dependent) random variables which may be structured as a *double array* of the form

$$\begin{pmatrix} X_{11}, & X_{12}, & \cdots, & X_{1k_1} \\ X_{21}, & X_{22}, & \cdots, & X_{2k_2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1}, & X_{n2}, & \cdots, & X_{nk_n} \end{pmatrix}$$

where the X_{nk} are row-wise independent. The case where $k_n = n$, $n \geq 1$, is usually termed a *triangular array* of random variables. This result is very useful in the field of ▶order statistics.

Theorem 5 *(Double array CLT) Let the random variables $\{Y_{nk}, 1 \leq k \leq k_n, n \geq 1\}$ where $k_n \to \infty$ as $n \to \infty$ be such that for each n, $\{Y_{nk}, 1 \leq k \leq k_n\}$ are independent. Then*

1. *$\{Y_{nk}, 1 \leq k \leq k_n, n \geq 1\}$ is an infinitesimal system of random variables, i.e., satisfies the UAN condition.*
2. *$Z_n = \sum_{k=1}^{k_n}Y_{nk} \xrightarrow{D} \mathcal{N}(0,1)$.*

hold simultaneously, if and only if, for every $\varepsilon > 0$, as $n \to \infty$ the following two conditions hold

1. $\sum_{k=1}^{k_n}P(|Y_{nk}| > \varepsilon) \to 0$.
2. $\sum_{k=1}^{k_n}\left\{\int_{\{|y|\leq\varepsilon\}} y^2 dP(Y_{nk} \leq y) - \left[\int_{\{|y|\leq\varepsilon\}} y dP(Y_{nk} \leq x)\right]^2\right\} \to 1$.

Linear regression and related models pose special problems since the underlying random variables are not identically distributed and in many cases, the exact functional form of their distributions is not completely specified. Least-squares methods (see ▶Least Squares) are attractive under these conditions, since they may be employed in a rather general setup. In this context, the following CLT specifies sufficient conditions on the explanatory variables such that the distributions of the least squares estimators of the regression parameters may be approximated by normal distributions.

Theorem 6 *(Hájek-Šidak CLT) Let* $\{Y_n, n \geq 1\}$ *be a sequence of i.i.d. random variables with mean* μ *and finite variance* σ^2*; let* $\{\boldsymbol{x}_n, n \geq 1\}$ *be a sequence of real vectors* $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nn})'$*. Then if Noether's condition holds, i.e., if*

$$\max_{1 \leq i \leq n} \left[x_{ni}^2 \Big/ \sum_{i=1}^{n} x_{ni}^2 \right] \to 0 \quad as \quad n \to \infty,$$

holds, it follows that

$$Z_n = \left[\sum_{i=1}^{n} x_{ni}(Y_{ni} - \mu) \right] \Big/ \left[\sigma^2 \sum_{i=1}^{n} x_{ni}^2 \right]^{1/2} \xrightarrow{D} \mathcal{N}(0,1).$$

As an illustration, consider the simple linear regression model (see ▶Simple Linear Regression)

$$y_{ni} = \alpha + \beta x_{ni} + e_{ni}, \; i = 1, \ldots, n,$$

where y_{ni} and x_{ni} represent observations of the response and explanatory variables, respectively, α and β are the parameters of interest and the e_{ni} correspond to uncorrelated random errors with mean 0 and variance σ^2. The least squares estimators of β and α are respectively $\widehat{\beta}_n = \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)(y_{ni} - \bar{y}_n) / \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2$ and $\widehat{\alpha}_n = \bar{y}_n - \widehat{\beta}_n \bar{x}_n$ where $\bar{x}_n$ and $\bar{y}_n$ correspond to the sample means of the explanatory and response variables. Irrespectively of the form of underlying distribution of e_{ni}, we may use standard results to show that $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ are unbiased and have variances given by $\sigma^2 \left[\sum_{i=1}^{n} x_{ni}^2 / \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2 \right]$ and $\sigma^2 \left[\sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2 \right]^{-1}$, respectively. Furthermore, the covariance between $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ is $-\sigma^2 \bar{x}_n / \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2$. When the underlying distribution of e_{ni} is normal, we may show that $(\widehat{\alpha}_n, \widehat{\beta}_n)$ follows a bivariate normal distribution. If Noether's condition holds and both $\bar{x}_n$ and $n^{-1} \sum_{i=1}^{n}(x_{ni} - \bar{x}_n)^2$ converge to finite constants as $n \to \infty$, we may use the Hájek-Šidak CLT and the Cramér-Wold Theorem to conclude that the same bivariate normal distribution specified above serves as an approximation of the true distribution of $(\widehat{\alpha}_n, \widehat{\beta}_n)$, whatever the form of the distribution of e_{ni}, provided that n is sufficiently large.

The results may also be generalized to cover alternative estimators obtained by means of generalized and weighted least-squares procedures as well as via robust M-estimation procedures. They may also be extended to generalized linear and nonlinear models. Details may be obtained in Sen et al. (2010), for example.

It is still possible to relax further the independence assumption on the underlying random variables. The following theorems constitute examples of CLT for dependent random variables having a martingale (or reverse martingale) structure. For further details, the reader is referred to Loynes (1970), Brown (1971), Dvoretzky (1971), or McLeish (1974).

Theorem 7 *(Martingale CLT) Consider a sequence* $\{X_k, k \geq 1\}$ *of random variables satisfying*

1. $\mathbb{E}(X_k) = 0$.
2. $\mathbb{E}\left(X_k^2\right) = \sigma_k^2 < \infty$.
3. $\mathbb{E}\{X_k | X_1, \ldots, X_{k-1}\} = 0$, $X_0 = 0$.

Also let $T_n = \sum_{k=1}^{n} X_k$, $\tau_n^2 = \sum_{k=1}^{n} \sigma_k^2$, $v_k^2 = \mathbb{E}\left(X_k^2 | X_1, \ldots, X_{k-1}\right)$ *and* $w_n^2 = \sum_{k=1}^{n} v_k^2$*. If*

1. $w_n^2 / \tau_n^2 \xrightarrow{P} 1$ *as* $n \to \infty$.
2. $\forall \varepsilon > 0$, $\tau_n^{-2} \sum_{k=1}^{n} \mathbb{E}\left[X_k^2 I(|X_k| > \varepsilon \tau_n) \right] \to 0$ *as* $n \to \infty$ *(Lindeberg-Feller condition),*

then the sequence $\{X_k, k \geq 1\}$ *is infinitesimal and* $Z_n = T_n / \tau_n \xrightarrow{D} \mathcal{N}(0,1)$.

Note that the terms v_k^2 are random variables since they depend on $X_1, \ldots, X_{k-1}$; condition (A) essentially states that all the information about the variability in the X_k is contained in $X_1, \ldots, X_{k-1}$. Also note that $\{T_n, n \geq 1\}$ is a *zero mean martingale* (See also ▶Martingale Central Limit Theorem.)

Theorem 8 *(Reverse Martingale CLT) Consider a sequence* $\{T_k, k \geq 1\}$ *of random variables such that*

$$\mathbb{E}(T_n | T_{n+1}, T_{n+2}, \ldots) = T_{n+1} \quad and \quad \mathbb{E}(T_n) = 0,$$

i.e., $\{T_k, k \geq 1\}$ *is a zero mean reverse martingale. Assume that* $\mathbb{E}\left(T_n^2\right) < \infty$ *and let* $Y_k = T_k - T_{k+1}, k \geq 1$, $v_k^2 = \mathbb{E}\left(Y_k^2 | T_{k+1}, T_{k+2}, \ldots\right)$ *and* $w_n^2 = \sum_{k=n}^{\infty} v_k^2$*. If*

1. $w_n^2 / \mathbb{E}\left(w_n^2\right) \xrightarrow{a.s.} 1$.
2. $w_n^{-2} \sum_{k=n}^{\infty} \mathbb{E}\left[Y_k^2 I(|Y_k| > \varepsilon w_n) \middle| T_{k+1}, T_{k+2}, \ldots \right] \xrightarrow{P} 0$, $\varepsilon > 0$ *or* $w_n^{-2} \sum_{k=n}^{\infty} Y_k^2 \xrightarrow{a.s.} 1$,

it follows that $T_n / \sqrt{\mathbb{E}\left(w_n^2\right)} \xrightarrow{D} \mathcal{N}(0,1)$.

Rates of Convergence to Normality

In the general context discussed above, a question of both theoretical and practical interest concerns the speed with which the convergence to the limiting normal distribution takes place. Although there are no simple answers to this question, the following result may be useful.

Theorem 9 *(Berry-Esséen) Let* $\{X_n, n \geq 1\}$ *be a sequence of i.i.d. random variables with* $\mathbb{E}(X_1) = \mu$, $\mathbb{V}ar(X_1) = \sigma^2$ *and suppose that* $\mathbb{E}(|X_1 - \mu|^{2+\delta}) = \nu_{2+\delta} < \infty$ *for*

some $0 < \delta \leq 1$. *Also let* $T_n = \sum_{i=1}^{n} X_i$ *and* $F_{(n)}(x) = P\left[(T_n - n\mu)/\left(\sigma\sqrt{n}\right) \leq x\right]$, $x \in \mathbb{R}$. *Then there exist a constant* C *such that*

$$\Delta_n = \sup_{x\in\mathbb{R}} |F_{(n)}(x) - \Phi(x)| \leq C\frac{\nu_{2+\delta} n^{-\delta/2}}{\sigma^{2+\delta}}$$

where Φ *denotes the standard normal distribution function.*

The reader is referred to Feller (1971) for details. Berry (1941) proved the result for $\delta = 1$ and Esséen (1956) showed that $C \geq 0.4097$. Although the exact value of the constant C is not known, many authors have proposed upper bounds. In particular, van Beeck (1972) showed that $C \leq 0.7975$ and more recently, Shevtsova (2007) concluded that $C \leq 0.7056$. The usefulness of the theorem, however, is limited, since the rates of convergence attained are not very sharp.

Alternatively, the rates of convergence of the sequence of distribution functions $F_{(n)}$ to Φ or of the density functions $f_{(n)}$ (when they exist) to φ (the density function of the standard normal distribution) may be assessed by *Gram-Charlier* or *Edgeworth expansions* as discussed in Cramér (1946), for example. Although this second approach might offer a better insight to the problem of evaluating the rate of convergence to normality than that provided by the former, it requires the knowledge of the moments of the parent distribution and, thus, is less useful in practical applications.

Convergence of Moments

Given that weak convergence has been established, a question of interest is whether the moments (e.g., mean and variance) of the statistics under investigation converge to the moments of the limiting distribution. Although the answer is negative in general, an important theorem, due to Cramér, indicates conditions under which the result is true. The reader is referred to Sen et al. (2010) for details.

Asymptotic Distributions of Statistics not Expressible as Sums of Random Variables

The *Slustky theorem* is a handy tool to prove weak convergence of statistics that may be expressed as the sum, product or ratio of two terms, the first known to converge weakly to some distribution and the second known to converge in probability to some constant. As an example, consider independent and identically distributed random variables $Y_1, \ldots, Y_n$ with mean μ and variance σ^2. Since the corresponding sample standard deviation S converges in probability to σ and the distribution of $\overline{Y}$ may be approximated by a $\mathcal{N}(\mu, \sigma^2/n)$ distribution, we may apply Slutsky's theorem to show that the large-sample distribution of $\sqrt{n}\,\overline{Y}/S = \left(\sqrt{n}\,\overline{Y}/\sigma\right) \times (\sigma/S)$ may be approximated by a $\mathcal{N}(\mu, 1)$ distribution. This allows us to construct approximate confidence intervals for and tests of hypotheses about μ using the standard normal distribution. A similar approach may be employed to the Bernoulli example by noting that p_n is a consistent estimator of π.

An important application of Slutky's Theorem relates to statistics that can be decomposed as a sum of a term for which some CLT holds and a term that converges in probability to 0. Assume, for example, that the variables Y_i have a finite fourth central moment γ and write the sample variance as

$$S^2 = [n/(n-1)]\left\{n^{-1}\sum_{i=1}^{n}[(Y_i - \mu)^2 - \sigma^2/n] + \left[\sigma^2 - \sum_{i=1}^{n}(\overline{Y} - \mu)^2\right]\right\}.$$

Since the first term within the $\{\}$ brackets is known to converge weakly to a normal distribution by the CLT and the second term converges in probability to 0, we conclude that the distribution of S^2 may be approximated by a $\mathcal{N}(\sigma^2, \gamma/n)$ distribution. This is the basis of the projection results suggested by Hoeffding (1948) and extensively explored by Jurečkova and Sen (1996) to obtain large-sample properties of ▶*U*-statistics as well as of more general classes of estimators.

Another convenient technique to obtain the asymptotic distributions of many (smooth) functions of asymptotically normal statistics is the *Delta-method*: if g is a locally differentiable function of a statistic T_n whose distribution may be approximated (for large samples) by a $\mathcal{N}(\mu, \tau^2)$ distribution, then the distribution of the statistic $g(T_n)$ may be approximated by a $\mathcal{N}\{g(\mu), [g'(\mu)]^2\tau^2\}$ distribution, where $g'(\mu)$ denotes the first derivative of g computed at μ. Suppose that we are interested in estimating the odds of a failed versus pass response, i.e., $\pi/(1-\pi)$ in an exam based on a sample of n students. A straightforward application of the De Moivre Laplace CLT may be used to show that the estimator of π, namely, k/n, where k is the number of students that failed the exam, follows an approximate $\mathcal{N}[\pi, \pi(1-\pi)/n]$ distribution. Taking $g(x) = x/(1-x)$, we may use the Delta-method to show that the distribution of the sample odds $k/(n-k)$ may be approximated by a $\mathcal{N}\{\pi/(1-\pi), \pi/[n(1-\pi)^3]\}$ distribution. This type of result has further applications in variance-stabilizing transformations used in cases (as the above example) where the variance of the original statistic depends on the parameter it is set to estimate.

For some important cases, like the Pearson χ^2-statistic or more general quadratic forms $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\mu}) = (\mathbf{Y} - \boldsymbol{\mu})^t \boldsymbol{A}(\mathbf{Y} - \boldsymbol{\mu})$ where $\mathbf{Y}$ is a p-dimensional random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$ and $\mathbf{A}$ is a p-dimensional square matrix of full rank, the (multivariate) Delta-method may not be employed because the derivative of Q computed at $\boldsymbol{\mu}$ is null. If $\mathbf{A}$ converges to an inverse of $\mathbf{V}$, a useful result known as the *Cochran theorem*, states that the distribution of Q may be approximated by a χ^2 instead of a normal distribution. In fact, the theorem holds even if $\mathbf{A}$ is not of full rank, but converges to a generalized inverse of $\mathbf{V}$. This is important for applications in categorical data.

The CLT also does not hold for extreme order statistics like the sample minimum or maximum; depending on some regularity conditions on the underlying random variables, the distribution of such statistics, conveniently normalized, may be approximated by one of three types of distributions, namely the extreme value distributions of the first, second or third type, which, in this context, are the only possible limiting distributions as shown by Gnedenko (1943).

Central Limit Theorems for Stochastic Processes

Empirical distribution functions and ▶order statistics have important applications in nonparametric regression models, resampling methods like the ▶*jackknife* and *bootstrap* (see ▶Bootstrap Methods), sequential testing as well as in Survival and Reliability analysis. In particular it serves as the basis for the well known goodness-of-fit *Kolmogorov-Smirnov* and *Cramér-von Mises statistics* and for *L*- and *R-estimators* like *trimmed* or *Winsorized means*. Given the sample observations $Y_1, \ldots, Y_n$ assumed to follow some distribution function F and a real number y, the empirical distribution function is defined as

$$F_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i \le y)$$

where $I(Y_i \le y)$ is an indicator function assuming the value 1 if $Y_i \le y$ and 0, otherwise. It is intimately related to the order statistics, $Y_{n:1} \le Y_{n:2} \le \ldots \le Y_{n:n}$ where $Y_{n:1}$ is the smallest among $Y_1, \ldots, Y_n$, $Y_{n:2}$ is the second smallest and so on. For each fixed sample, F_n is a distribution function when considered as a function of y. For every fixed y, when considered as a function of $Y_1, \ldots, Y_n$, $F_n(y)$ is a random variable; in this context, since the $I(Y_i \le y), i = 1, \ldots, n$, are independent and identically distributed zero-one valued random variables, we may apply the classical CLT to conclude that for each fixed y the distribution of $F_n(y)$ may be approximated by a $\mathcal{N}\{F(y), F(y)[1-F(y)]/n\}$ distribution provided that n is sufficiently large. In fact, using standard asymptotic results, we may show that given any finite number m of points $y_1, \ldots, y_m$, the distribution function of the vector $[F_n(y_1), \ldots, F_n(y_m)]$ may be approximated by a multivariate normal distribution function. This property is known as *convergence of finite-dimensional distributions*.

On the other hand, $F_n - F = \{F_n(y) - F(y) : y \in \mathbb{R}\}$ is a random function defined on the set of real numbers, and, hence, to study its various properties we may need more than the results considered so far. Note that as the sample size n increases, so does the cardinality of the set of order statistics used to define the empirical distribution function and we may not be able to approximate this n-dimensional joint distribution by an m-dimensional one unless some further *tightness* or *compactness* conditions are imposed on the underlying distributions. This is the basis of the weak invariance principles necessary to show the convergence of empirical and other ▶stochastic processes to *Brownian bridge* or *Brownian motion* processes. An outline of the rationale underlying these results follows.

Let $t = F(y)$ and $W_n^0(t) = \sqrt{n}[G_n(t) - t]$, $t \in (0,1)$ where $G_n(t) = F_n[F^{-1}(t)] = F_n(y)$ with $F^{-1}(x) = \inf\{y : F(y) > x\}$, so that $\{W_n^0(t),\ t \in (0,1)\}$ is a stochastic process with $\mathbb{E}[W_n^0(t)] = 0$ and $\mathbb{E}[W_n^0(s)W_n^0(t)] = \min(s,t) - st$, $0 \le s,t \le 1$. Using the multivariate version of the CLT we may show that as $n \to \infty$, for all $m \ge 1$, given $0 \le t_1 \le \ldots \le t_m \le 1$, the vector $\boldsymbol{W}_{nm}^0 = [W_n^0(t_1), \ldots, W_n^0(t_m)] \xrightarrow{D} [W^0(t_1), \ldots, W^0(t_m)] = \boldsymbol{W}_m^0$ where $\boldsymbol{W}_m^0$ follows a $\mathcal{N}_m(\mathbf{0}, \boldsymbol{\Gamma}_m)$ distribution with $\boldsymbol{\Gamma}_m$ denoting a positive definite matrix with elements $\min(t_i, t_j) - t_i t_j$, $i,j = 1, \ldots, m$.

Now, define a stochastic process $\{Z(t),\ t \in (0,1)\}$ with independent and homogeneous increments such that, for every $0 \le s < t \le 1$, the difference $Z(t) - Z(s)$ follows a $\mathcal{N}(0, t-s)$ distribution. Then, it follows that $\mathbb{E}[Z(s)Z(t)] = \min(s,t)$. This process is known as a *standard Brownian motion* or *standard Wiener process*. Furthermore, letting $W^0(t) = Z(t) - tZ(1)$, $0 \le t \le 1$, it follows that $\{W^0(t),\ t \in (0,1)\}$ is also a Gaussian stochastic process such that $\mathbb{E}[W^0(t)] = 0$ and $\mathbb{E}[W^0(s)W^0(t)] = \min(s,t) - st$, $0 \le s,t \le 1$. Then for all $m \ge 1$, given $0 \le t_1 \le \ldots \le t_m \le 1$, the vector $\boldsymbol{W}_m^0 = [W^0(t_1), \ldots, W^0(t_m)]$ also follows a $\mathcal{N}_m(\mathbf{0}, \boldsymbol{\Gamma}_m)$ distribution. Since $W^0(0) = W^0(1) = 0$ with probability 1, this process is called a *tied down Wiener process* or *Brownian bridge*.

Using *the Kolmogorov maximal inequality*, we may show that $\{W_n^0(t),\ t \in (0,1)\}$ is *tight* and referring to standard results in weak convergence of probability measures, we may conclude that $\{W_n^0(t),\ t \in (0,1)\} \xrightarrow{D}$

$\{W^0(t),\ t \in (0,1)\}$. Details and extensions to statistical functionals may be obtained in Jurečkova and Sen (1996) among others.

About the Author

Dr. Julio da Motta Singer is Professor of Biostatistics and Head, Centre for Applied Statistics, University of São Paulo, Brazil. He has obtained his Ph.D. at the University of North Carolina in 1983 (advisor P.K. Sen). He has coauthored over 80 refereed papers and several books including, *Large Sample Methods in Statistics: An Introduction with Applications* (with P.K. Sen, Chapman and Hall, 1993), and *From finite sample to asymptotic methods in Statistics* (with P.K. Sen and A.C. Pedroso-de-Lima, Cambridge University Press, 2010).

Cross References

►Almost Sure Convergence of Random Variables
►Approximations to Distributions
►Asymptotic Normality
►Asymptotic Relative Efficiency in Estimation
►Asymptotic Relative Efficiency in Testing
►Empirical Processes
►Limit Theorems of Probability Theory
►Martingale Central Limit Theorem
►Normal Distribution, Univariate
►Probability Theory: An Outline

References and Further Reading

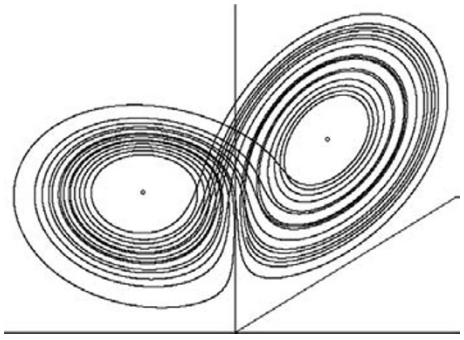
Billingsley P (1968) Convergence of probability measures. Wiley, New York
Berry AC (1941) The accuracy of the Gaussian approximation to the sum of independent variates. Trans Am Math Soc 49:122–136
Brown BM (1971) Martingale central limit theorems. Ann Math Stat 42:59–66
Chow YS, Teicher H (1978) Probability theory: independence, interchangeability, martingales Springer, New York
Cramér H (1946) Mathematical methods of statistics. Princeton University Press, Princeton
Dvoretzky A (1971) Asymptotic normality for sums of dependent random variables. In: Proceedings of the sixth berkeley symposium on mathematical statistics and probability. University of California Press, Berkeley, vol 2, pp 513–535
Esséen CG (1971) A moment inequality with an application to the central limit theorem. Skandinavisk Aktuarietidskrift 39: 160–170
Feller W (1971) An introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York
Ferguson TS (1996) A course in large sample theory. Chapman & Hall, London
Gnedenko BV (1943) Sur la distribution limite du terme maximum d'une série aléatoire. Ann Math 44:423–453
Hoeffding W (1948) A class of statistics with asymptotically normal distributions. Ann Math Stat 19:293–325
Jurečková J, Sen PK (1996) Robust statistical procedures. Wiley, New York
Lehmann EL (2004) Elements of large sample theory. Springer, New York
Loynes RM (1970) An invariance principle for reversed martingales. Proc Am Math Soc 25:56–64
McLeish DL (1974) Dependent central limit theorems and invariance principles. Ann Probab 2:620–628
Reiss RD (1989) Approximate distributions of order statistics with applications to nonparametric statistics. Springer, New York
Sen PK (1981) Sequential nonparametrics: invariance principles and statistical inference. Wiley, New York
Sen PK, Singer JM, Pedroso-de-Lima AC (2010) From finite sample to asymptotic methods in statistics. Cambridge University Press, Cambridge
Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York
Shevtsova IG (1974) Sharpening the upper bound of the absolute constant in the Berry–Esséen inequality. Theory Probab Appl 51:549–533
van Beeck P (1972) An application of Fourier methods to the problem of sharpening the Berry–Esséen inequality. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 23: 187–197
van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, New York

Chaotic Modelling

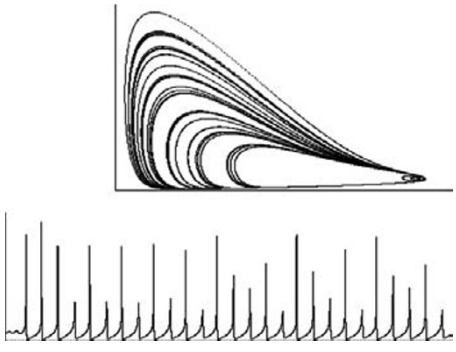
Christos H. Skiadas
Professor, Director of the Data Analysis and Forecasting Laboratory
Technical University of Crete, Chania, Greece

Chaotic modeling is a term used to express the representation of the state of a system or a process by using chaotic models or tools developed in the chaotic literature and the related scientific context. In the following we present the main elements of the chaotic modeling including chaotic terms, differential and difference equations and main theorems (Skiadas 2009).

Chaos is a relatively new science mainly developed during last decades with the use of computers and supercomputers. It touches almost all the scientific fields. However, the basic elements can be found at the end of the nineteenth century and the attempts to solve the famous three-body problem by Henri Poincaré (1890). Although he succeeded to solve only the special case when the three bodies move in the same plane, he could explore the main characteristics of the general three-body problem and to see the unpredictability of the resulting paths in space. In other words he could realize the main characteristic of



Chaotic Modelling. Fig. 1 The Lorenz attractor (xyz view)



Chaotic Modelling. Fig. 2 Autocatalytic attractor and chaotic oscillations

a chaotic process that very small changes in initial conditions have significant impact to the future states of a system.

This was verified by Edwin Lorenz in 1963 with his work on modeling the atmospheric changes. He reduced the Navier-Stokes equations, used to express fluid flows, to a system of three nonlinear coupling differential equations and performed simulations in a computer trying to model the weather changes. He surprisingly found that the system was very sensitive to small changes of initial conditions thus making the forecasts of the future weather unpredictable. Famous are the forms of his simulated paths that look like a butterfly with open wings. The three-dimensional model which he proposed has the form (σ, r and b are parameters):

$$\dot{x} = -\sigma x + \sigma y,\ \dot{y} = -xz + rx - y,\ \dot{z} = xy - bz.$$

The famous Lorenz attractor also known as the butterfly attractor is illustrated in Fig. 1.

Several years later Rössler (1976) proposed a simpler three-dimensional model including only one nonlinear term thus verifying the assumption that a set of simple differential equations with only one nonlinear term may express chaotic behavior. The Rössler system is the following (e, f and m are parameters):

$$\dot{x} = -y - z,\ \dot{y} = x - ez,\ \dot{z} = f + xz - mz.$$

It can be verified that the number of chaotic parameters is equal to the number of the equations.

Chemical chaotic oscillations where observed by Belousov (1959) and later on by Zhabotinsky (1964) when they where working with chemical autocatalytic reactions. The Nobel Prize in chemistry (1977) was awarded to Prigogine for his work on dynamics of dissipative systems (see Prigogine 1961) including the mathematical representation of autocatalytic reactions. A simple autocatalytic reaction is expressed by the following set of three differential equations:

$$\dot{x} = \left(\frac{1}{1+k} + m\right)(k+z) - xy^2 - x,\ \dot{y} = \frac{xy^2 + x - y}{e}, \dot{z} = y - z$$

This model is illustrated in Fig. 2; the parameters set are: $e = 0.013, k = 2.5, m = 0.017$.

The use of computing power gave rise to the exploration of chaos in astronomy and astrophysics. A paper that influenced much the future developments of the chaotic applications was due to Hénon and Heiles in 1964. They had predicted chaos in Hamiltonian systems that could apply to astronomy and astrophysics. Few years before George Contopoulos (1958) had also found chaotic behavior when he explored the paths of stars in a galaxy. That it was most important was that they had shown that the computer experiments had much more to show than simply verify the results coming from the mathematical formulations. Hidden and unexplored scientific fields would emerge by the use of computers.

It was found that chaos could emerge from a system of three or more differential equations with at list one nonlinear term. This comes from the Poincaré–Bendixson theorem which states that a two dimensional system of nonlinear equations may have a regular behavior.

Another theorem is the famous *KAM* theorem from the initials of the names of Kolmogorov, Arnold and Moser. This theorem applies to dynamical systems and may explain the stability or not of these systems to small perturbations. It is interesting that the chaotic forms could be quite stable as it happens for vortex and tornados.

However, the main scientific discovery on chaos came only in 1978 by Michel Feigenbaum when he found that the simple logistic map could produce a chaotic sequence. Feigenbaum tried a difference equation instead of the differential equations that where used in the previous works on chaos. That is different is that chaos can emerge even

from only one difference equation with at list one nonlinear term. This is because a difference equation defines a recurrence scheme which is a set of numerous equations in which every equation uses the outcomes from the preceding one. The complexity resulting from a nonlinear difference equation is large and it can be measured with a power law of the number of iterations.

In the logistic model a mapping into itself is defined by the difference equation and gives rise to period doubling bifurcations and chaos for a specific range of the chaotic parameter. The logistic map is of the form: $x_{n+1} = bx_n(1 - x_n)$, where b is the chaotic parameter and x_n is the chaotic function (see a (x_{n+1}, x_n) diagram of the Logistic model in Fig. 3; $b = 2.9$).

For the logistic map as also for other maps there exists the bifurcation diagram. This is a diagram, usually two dimensional, defining the bifurcation points with respect to the chaotic parameter or parameters (see Fig. 4).

The chaotic modeling has also to do with *strange attractors* by means forms in space that have a great detail and complexity. These forms can arise in nature and also can be simulated from chaotic equations. A very interesting future of a chaotic attractor is that for a variety of initial conditions the chaotic system leads the final results or solutions to a specific area, the strange or chaotic attractor.

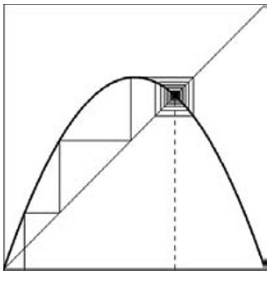
Chaos may also arise from a set of two or more difference equations with at least one nonlinear term. The most popular model is the Hénon (1976) model given by:

$$x_{n+1} = y_n + 1 - ax_n^2,\ y_{n+1} = bx_n.$$

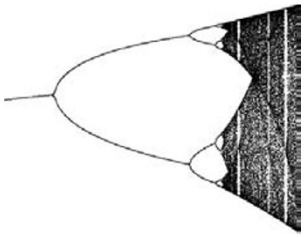
The Jacobian determinant of this model is:

$$\det J = \begin{vmatrix} \frac{\partial x_{n+1}}{\partial x_n} & \frac{\partial y_{n+1}}{\partial x_n} \\ \frac{\partial x_{n+1}}{\partial y_n} & \frac{\partial y_{n+1}}{\partial y_n} \end{vmatrix} = -b.$$

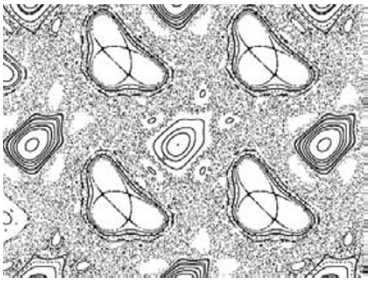
The system is stable for $0 < b < 1$. When $b = 1$ the system is area preserving, but it is unstable.



Chaotic Modelling. Fig. 3 The logistic model



Chaotic Modelling. Fig. 4 The bifurcation diagram



Chaotic Modelling. Fig. 5 A carpet-like form

An alternative of the Hénon map is provided by the following cosine model:
$x_{n+1} = by_n + 2a\cos(x_n) - 2a + 1,\ y_{n+1} = x_n$. This map provides a carpet-like form (see Fig. 5) for $b = -1$ and $a = -0.6$.

Very many cases in nature have to do with delays. This mathematically can be modeled by a delay differential or difference equation. Simpler is to use difference equations to express delay cases. An example is the transformation of the previous Hénon map to the corresponding delay difference equation of the form:

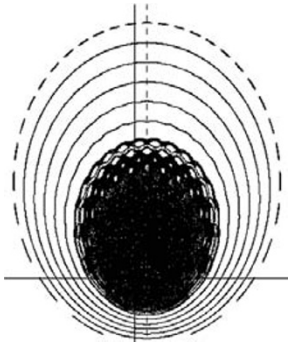
$$x_{n+1} = bx_{n-1} + 1 - ax_n^2.$$

This delay differential equation has the same properties of the Hénon map. In general modeling delays leads to differential or difference equations which produce oscillations and may produce chaos for appropriate selection of the parameters. One of the first proposed chaotic models including delays is the famous Mackey-Glass (1977) model regarding oscillation and chaos in physiological control systems.

Ikeda found his famous attractor in 1979 (see Fig. 6; parameters $a = 1, b = 0.83, c = 0.4$ and $d = 6$) when he was experimenting on the light transmitted by a ring cavity system. The equations' set is:

$$x_{n+1} = a + b(x_n\cos(\varphi_n) - y_n\sin(\varphi_n)),$$
$$y_{n+1} = b(x_n\sin(\varphi_n) + y_n\cos(\varphi_n)),$$

Chaotic Modelling. Fig. 6 The Ikeda attractor



Chaotic Modelling. Fig. 7 Chaotic rotating forms

where the rotation angle is: $\varphi_n = c - \dfrac{d}{1 + x_n^2 + y_n^2}$

The last form of difference equations express a rotation-translation phenomenon and can give very interesting chaotic forms (see Skiadas 2009). Figure 7 illustrates such a case where the rotation angle follows an inverse low regarding the distance r from the origin: $\varphi_n = \dfrac{c}{\sqrt{x_n^2 + y_n^2}} = \dfrac{c}{r}$. A chaotic bulge is located in the central part followed by elliptic trajectories in the outer part (the parameters are: $a = 0.6$ and $b = c = 1$).

Other interesting aspects of chaotic modeling are found in numerous publications regarding control of chaos with applications in various fields.

Chaotic mixing and chaotic advection have also studied with chaotic models as well as economic and social systems.

About the Author

Christos H. Skiadas is the Founder and Director of the Data Analysis and Forecasting Laboratory, Technical University of Crete (TUC). He was Vice-Rector of the Technical University of Crete (1997–1999), Chairman of the Production Engineering and Management Department, TUC (1995–1997) and participated in many committees and served as a consultant in various firms while he directed several scientific and applied projects. He was a visiting fellow in the University of Exeter, UK (1986) and Université Libre de Bruxelles, Belgium (1993–1994). He has been the guest-editor of the journals *Communications in Statistics*, *Methodology and Computing in Applied Probability* (MCAP) and *Applied Stochastic Models and Data Analysis*. He is a member of the Committee and Secretary of ASMDA International Society and was the Chair or co-chair of the International Conferences: 6th ASMDA 1993, 12th ASMDA 2007. As a member of the Greek Statistical Institute he organised and co-chaired two Annual National Conferences of the Institute (11th Chania 1998 and 22nd Chania 2009). Professor Skiadas contributions mainly are directed to the modeling and simulation of innovation diffusion and application of the findings in various fields as finance and insurance, energy consumption, forecasting and modeling. The related publications include more than 90 papers and 10 books, including *Chaotic Modelling and Simulation: Analysis of Chaotic Models, Attractors and Forms* (with Charilaos Skiadas, Chapman and Hall/CRC Press, 2009).

Cross References

►Stochastic Modeling, Recent Advances in

References and Further Reading

Belousov BP (1959) Периодически действующая реакция и ее механизм. [A periodic reaction and its mechanism]. Сборник рефератов по радиационной медицине (*Compilation of Abstracts on Radiation Medicine*) 147:145

Contopoulos G (1958) On the vertical motions of stars in a galaxy. *Stockholm Ann* 20(5):20

Hénon M (1976) A two-dimensional mapping with a strange attractor. *Commun Math Phys* 50:69–77

Hénon M, Heiles C (1964) The applicability of the third integral of motion: some numerical experiments. *Astron J* 69:73–79

Feigenbaum MJ (1978) Quantitative universality for a class of nonlinear transformations. *J Stat Phys* 19:25–52

Ikeda K (1979) Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system. *Opt Commun* 30:257–261

Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20:130–141

Mackey MC, Glass L (1977) Oscillation and chaos in physiological control systems. *Science* 197:287–289

Poincaré H (1890) Sur les équations de la dynamique et le problème de trois corps. *Acta Math* 13:1–270

Prigogine I (1961) *Thermodynamics of irreversible processes*, 2nd edn. Interscience, New York

Rössler OE (1976) An equation for continuous chaos. *Phys Lett A* 57:397–398

Skiadas CH, Skiadas C (2009) *Chaotic modeling and simulation: analysis of chaotic models attractors and forms.* Taylor & Francis/CRC Press, London

Zhabotinsky AM (1964) Периодический процесс окисления малоновой кислоты растворе (исследование кинетики реакции Белоусова). [Periodic processes of malonic acid oxidation in a liquid phase.] Биофизика [*Biofizika*] 9:306–311

Characteristic Functions

Miljenko Huzak
University of Zagreb, Zagreb, Croatia

Characteristic functions play an outstanding role in the theory of probability and mathematical statistics (Ushakov 1999). The characteristic function (c.f.) of a probability distribution function (d.f.) is the Fourier–Stieltjes transform of the d.f. More precisely, if F is a probability d.f. on d-dimensional real space R^d ($d \geq 1$), then its c.f. is a complex function $\phi : \mathrm{R}^d \to \mathrm{C}$ such that for any $\mathbf{t} = (t_1, \ldots, t_d) \in \mathrm{R}^d$,

$$\begin{aligned}\phi(\mathbf{t}) &= \int_{\mathrm{R}^d} e^{i\sum_{j=1}^d t_j x_j}\, dF(x_1, \ldots, x_d) := \\ &= \int_{\mathrm{R}^d} \cos\left(\sum_{j=1}^d t_j x_j\right) dF(x_1, \ldots, x_d) \\ &\quad + i \int_{\mathrm{R}^d} \sin\left(\sum_{j=1}^d t_j x_j\right) dF(x_1, \ldots, x_d),\end{aligned}$$

where the integrals are Lebesgue–Stieltjes integrals with respect to d.f. F.

If $\mathbf{X} = (X_1, \ldots, X_d)$ is a d-dimensional random vector, then c.f. $\phi = \phi_{\mathbf{X}}$ associated to $\mathbf{X}$ is the c.f. of its d.f. $F = F_{\mathbf{X}}$. Hence

$$\phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[e^{i\sum_{j=1}^d t_j X_j}\right], \quad \mathbf{t} = (t_1, \ldots, t_d) \in \mathrm{R}^d. \tag{1}$$

Particularly, c.f. $\phi = \phi_X : \mathrm{R} \to \mathrm{C}$ of a random variable (r.v.) X is equal to

$$\phi(t) = \mathbb{E}[e^{itX}], \quad t \in \mathrm{R}.$$

Examples of c.f.s of some r.v.s are in Table 1.

C.f.s have many good properties (see Table 2). One of the most important properties of c.f.s is that there is a one-to-one correspondence between d.f.s and their c.f.s, which is a consequence of the *Lévy inversion formula* (see Chow and Teicher 1988 or Feller 1971). Since it is usually simpler to manipulate with c.f.s than with corresponding d.f.s, this property makes c.f.s useful in proving many theorems on probability distributions. For example, it can be proved that the components of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ are independent r.v.s if and only if

$$\begin{aligned}&(\forall t_1, \ldots, t_d \in \mathrm{R}) \;\; \phi_{\mathbf{X}}(t_1, \ldots, t_d) \\ &= \phi_{X_1}(t_1) \cdot \phi_{X_2}(t_2) \ldots \phi_{X_d}(t_d).\end{aligned}$$

Moreover, since for any independent r.v.s $X_1, X_2, \ldots, X_n$, c.f. of their sum $S_n = X_1 + \ldots + X_n$ is equal to the product of their c.f.s, to obtain the d.f. of S_n, it is usually easier

Characteristic Functions. Table 1 Characteristic functions of some univariate probability distributions

Distribution	Density $f(x)$	c.f. $\phi(t)$
Degenerate at c		e^{itc}
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$	$(pe^{it} + 1 - p)^n$
Poisson	$e^{-\lambda}\frac{\lambda^x}{x!}$	$\exp\{\lambda(e^{it} - 1)\}$
Normal	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$	$e^{i\mu t - \sigma^2 t^2/2}$
Symmetric uniform over $(-\theta, \theta)$	$\frac{1}{2\theta}$	$\frac{\sin \theta t}{\theta t}$
Gamma	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$	$(1 - it\beta)^{-\alpha}$
Cauchy	$\frac{\alpha}{\pi(\alpha^2 + x^2)}$	$e^{-\alpha\lvert t\rvert}$

Characteristic Functions. Table 2 List of properties of characteristic functions $\phi_{\mathbf{X}}(\mathbf{t})$ given by (1) (the list follows one from Ferguson (1996))

(1)	$\phi_{\mathbf{X}}(\mathbf{t})$ exists for all $\mathbf{t} \in \mathrm{R}^d$ and is continuous.
(2)	$\phi_{\mathbf{X}}(\mathbf{0}) = 1$ and $\lvert\phi_{\mathbf{X}}(\mathbf{t})\rvert \leq 1$ for all $\mathbf{t} \in \mathrm{R}^d$.
(3)	For a scalar a, $\phi_{a\mathbf{X}}(\mathbf{t}) = \phi_{\mathbf{X}}(a\mathbf{t})$.
(4)	For a matrix A and a vector $\mathbf{c}$, $\phi_{A\mathbf{X}+\mathbf{c}}(\mathbf{t}) = e^{i\mathbf{t}^\tau \mathbf{c}} \cdot \phi_{\mathbf{X}}(A^\tau \mathbf{t})$.
(5)	For $\mathbf{X}$ and $\mathbf{Y}$ independent, $\phi_{\mathbf{X}+\mathbf{Y}}(\mathbf{t}) = \phi_{\mathbf{X}}(\mathbf{t})\phi_{\mathbf{Y}}(\mathbf{t})$.
(6)	If $\mathbb{E}\lvert\mathbf{X}\rvert < \infty$, $\dot{\phi}_{\mathbf{X}}(\mathbf{t})$ exists and is continuous and $\dot{\phi}_{\mathbf{X}}(\mathbf{0}) = i\,\mathbb{E}\,\mathbf{X}^\tau$.
(7)	If $\mathbb{E}[\lvert\mathbf{X}\rvert^2] < \infty$, $\ddot{\phi}_{\mathbf{X}}(\mathbf{t})$ exists and is continuous and $\ddot{\phi}_{\mathbf{X}}(\mathbf{0}) = -\mathbb{E}[\mathbf{X}\mathbf{X}^\tau]$.
(8)	If $P(\mathbf{X} = \mathbf{c}) = 1$ for a vector $\mathbf{c}$, $\phi_{\mathbf{X}}(\mathbf{t}) = e^{i\mathbf{t}^\tau \mathbf{c}}$.
(9)	If $\mathbf{X}$ is normal r. vec. with $\boldsymbol{\mu} = \mathbb{E}\,\mathbf{X}$ and $\mathrm{cov}(\mathbf{X}) = \Sigma$, $\phi_{\mathbf{X}}(\mathbf{t}) = \exp\left\{i\mathbf{t}^\tau \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\tau \Sigma \mathbf{t}\right\}$.

to find the c.f. of their sum and to apply the Lévy inversion formula than to find the convolution of their d.f.s.

Another very important property of c.f.s comes from the *continuity theorem* (see Chow and Teicher 1988 or Feller 1971): r.v.s X_n, $n \geq 1$, with corresponding c.f.s ϕ_n, $n \geq 1$, converge in law to a r.v. X with c.f. ϕ if and only if c.f.s ϕ_n, $n \geq 1$, converge to ϕ pointwise. For example, this property makes proving ►central limit theorems easier if not only possible.

C.f.s have been important tools in developing theories of infinite divisible and particularly stable distributions (e.g., see Feller 1971; Chow and Teicher 1988).

Cross References

►Probability on Compact Lie Groups

References and Further Reading

Chow YS, Teicher H (1988) Probability theory, independence, interchangeabilty, martingales. 2nd edn. Springer-Verlag, New York

Feller W (1971) An introduction to probability theory and its applications, Vol 2, 2nd edn. Wiley, New York

Ferguson TS (1996) A course in large sample theory. Chapman & Hall, London

Lukacs E (1970) Characteristic functions. 2nd edn. Griffin, London

Ushakov NG (1999) Selected topics in characteristic functions. Brill Academic Publishers, Leiden

Chebyshev's Inequality

Gerold Alsmeyer
Professor
Institut für Mathematische Statistik, Münster, Germany

Chebyshev's inequality is one of the most common inequalities used in probability theory to bound the tail probabilities of a random variable X having finite variance $\sigma^2 = \mathrm{Var}X$. It states that

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{for all } t>0, \tag{1}$$

where $\mu = \mathbb{E}X$ denotes the mean of X. Of course, the given bound is of use only if t is bigger than the standard deviation σ. Instead of proving (1) we will give a proof of the more general *Markov's inequality* which states that for any nondecreasing function $g : [0,\infty) \to [0,\infty)$ and any nonnegative random variable Y

$$\mathbb{P}(Y \geq t) \leq \frac{\mathbb{E}g(Y)}{g(t)} \quad \text{for all } t>0. \tag{2}$$

Indeed, choosing $Y = |X-\mu|$ and $g(x) = x^2$ gives (1). The proof of Markov's inequality is very easy: For any $t > 0$,

$$\mathbb{P}(Y \geq t) = \int 1_{\{Y\geq t\}} d\mathbb{P} \leq \int_{\{Y\geq t\}} \frac{g(Y)}{g(t)} d\mathbb{P} \leq \frac{\mathbb{E}g(X)}{g(t)}.$$

Plainly, (1) provides us with the same bound $\sigma^2 t^{-2}$ for the one-sided tail probability $\mathbb{P}(X-\mu > t)$, but in this case an improvement is obtained by the following consideration: For any $c \geq 0$, we infer from Markov's inequality with $g(x) = x^2$

$$\mathbb{P}(X-\mu \geq t) = \mathbb{P}(X-\mu+c \geq t+c) \leq \frac{\mathbb{E}(X-\mu+c)^2}{(t+c)^2} = \frac{\sigma^2+c^2}{(t+c)^2}.$$

The right-hand side becomes minimal at $c = \sigma^2/t$ giving the one-sided tail bound

$$\mathbb{P}(X-\mu > t) \leq \frac{\sigma^2}{\sigma^2+t^2} \quad \text{for all } t>0, \tag{3}$$

sometimes called *Cantelli's inequality*.

Although Chebyshev's inequality may produce only a rather crude bound its advantage lies in the fact that it applies to any random variable with finite variance. Moreover, within the class of all such random variables the bound is indeed tight because, if X has a symmetric distribution on $\{-a,0,a\}$ with $\mathbb{P}(X = \pm a) = 1/(2a^2)$ and $\mathbb{P}(X=0) = 1-1/a^2$ for some $a > 1$, then $\mu = 0$, $\sigma^2 = 1$ and

$$\mathbb{P}(|X| \geq a) = \mathbb{P}(|X| = a) = \frac{1}{a^2},$$

which means that equality holds in (1) for $t = a$.

On the other hand, tighter bounds can be obtained when imposing additional conditions on the considered distributions. On such example is the following *Vysočanskiĭ-Petunīn inequality* for random variables X with an unimodal distribution:

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{4\sigma^2}{9t^2} \quad \text{for all } t > \sqrt{3/8}\,\sigma, \tag{4}$$

This improves (1) by a factor 4/9 for sufficiently large t.

One of the most common applications of Chebyshev's inequality is the weak law of large numbers (WLLN). Suppose we are given a sequence $(S_n)_{n\geq 1}$ of real-valued random variables with independent increments $X_1, X_2, \ldots$ such that $\mu_n := \mathbb{E}X_n$ and $\sigma_n^2 := \mathrm{Var}X_n$ are finite for all $n \geq 1$. Defining

$$m_n := \mathbb{E}S_n = \sum_{k=1}^n \mu_k \quad \text{and} \quad s_n^2 := \mathrm{Var}S_n = \sum_{k=1}^n \sigma_k^2$$

and assuming *Markov's condition*

$$\lim_{n\to\infty} \frac{s_n^2}{n^2} = 0 \tag{5}$$

we infer by making use of (1) that, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n - m_n}{n}\right| \geq \epsilon\right) \leq \frac{s_n^2}{\epsilon^2 n^2} \to 0 \quad \text{as } n \to \infty$$

and therefore

$$\frac{S_n - m_n}{n} \to 0 \quad \text{in probability.} \tag{WLLN}$$

This result applies particularly to the case of i.i.d. $X_1, X_2, \ldots$ Then $m_n = n\mu$ and $s_n^2 = n\sigma^2$ where $\mu := \mathbb{E}X_1$ and $\sigma^2 := \mathrm{Var}X_1$. In this case, Chebyshev's inequality further gives, for all $\epsilon, \beta > 0$, that

$$\sum_{n\geq 1} \mathbb{P}\left(\left|\frac{S_n - n\mu}{n}\right| \geq \epsilon \log^\beta n\right) \leq \sum_{n\geq 1} \frac{\sigma^2}{\epsilon^2 n \log^{2\beta} n} < \infty$$

and thus, by invoking the Borel-Cantelli lemma (see ►Borel–Cantelli Lemma and Its Generalizations),

$$\frac{S_n - n\mu}{n \log^\beta n} \to 0 \quad \text{a.s. for all } \beta > 0 \tag{6}$$

This is not quite the strong law of large numbers ($\beta = 0$) but gets close to it. In fact, in order for this to derive, a stronger variant of Chebyshev's inequality, called *Kolmogorov's inequality*, may be employed which states that

$$\mathbb{P}\left(\max_{1\leq k\leq n} |S_k - m_k| \geq t\right) \leq \frac{s_n^2}{t^2} \quad \text{for all } t > 0$$

under the same independence assumptions stated above for the WLLN. Notice the similarity to Chebyshev's inequality in that only $S_n - m_n$ has been replaced with $\max_{1\leq k\leq n}(S_k - m_k)$ while retaining the bound.

Let us finally note that, if X has mean μ, median m and finite variance σ^2, then the one-sided version (3) of Chebyshev's inequality shows that

$$\mathbb{P}(X - \mu \geq \sigma) \leq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X - \mu \leq -\sigma) \leq \frac{1}{2},$$

in other words, the median of X is always with one standard deviation of it mean.

Bibliographical notes: (1) dates back to Chebyshev's original work (Chebyshev 1867), but is nowadays found in any standard textbook on probability theory, like (Feller 1971). The latter contains also a proof of the one-sided version (3) which differs from the one given here. (4) for unimodal distributions is taken from Vysočanskiĭ and Petunĭn (1979), see also Sellke and Sellke (1997). For multivariate extensions of Chebyshev's inequality see Olkin and Pratt (1958) and Monhor (2007).

About the Author

Dr. Alsmeyer is Professor at the Department of Mathematics and Computer Science of the University of Münster and was the Chairman from 2000 till 2008. He has written more than 50 research articles and one book. He has supervised eight Ph.D. students.

Cross References

►Borel–Cantelli Lemma and Its Generalizations
►Expected Value
►Laws of Large Numbers
►Standard Deviation

References and Further Reading

Chebyshev P (1867) Des valeurs moyennes. Liouville's J Math Pure Appl 12:177–184
Feller W (1971) An introduction to probability theory and its applications, vol II, 2nd edn. Wiley, New York
Monhor D (2007) A Chebyshev inequality for multivariate normal distribution. Prob Eng Inform Sci 21(2):289–300
Olkin I, Pratt JW (1958) A multivariate Tchebycheff inequality. Ann Math Stat 29:226–234
Sellke TM, Sellke SH (1997) Chebyshev inequalities for unimodal distributions. Am Stat 51(1):34–40
Vysočanskiĭ DF, Petunĭn JĬ (1979) Proof of the 3σ rule for unimodal distributions. Teor Veroyatnost i Mat Stat 21:23–35

Chemometrics

Rolf Sundberg
Professor of Mathematical Statistics
Stockholm University, Stockholm, Sweden

The role of statistics in chemistry is over a century old, going back to the Guinness brewery chemist and experimenter Gosset, more well-known under the pseudonym "Student." For his applications, he was in need of small-sample statistical methods. Until the 1970s, chemistry methods and instruments were typically univariate, but in that decade analytical chemistry and some other branches of chemistry had to start handling data of multivariate character. For example, instead of measuring light intensity at only a single selected wavelength, instruments became available that could measure intensities at several different wavelengths at the same time. The instrumental capacity rapidly increased, and the multivariate spectral dimension soon exceeded the number of chemical samples analysed (the "$n < p$" problem). In parallel, other chemists worked with Quantitative Structure–Activity Relationships (QSAR), where they tried to explain and predict biological activity or similar properties of a molecule from

a large number of structural physical-chemical characteristics of the molecule, but having an empirical data set of only a moderate number of different molecules. Generally, as soon as multivariate data are of high dimension, we must expect near collinearities among the variables, and when $n < p$, there are necessarily exact collinearities. These were some of the problems faced, long before statisticians got used to $n < p$ in genomics, proteomics etc. This was the birth of the field of chemometrics, a name coined by Svante Wold to characterize these research and application activities.

A standard definition of *Chemometrics* would be of type "The development and use of mathematical and statistical methods for applications in chemistry," with more weight on statistical than mathematical. Another characterization, formulated by Wold, is that the aim of chemometrics is to provide methods for

- How to get chemically relevant information out of measured chemical data
- How to get it into data
- How to represent and display this information

and that in order to achieve this, chemometrics is heavily dependent on statistics, mathematics and computer science. The first task is much concerned with analysis of dependencies and relationships (regression, calibration, discrimination, etc.) within a multivariate framework, because complex chemical systems are by necessity multidimensional. The second task is largely represented by experimental design, both classical and newer, where chemometrics has contributed the idea of design in latent factors (principal variates). For representation of high-dimensional data, projection on a low-dimensional latent variable space is the principal tool. Using diagrams in latent factors from PCA or other dimension-reducing methods is also a way of displaying the information found.

Another type of definition, often quoted, is that "Chemometrics is what chemometricians do." This is not only to laugh at. A vital part of chemometrics is connected with chemistry, but the methods developed might be and are applied in quite different fields, where the data analysis problems are similar, such as metabolomics, food science, sensometrics, and image analysis. This could motivate to distinguish chemometrics and chemometric methods, where the latter could as well be described as statistical methods originally inspired by problems in chemistry.

A statistician's look at the contents of *Journal of Chemometrics* for the period 2003–2005 (150 papers) showed that regression and calibration dominated, covering a third of the contents. Much of this was on regularized regression methods, such as PCR (Principal Components Regression) and PLSR (Partial Least Squares Regression). Other statistical areas represented were multiway methods (where each observation is a matrix or an even higher-dimensional array, see Smilde et al. 2004), classification (discrimination and clustering), multivariate statistical process control, and occasionally other areas, for example experimental design, wavelets, genetic algorithms.

A difference between chemometrics and biometrics (►biostatistics) is that chemometricians are mostly chemists by principal education, with more or less of additional statistical education, whereas biometricians are typically statisticians by education. This has had several consequences for chemometrics. Statistical methods are sometimes reinvented. Metods are sometimes proposed without a theoretical underpinning. The popular method of partial least squares (see ►Partial Least Squares Regression Versus Other Methods) is a good such example, nowadays relatively well understood, but proposed and advocated as a computational algorithm, that was widely regarded with suspicion among statisticians. Thus there is often a role for theoretical statistical studies to achieve a deeper understanding of the chemometric methods and their properties, not least to reveal how various suggested methods relate to each others.

About the Author

Professor Sundberg is Past President of the Swedish Statistical Association (1979–1981). He is an Elected member of the International Statistical Institute (1984). He was an Associate editor for *Scandinavian Journal of Statistics* (1987–1994). In 2004 he was awarded (with Marie Linder) the Kowalski prize for best theoretical paper in *Journal of Chemometrics* (2002–2003).

Cross References

►Sensometrics

References and Further Reading

Regression and calibration

Brown PJ (1993) Measurement, regression, and calibration. Oxford University Press, Oxford

Martens H, Næs T (1989) Multivariate calibration. Wiley, Chichester

Sundberg R (1999) Multivariate calibration – direct and indirect regression methodology (with discussion). Scand J Stat 26: 161–207

Other fields of chemometrics

Carlson R, Carlson JE (2005) Design and optimization in organic synthesis, 2nd edn. Elsevier, Amsterdam

Martens H, Martens M (2001) Multivariate analysis of quality. An introduction. Wiley, Chichester

Smilde A, Bro R, Geladi P (2004) Multi-way analysis with applications in the chemical sciences. Wiley, Chichester

Two journals are devoted to chemometrics, started in 1986/87
Journal of Chemometrics. John Wiley & Sons.
Chemometrics and Intelligent Laboratory Systems. Elsevier.

There are several introductions to chemometrics written for chemists, not listed here. Not all of them are satisfactory in their more statistical parts.

Chernoff Bound

Herman Chernoff
Professor Emeritus
Harvard University, Cambridge, MA, USA

The Chernoff Bound, due to Herman Rubin, states that if $\overline{X}$ is the average of n independent observations on a random variable X with mean $\mu < a$ then, for all t,

$$P(\overline{X} > a) \le \left[E(e^{t(X-a)})\right]^n.$$

The proof which follows shortly is a simple application of the Markov inequality that states that for a positive random variable $Y, P(Y \ge b) \le E(Y)/b$, for $b > 0$. The Chernoff bound was a step in the early development of the important field "Large Deviation Theory." It became prominent among computer scientists because of its usefulness in Information Theory.

The Markov inequality is derived from the fact that for $b > 0$,

$$E(Y) = \int ydF(Y) \ge \int_b^\infty ydF(y) \ge bP(Y \ge b)$$

where F is the cumulative distribution of Y.

We observe that

$$E(e^{nt(\overline{X}-a)}) = \left[E(e^{t(X-a)})\right]^n$$

and hence $P(e^{nt(\overline{X}-a)} \ge 1)$ is less than or equal to the bound. This implies the Chernoff bound for $t > 0$. For $t \le 0$ the inequality is automatically satisfied because the bound is at least one. That follows form the fact that the ►moment generating function $M(t) = E(e^{tZ})$ is convex with $M(0) = 1, M'(0) = E(Z)$ and $E(X - a) < 0$.

The prominence of the bound is due to a natural inclination to extend beyond its proper range of applicability. The Central Limit Theorem (see ►Central Limit Theorems), for which an informal statement is that $\overline{X}$ is approximately normally distributed with mean $\mu = E(X)$ and variance σ^2/n where σ is the standard deviation of X. For large deviations (see ►Large Deviations and Applications), or many standard deviations from the mean, the theorem implies that the probability of exceeding a would approach zero, but a naive interpretation would state that this probability would be approximately $\exp(-na^2/2)(2\pi na)^{-1/2}$ and could be seriously wrong.

In 1951, for a special problem of testing a simple hypothesis versus a simple alternative using a statistic of the form $\overline{X}$ where X could take on a few integer values, I realized that the normal approximation was inappropriate. I derived (Chernoff 1952), for $a > E(X)$,

$$n^{-1} \log P(\overline{X} > a) \to \inf_t E(e^{t(X-a)})$$

which was, as far as I know, the first application of Large Deviation Theory to Statistical Inference. This result was used to define a measure of information useful for experimental design and to show that the Kullback-Leibler information numbers (Kullback and Leibler 1951; Chernoff 1956) measure the exponential rate at which one error probability approaches zero when the other is held constant.

At the time I was informed of Cramér's (1938) earlier elegant derivation of more encompassing results, using exponentially tilted distributions. Cramér dealt with deviations which were not limited to those of order square root of n standard deviations, but required a condition that excluded the case which I needed, where the range of X was a multiple of the integers. Blackwell and Hodges (1959) later dealt with that case.

One of my colleagues, Herman Rubin, claimed that he could derive my results more simply, and when I challenged him, he produced the upper bound that I included in my manuscript. At the time the proof seemed so trivial, that I did not mention that it was his. I made two serious errors. First, the inequality is stronger than the upper limit implied by my result, and therefore deserves mention of authorship even though the derivation is simple. Second, because I was primarily interested in the exponential rate at which the probability approached zero, it did not occur to me that this trivially derived bound could become prominent.

About the Author

For biography *see* the entry ►Chernoff–Savage Theorem.

Cross References

►Central Limit Theorems
►Chebyshev's Inequality
►Kullback-Leibler Divergence
►Large Deviations and Applications

References and Further Reading

Blackwell D, Hodges JL (1959) The probability in the extreme tail of a convolution. Ann Math Stat 30:1113–1120

Chernoff H (1952) A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann Math Stat 23:495–507

Chernoff H (1956) Large-sample theory: parametric case. Ann Math Stat 27:1–22

Cramér H (1938) Sus un nouveau théorème-limite de la théorie des probabilités. Acutalités Scientifiques et Industrielles, 736, Paris

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Chernoff Faces

Herman Chernoff
Professor Emeritus
Harvard University, Cambridge, MA, USA

The graphical representation of two dimensional variables is rather straightforward. Three dimensional variables presents more of a challenge, but dealing with higher dimensions is much more difficult. Two methods using profiles and stars suffers from a confusion of which variable is represented when the dimensionality is greater than six.

The method called "Chernoff Faces" (Chernoff 1973) involves a computer program which draws a caricature of a face when given 18 numbers between 0 and 1. These numbers correspond to features of the face. Thus one may represent the length of the nose, another curvature of the mouth, and a third the size of the eyes. If we have 12 dimensional data, we can adjoin 6 constants to get points in 18 dimensional space, each represented by a face. As the point moves in 18 dimensional space the face changes.

The method was developed in response to a problem in cluster analysis (see ►Cluster Analysis: An Introduction). There are many methods proposed to do clustering. It seems that an appropriate method should depend on the nature of the data, which is difficult to comprehend without visualization. The grouping of faces which look alike serves as a preliminary method of clustering and of recognizing which features are important in the clustering.

In the two original applications of the method, the scientists involved claimed that the implementation was lucky because the features which were most important were represented respectively by the size of the eyes and the shape of the face, both of which are prominent features. I claimed that it did not matter which features were selected for the various variables and challenged the scientists to select an alternative choice of features for the variables to degrade the effect of the faces. Their candidate choices had little degradation effect.

To test the conjecture that the choice of variables would have no effect, Rizvi and I carried out an experiment (Chernoff and Rizvi 1975). Of course it is clear that the conjecture cannot be absolutely sound, since the position of the pupils in the eyes cannot be detected if the eyes are small and other features interact similarly. However we set up an experiment where subjects were supposed to cluster 36 faces into two groups of approximately 18 each. The faces were generated from two six dimensional ►multivariate normal distributions with means δ units apart, in Mahalanobis distance, and identity covariance matrix. These data were then subjected to a linear transformation to an 18-dimensional space, and 12 feature selections were made at random. The subjects were given three clustering problem. For the first δ was so large that there was no problem recognizing the clusters. That was a practice problem to train the students in the experiment. For the other two problems two choices of δ were made to establish greater difficulty in separating the two distributions. The result of this experiment was that when the error rate in clustering varies from 8% to 22%, the typical random permutations could change the error rate by a proportion which decreases from 45% to 18%.

Originally, Faces were designed to serve to understand which variables were important and which interacted with each other. Once such relations are understood, analytic methods could be used to probe further. In many applications, Faces could also be used to comprehend data where the roles of the various factors were well understood. For example, in business applications, a smiling face could indicate that some aspect of the business was doing well. With training of the users, such applications could be useful in providing a quick and easy comprehension of a moderately complicated system. For example, one could use a face to represent the many meters an airplane pilot watches, so that he could be alerted when the face begins to look strange. The method of stars could also serve such a function.

Jacob (1978) used faces to represent five particular scales of the Minnesota Multiphasic Personality Inventory (MMPI). The scales represented Hypochondriasis, Depression, Paranoia, Schizophrenia and Hypomania. Realizing that training a psychologist to recognize a smiling face as belonging to a depressed patient would be difficult, he developed an innovative approach to selecting features for the five scales. He presented a random selection of faces

to some psychologists and asked them to rate these faces on the MMPI scales. Then he used regression techniques to decide how the numerical values of an MMPI scale should be translated into features of the face, so that the face presented to a psychologist would resemble that of a person with those scaled values. This would facilitate the process of training psychologists to interpret the faces.

The method of Faces handles many dimensions well. For more than 18 variables, one could use a pair of faces. It does not deal so well with a large number of faces unless we have a time series in which they appear in succession. In that case they can be used to detect changes in characteristics of important complicated systems.

It seems that face recognition among humans is handled by a different part of the brain than that handling other geometrical data and humans are sensitive to very small changes in faces. Also, it seems that cartoons and caricatures of faces are better remembered than realistic representations.

Before the computer revolution, graphical representations, such as nomograms, could be used to substitute for accurate calculations. The Faces are unlikely to be useful for calculation purposes.

About the Author

For biography *see* the entry ▶Chernoff–Savage Theorem.

Cross References

▶Cluster Analysis: An Introduction

References and Further Reading

Chernoff H (1973) The use of faces to represent points in k-dimensional space graphically. J Am Stat Assoc 68:301–308

Chernoff H, Rizvi MH (1975) Effect on classification error of random permutations of features in representing multivariate data by faces. J Am Stat Assoc 70:548–554

Jacob RJK (1978) Facial representation of multivariate data. In: Wang PCC (ed) Graphical representation of multivariate data. Academic, New York, pp 143–168

Chernoff-Savage Theorem

Herman Chernoff
Professor Emeritus
Harvard University, Cambridge, MA, USA

Hodges and Lehmann (1956) conjectured in 1956 that the nonparametric competitor to the t-test, the Fisher-Yates-Terry-Hoeffding or c_1 test (Terry 1952), was as efficient as the t-test for normal alternatives and more efficient for nonnormal alternatives.

To be more precise, we assume that we have two large samples, of sizes m and n with $N = m + n$, from two distributions which are the same except for a translation parameter which differs by an amount δ. To test the hypothesis that $\delta = 0$ against one sided alternatives, we use a test statistic of the form

$$T_N = m^{-1} \sum_{i=1}^{N} E_{Ni} z_{Ni}$$

where z_{Ni} is one or zero depending on whether the ith smallest of the N observations is from the first or the second sample. For example the Wilcoxon test is of the above form with $E_{Ni} = i/N$. It was more convenient to represent the test in the form

$$T_N = \int_{-\infty}^{\infty} J_N[H_N(x)]\, dF_m(x).$$

where F_m and G_n are the two sample cdf's, $\lambda_N = m/N$ and $H_N = \lambda_N F_m + (1 - \lambda_N) G_n$. These two forms are equivalent when $E_{Ni} = J_N(i/N)$.

The proof of the conjecture required two arguments. One was the ▶asymptotic normality of T when $\delta \neq 0$. The Chernoff-Savage theorem (Chernoff and Savage 1958) establishes the asymptotic normality, under appropriate regularity conditions on J_N, satisfied by c_1, using an argument where F_m and G_n are approximated by continuous time ▶Gaussian Processes, and the errors due to the approximation are shown to be relatively small.

The second argument required a variational result using the Pitman measure of local efficacy of the test of $\delta = 0$, which may be calculated as a function of the underlying distribution. For distributions with variance 1, the efficiency of the test relative to the t-test is minimized with a value of 1 for the normal distribution. It follows that the c_1 test is as efficient as the t-test for normal translation alternatives and more efficient for nonnormal translation alternatives.

About the Author

Dr. Herman Chernoff (born in New York City on July 1, 1923) is Professor Emeritus of Statistics at Harvard University and Emeritus Professor at M.I.T. He received a PhD in Applied Mathematics at Brown University in 1948 under the supervision of Abraham Wald (at Columbia University). Dr. Chernoff worked for the Cowles Commission at the University of Chicago and then spent three years in the Mathematics Department at the University of Illinois before joining the Department of Statistics at Stanford University in 1952, where he remained for 22 years.

He moved to M.I.T. in 1974, where he founded the Statistics Center. Since 1985 he has been in the Department of Statistics at Harvard. He retired from Harvard in 1997. Professor Chernoff was President of the Institute of Mathematical Statistics (1967–1968) and is an Elected member of both the American Academy of Arts and Sciences and the National Academy of Sciences. He has been honored for his contributions in many ways. He is a recipient of the Townsend Harris Medal and Samuel S. Wilks Medal "for outstanding research in large sample theory and sequential analysis, for extensive service to scholarly societies and on government panels, for effectiveness and popularity as a teacher, and for his continuing impact on the theory of statistics and its applications in diverse disciplines" (1987). He was named Statistician of the Year, Boston Chapter of the ASA (1991). He holds four honorary doctorates. Professor Chernoff is the co-author, with Lincoln Moses, of a classic text, now a Dover Reprint, entitled *Elementary Decision Theory*. He is also the author of the SIAM monograph 8 entitled *Sequential Analysis and Optimal Design*. The book *Recent Advances in Statistics* (MH Rizvi, J Rustagi and D Siegmund (Eds.), Academic Press, New York, 1983) published in honor of his 60th birthday in 1983 contained papers in the fields where his influence as a researcher and teacher has been strong: design and sequential analysis, optimization and control, nonparametrics, large sample theory and statistical graphics.

Cross References

►Asymptotic Normality
►Asymptotic Relative Efficiency in Testing
►Gaussian Processes
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test
►Wilcoxon-Signed-Rank Test

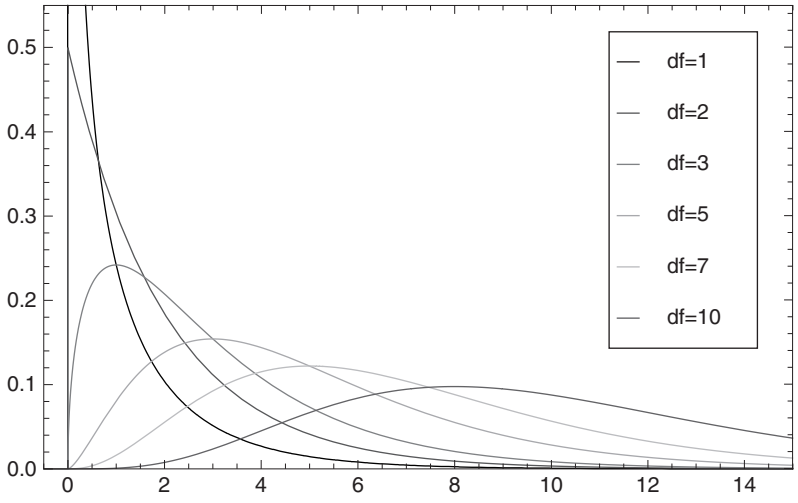
References and Further Reading

Chernoff H, Savage IR (1958) Asymptotic normality and efficiency of certain nonparametric test statistics. Ann Math Stat 29:972–994
Hodges JL, Lehmann EL (1956) The efficiency of some nonparametric competitors to the *t*-test. Ann Math Stat 27:321–325
Terry ME (1952) Some rank order tests which are most powerful against specific parametric alternatives. Ann Math Stat 23: 346–366

Chi-Square Distribution

Miljenko Huzak
University of Zagreb, Zagreb, Croatia

The chi-square distribution is one of the most important continuous probability distributions with many uses in statistical theory and inference. According to O. Sheynin (1971), Ernst Karl Abbe obtained it in 1863, Maxwell formulated it for three degrees of freedom in 1860, and Boltzman discovered the general expression in 1881. Lancaster (1966) ascertained that Bienaymé derived it as early as in 1838. However, their derivations "had no impact on the progress of the mainstream statistics" (R. L. Plackett 1983, p. 68)



Chi-Square Distribution. Fig. 1 Densities of χ^2 -distributions with 1, 2, 3, 5, 7, and 10 degrees of freedom (df)

since chi-square is not only a distribution, but also a statistic and a test procedure, all of which arrived simultaneously in the seminal paper written by Karl Pearson in 1900.

Let $n \geq 1$ be a positive integer. We say that a random variable (r.v.) has χ^2 (*chi-square*, χ is pronounced ki as in kind) *distribution with n degrees of freedom* (d.f.) if it is absolutely continuous with respect to the Lebesgue measure with density:

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \Gamma\left(\frac{n}{2}\right)^{-1} 2^{-n/2} x^{n/2-1} e^{-x/2} & \text{if } x > 0 \end{cases}$$

where Γ denotes the Gamma function.

Figure 1 shows some of the densities.

Hence, the χ^2-distribution (with n d.f.) is equal to the Γ - distribution with the parameters $(n/2, 2)$, that is, with the mean and variance equal to n and $2n$ respectively.

The χ^2-distribution is closely connected with the normal distribution. It turns out that the sample variance S^2 of a random sample from a normally distributed population has, up to the constant, the χ^2 – sample distribution. More precisely, if $X_1, \ldots, X_n$ are independent and identically distributed normal r.v.s with the population variance σ^2, then

$$\frac{n-1}{\sigma^2} \cdot S^2 = \frac{1}{\sigma^2}\left((X_1 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2\right) = \frac{1}{\sigma^2}\left(X_1^2 + X_2^2 + \cdots + X_n^2 - n\overline{X}^2\right)$$

is a χ^2-distributed r.v. with $n - 1$ d.f. (see e.g., Shorack 2000). This is a consequence of a more general property of the normality (Feller 1971). For example, let $\mathbf{X}$ be an n-dimensional standard normal vector, that is, a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ such that its components $X_1, \ldots, X_n$ are independent and normally distributed with mean and variance equal to 0 and 1 respectively. Then the square of the Euclidean norm of $\mathbf{X}$, $|\mathbf{X}|^2 = X_1^2 + \ldots + X_n^2$, is χ^2-distributed with n d.f. If means of the components of $\mathbf{X}$ are non-zero, then $|\mathbf{X}|^2$ has *non-central* χ^2-distribution with n d.f. and *non-centrality* parameter equal to the square of the mean of $\mathbf{X}$. In this generality, χ^2-distribution is the *central* χ^2-distribution, that is, a χ^2-distribution with non-centrality parameter equal to 0.

In statistics, many test statistics have a χ^2 or asymptotic χ^2-distribution. For example, goodness of fit χ^2-tests are based on the so-called Pearson's χ^2-statistics or general χ^2-statistics that have, under appropriate null-hypothesis, asymptotic χ^2-distributions; The Friedman test statistic and likelihood ratio tests are also based on asymptotically χ^2-distributed test statistic (see Ferguson 1996). Generally, appropriately normalized quadratic forms of normal (and asymptotic normal) statistics have χ^2 (and asymptotic χ^2) distributions.

Non-central χ^2-distributions are used for calculating the power function of tests based on quadratic forms of normal or asymptotic normal statistics.

Cross References

►Categorical Data Analysis
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Test: Analysis of Contingency Tables
►Chi-Square Tests
►Continuity Correction
►Gamma Distribution
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Tests for Homogeneity of Variance

References and Further Reading

Feller W (1971) An introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York

Ferguson TS (1996) A course in large sample theory. Chapman & Hall, London

Lancaster HO (1966) Forerunners of the Pearson χ^2. Aust J Stat 8: 117–126

Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. Philos Mag 5(1):157–175

Plackett RL (1983) Karl Pearson and the Chi-squared test. Int Stat Rev 51(1):59–72

Sheynin OB (1971) Studies in the history of probability and statistics. XXV. On the history of some statistical laws of distribution. Biometrika 58(1):234–236

Shorack GR (2000) Probability for statisticians, Springer-Verlag, New York

Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements

Vassiliy Voinov[1], Mikhail Nikulin[2]
[1]Professor
Kazakhstan Institute of Management, Economics and Strategic Research, Almaty, Kazakhstan
[2]Professor
University of Victor Segalen, Bordeaux, France

The famous chi-squared goodness-of-fit test was discovered by Karl Pearson in 1900. If the partition of a sample space is such that observations are grouped over r disjoined

intervals Δ_i, and denoting ν_i observed frequencies and $np_i(\theta)$ expected that correspond to a multinomial scheme, the Pearson's sum is written

$$\chi^2 = X_n^2(\theta) = \sum_{i=1}^{r} \frac{(\nu_i - np_i(\theta))^2}{np_i(\theta)} = \mathbf{V}^T(\theta)\mathbf{V}(\theta), \quad (1)$$

where $\mathbf{V}(\theta)$ is a vector with components $v_i(\theta)=(\nu_i - np_i(\theta))(np_i(\theta))^{-1/2}$, $i=1,\ldots,r$. If the number of observations $n \to \infty$, the statistic (1) for a simple null hypothesis, specifying the true value of θ, will follow chi-squared probability distribution with $r-1$ degrees of freedom.

Until 1934, Pearson believed that the limit distribution of his chi-squared statistic would be the same if unknown parameters of the null hypothesis were replaced by estimates based on a sample (Stigler (2008), p. 266). Stigler noted that this major error of Pearson "has left a positive and lasting impression upon the statistical world." It would be better to rephrase this sentence as follows: "has left a positive (because it inspired the further development of the theory of chi-squared test) and lasting 'negative' impression". Fisher (1924) clearly showed that the number of degrees of freedom of the Pearson's test must be reduced by the number of parameters estimated by a sample. The Fisher's result is true if and only if parameters are estimated by grouped data (minimizing Pearson's chi-squared sum, using multinomial maximum likelihood estimates (MLEs) for grouped data, or by any other asymptotically equivalent procedure).

Nowadays, the Pearson's test with unknown parameters replaced by grouped data estimates $\hat{\theta}_n$ is known as Pearson-Fisher test $X_n^2(\hat{\theta}_n)$. Chernoff and Lehmann (1954) showed that replacing unknown parameters in (1) by their maximum likelihood estimates based on non-grouped data would dramatically change the limit distribution. In this case, it will follow a distribution that in general depends on unknown parameters and, hence, cannot be used for testing. What is difficult to understand for those who apply chi-squared tests is that an estimate is a realization of a random variable with its own probability distribution and that a particular estimate can be too far from the actual unknown value of a parameter or parameters. This misunderstanding is rather typical for those who apply both parametric and non-parametric tests for compound hypotheses.

Roy (1956) extended Chernoff and Lehmann's result to the case of random grouping intervals. Molinari (1977) derived the limit distribution of Pearson's sum if moment type estimates (MMEs) based on raw data are used. Like the case of MLEs it depends on unknown parameters. Thus, a problem of deriving a test statistic, where limiting distribution will not depend on parameters, is aroused. Dahiya and Gurland (1972) showed that for location and scale families with properly chosen random cells, the limit distribution of Pearson's sum may not depend on unknown parameters but on the null hypothesis. Being distribution-free, such tests can be used in practice, but for each specific null distribution one has to evaluate corresponding critical values. So, two ways of constructing distribution-free Pearson's type tests are to use proper estimates of unknown parameters (e.g., based on grouped data), or to use specially constructed grouping intervals. Another possible way is to modify the Pearson's sum such that its limit probability distribution would not depend on unknowns. Nikulin (1973), using a very general theoretical approach (nowadays known as Wald's method (see Moore 1977)), solved the problem in full for any continuous probability distribution if one will use random cells based on predetermined probabilities to fall into a cell with random boundaries depending on efficient estimates (MLEs or best asymptotically normal (BAN) estimates) of unknown parameters. Rao and Robson (1974), using a much less general heuristic approach, confirmed the result of Nikulin for a particular case of exponential family of distributions. Formally their result fully coincides with that of Nikulin (1973)

$$Y1_n^2(\hat{\theta}_n) = X_n^2(\hat{\theta}_n) + \mathbf{V}^T(\hat{\theta}_n)\mathbf{B}(\mathbf{J} - \mathbf{J}_g)^{-1}\mathbf{B}^T\mathbf{V}(\hat{\theta}_n), \quad (2)$$

where $\mathbf{J}$ and $\mathbf{J}_g = \mathbf{B}^T\mathbf{B}$ are Fisher information matrices for non-grouped and grouped data correspondingly, and $\mathbf{B}$ is a matrix with elements $b_{ij} = \frac{1}{\sqrt{p_i(\theta)}}\frac{\partial p_i(\theta)}{\partial \theta_j}$, $i = 1,\ldots,r$, $j = 1,\ldots,s$. The statistic (2) can be presented also as (Moore and Spruill (1975))

$$Y1_n^2(\hat{\theta}_n) = \mathbf{V}^T(\hat{\theta}_n)(\mathbf{I} - \mathbf{B}\mathbf{J}^{-1}\mathbf{B}^T)^{-1}\mathbf{V}(\hat{\theta}_n). \quad (3)$$

The statistic (2) or (3), suggested first by Nikulin (1973a) for testing the normality, will be referred to subsequently as Nikulin-Rao-Robson (NRR) test. Nikulin (1973) assumed that only asymptotically efficient estimates of unknown parameters (e.g., MLEs based on non-grouped data or BAN estimates) are used for testing. Singh (1987), Spruill (1976), and Lemeshko et al. (2001) showed that the NRR test is asymptotically optimal in some sense. This optimality is not surprising because the second term of (2) depends on the difference between Fisher's matrices for grouped and non-grouped data that possibly takes the information lost in full (Voinov (2006)). Dzhaparidze and Nikulin (1992) generalized Fisher's idea to improve any $\sqrt{n}$- consistent estimator to make it asymptotically as efficient as

MLE. This gives the following way of chi-squared test modification: improve an estimator first and then use the NRR statistic. Since this way is not simple computationally, it is worth considering other modifications. At this point it is important to note that the NRR test is very suitable for describing censored data (Habib and Thomas (1986)).

Dzhaparidze and Nikulin (1974) proposed a modification of the standard Pearson's statistic valid for any square root of n consistent estimate $\tilde{\theta}_n$ of an unknown parameter $U_n^2(\tilde{\theta}_n) = \mathbf{V}^T(\tilde{\theta}_n)\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}(\tilde{\theta}_n)$. This test (the DN test), like the asymptotically equivalent Pearson-Fisher one, is not powerful for equiprobable cells (McCulloch (1985), Voinov et al. (2009)) but it can be rather powerful if an alternative hypothesis is specified and one uses the Neyman-Pearson classes for data grouping. Having generalized the idea of Dzhaparidze and Nikulin (1974), Singh (1987) suggested a generalization of the RRN test (3) valid for any $\sqrt{n}$ - consistent estimator $\tilde{\theta}_n$ of an unknown parameter $Q_s^2(\tilde{\theta}_n) = \mathbf{V}_*^T(\tilde{\theta}_n)(\mathbf{I} - \mathbf{B}\mathbf{J}^{-1}\mathbf{B}^T)^{-1}\mathbf{V}_*(\tilde{\theta}_n)$, where $\mathbf{V}_*(\tilde{\theta}_n) = \mathbf{V}(\tilde{\theta}_n) - \mathbf{B}\mathbf{J}^{-1}\mathbf{W}(\tilde{\theta}_n)$, and $\mathbf{W}(\tilde{\theta}_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \ln f(X_i,\theta)}{\partial\theta}|_{\theta=\tilde{\theta}_n}$.

A unified large-sample theory of general chi-squared statistics for tests of fit was developed by Moore and Spruill (1975). Moore (1977), based upon Wald's approach, formulated a general recipe for constructing modified chi-squared tests for any square root of n consistent estimator that actually is a generalization of Nikulin's idea. He was first to show that a resulting Wald's quadratic form does not depend on the way of limit covariance matrix of generalized frequencies inverting.

Hsuan and Robson (1976) showed that a modified statistic will not be the same as (3) in the case of moment type estimates (MMEs) of unknown parameters. They succeeded in deriving the limit covariance matrix for generalized frequencies and proved the theorem that a corresponding Wald's quadratic form will follow in the limit the chi-squared distribution. Hsuan and Robson provided the test statistic explicitly for the exponential family of distributions, when MMEs coincide with MLEs, thus confirming the already known result of Nikulin (1973). Hsuan and Robson have not derived the general modified test based on MMEs $\bar{\theta}_n$ explicitly. This was done later by Mirvaliev (2001). Taking into account the input of Hsuan and Robson, and Mirvaliev, this test will be referred to subsequently as the Hsuan-Robson-Mirvaliev (HRM) statistic

$$Y2_n^2(\bar{\theta}_n) = X_n^2(\bar{\theta}_n) + R_n^2(\bar{\theta}_n) - Q_n^2(\bar{\theta}_n). \qquad (4)$$

Explicit expressions for quadratic forms $R_n^2(\bar{\theta}_n)$ and $Q_n^2(\bar{\theta}_n)$ are given, e.g., in Voinov et al. (2009). The approach, based on Wald's transformation, was also used by Bol'shev and Mirvaliev (1978), Nikulin and Voinov (1989), Voinov and Nikulin (1994), and by Chichagov (2006) for minimum variance unbiased estimators (MVUEs).

It is important to mention two types of decompositions of classical and modified chi-squared tests. The first way decomposes a modified test on a sum of the classical Pearson's test and a correcting term that makes the test chi-squared distributed being distribution free in the limit (Nikulin (1973)). A much more important decomposition was first suggested by McCulloch (1985) (see also Mirvaliev (2001)). This is a decomposition of a test on a sum of the DN statistic and an additional quadratic form being asymptotically independent on the DN statistic. Denoting $W_n^2(\theta) = \mathbf{V}^T(\theta)\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}(\theta)$ and $P_n^2(\theta) = \mathbf{V}^T(\theta)\mathbf{B}(\mathbf{J} - \mathbf{J}_g)^{-1}\mathbf{B}^T\mathbf{V}(\theta)$ the decomposition of the NRR statistic (2) in case of MLEs will be $Y1_n^2(\hat{\theta}_n) = U_n^2(\hat{\theta}_n) + \left(W_n^2(\hat{\theta}) + P_n^2(\hat{\theta}_n)\right)$, where $U_n^2(\hat{\theta}_n)$ is asymptotically independent on $\left(W_n^2(\hat{\theta}) + P_n^2(\hat{\theta}_n)\right)$, and on $W_n^2(\hat{\theta})$. The decomposition of the HRM statistic (4) is $Y2_n^2(\bar{\theta}_n) = U_n^2(\bar{\theta}_n) + \left(W_n^2(\bar{\theta}) + R_n^2(\bar{\theta}_n) - Q_n^2(\bar{\theta}_n)\right)$, where $U_n^2(\bar{\theta}_n)$ is asymptotically independent on $\left(W_n^2(\bar{\theta}) + R_n^2(\bar{\theta}_n) - Q_n^2(\bar{\theta}_n)\right)$, but is asymptotically correlated with $W_n^2(\bar{\theta})$.

The decomposition of a modified chi-squared test on a sum of the DN statistic and an additional term is of importance because the DN test based on non-grouped data is asymptotically equivalent to the Pearson-Fisher's (PF) statistic for grouped data. Hence, that additional term takes into account the Fisher's information lost due to grouping. Later it was shown (Voinov et al. (2009)) that the DN part, like the PF test, is (for equiprobable cells, for example) insensitive to some alternative hypothesis in case of equiprobable cells (fixed or random) and would be sensitive to it for, e.g., non-equiprobable two Neyman-Pearson classes. For equiprobable cells this suggests using the difference between the modified statistic and the DN part that will be the most powerful statistic in case of equiprobable cells (McCulloch (1985), Voinov et al. (2009)). It became clear that the way of sample space partitioning essentially influences power of a test.

Ronald Fisher (1925) was the first to note that "in some cases it is possible to separate the contributions to χ^2 made by the individual degrees of freedom, and so to test the separate components of a discrepancy." Cochran (1954) wrote "that the usual χ^2 tests are often insensitive, and do not indicate significant results when the null hypothesis is actually false" and suggested to "use a single degree of freedom, or a group of degrees of freedom, from the total χ^2," to obtain more powerful and appropriate test. The

problem of implementing the idea of Fisher and Cochran was that decompositions of Pearson's sum and modified test statistics were not known at that time. Anderson (1994) (see also Boero et al. (2004)) was possibly the first who to decompose the Pearson's χ^2 for a simple null hypothesis into a sum of independent χ^2_1 random variables in case of equiprobable grouping cells. A parametric decomposition of Pearson's χ^2 in case of non-equiprobable cells based on ideas of Mirvaliev (2001) was obtained by Voinov et al. (2008) in an explicit form. At the same time Voinov et al. (2008) presented parametric decompositions of NRR and HRM statistics. Voinov (2010) and Voinov and Pya (2010) introduced vector-valued goodness-of-fit tests that, in some cases, can provide a gain in power for specified alternatives.

About the Authors

Vassiliy Voinov is a Professor of the Operations Management and Information Systems Department, Kazakhstan Institute of Management, Economics and Strategic Research (KIMEP). He received his engineering diploma at Tomsk State University; Candidate of Science degree in Kazakh Academy of Science; Doctor of Science degree in Joint Institute for Nuclear Research, Dubna, Moscow region; Professor degree in Kazakh State Polytechnic University, and also in 1998 received a Professor's degree in KIMEP. He has professional experience as an invited professor in statistics at the University Victor Segalen Bordeaux, France. Vassiliy Voinov participated in many international conferences and has more than 120 research papers and books, including: *Unbiased Estimators and Their Applications*, Volume 1: *Univariate Case*, and Volume 2: *Multivariate Case* (with M. Nikulin, Kluwer Academic Publishers: Dordrecht, 1993 and 1996).

Mikhail S. Nikulin (Nikouline) is Professor of Statistics at the University Victor Segalen, Bordeaux 2. He earned his doctorate in the Theory of Probability and Mathematical Statistics from the Steklov Mathematical Institute in Moscow in 1973, under supervision of Professor L. N. Bol'shev. He was Dean of the the Faculty l'UFR "Sciences and Modélisation", the University Victor Segalen (1996–2001) and Head of the Laboratory EA 2961 "Mathematical Statistics and its Applications" (1999–2007). Professor Nikulin has (co-)authored over 250 papers, 28 edited volumes and 13 books, including *A Guide to Chi-Squared Testing* (with P.E. Greenwood, Wiley, 2004), *Probability, Statistics and Modelling in Public Health* (with D. Commenges, Springer, 2005), and was the editor of the volume *Advances in Degradation Modeling: Applications to Reliability, Survival Analysis and Finance* (with N. Balakrishnan, W. Kahle, N. Limnios and C. Huber-Carol, Birkhäuser, 2009). His name is associated with several statistics terms: Dzhaparidze-Nikulin statistic, Rao-Robson-Nikulin statistic, Bagdonavičius-Nikulin model, and Bagdonavičius-Nikulin estimator.

Cross References

►Chi-Square Distribution
►Chi-Square Test: Analysis of Contingency Tables
►Chi-Square Tests

References and Further Reading

Anderson G (1994) Simple tests of distributional form. J Economet 62:265–276

Boero G, Smith J, Wallis KF (2004) The sensitivity of chi-squared goodness of-fit tests to the partitioning of data. Economet Rev 23:341–370

Bol'shev LN, Mirvaliev M (1978) Chi-square goodness-of-fit test for the Poisson, binomial, and negative binomial distributions. Theory Probab Appl 23:481–494 (in Russian)

Chernoff H, Lehmann EL (1954) The use of maximum likelihood estimates in tests for goodness of fit. Ann Math Stat 25: 579–589

Chichagov VV (2006) Unbiased estimates and chi-squared statistic for one-parameter exponential family. In: Statistical methods of estimation and hypotheses testing, vol 19. Perm State University, Perm, Russia, pp 78–89

Cohran G (1954) Some methods for strengthening the common χ^2 tests. Biometrics 10:417–451

Dahiya RC, Gurland J (1972) Pearson chi-squared test of fit with random intervals. Biometrica 59:147–153

Dzhaparidze KO, Nikulin MS (1974) On a modification of the standard statistic of Pearson. Theory Probab Appl 19: 851–853

Dzhaparidze KO, Nikulin MS (1992) On evaluation of statistics of chi-square type tests. In: Problem of the theory of probability distributions, vol 12. Nauka, St. Petersburg, pp. 59–90

Fisher RA (1924) The condition under which χ^2 measures the discrepancy between observation and hypothesis. J R Stat Soc 87:442–450

Fisher RA (1925) Partition of χ^2 into its components. In: Statistical methods for research workers. Oliver and Boyd, Edinburg

Habib MG, Thomas DR (1986) Chi-square goodness-of-fit tests for randomly censored data. Ann Stat 14:759–765

Hsuan TA, Robson DS (1976) The goodness-of-fit tests with moment type estimators. Commun Stat Theory Meth A5:1509–1519

Lemeshko BYu, Postovalov SN, Chimitiva EV (2001) On the distribution and power of Nikulin's chi-squared test. Ind Lab 67:52–58 (in Russian)

McCulloch CE (1985) Relationships among some chi-squared goodness of fit statistics. Commun Stat Theory Meth 14:593–603

Mirvaliev M (2001) An investigation of generalized chi-squared type statistics. Academy of Sciences of the Republic of Uzbekistan, Tashkent, Doctoral dissertation

Molinari L (1977) Distribution of the chi-squared test in non-standard situations. Biometrika 64:115–121

Moore DS (1977) Generalized inverses, Wald's method and the construction of chisquared tests of fit. J Am Stat Assoc 72: 131–137

Moore DS, Spruill MC (1975) Unified large-sample theory of general chisquared statistics for tests of fit. Ann Stat 3:599–616
Nikulin MS (1973a) Chi-square test for continuous distributions. Theory Probab Appl 18:638–639
Nikulin MS (1973b) Chi-square test for continuous distributions with shift and scale parameters. Theory Probab Appl 18: 559–568
Nikulin MS, Voinov VG (1989) A chi-square goodness-of-fit test for exponential distributions of the first order. Springer-Verlag Lect Notes Math 1412:239–258
Rao KC, Robson DS (1974) A chi-squared statistic for goodness-of-fit tests within the exponential family. Commun Stat 3: 1139–1153
Roy AR (1956) On χ^2 statistics with variable intervals. Technical report N1, Stanford University, Statistics Department
Singh AC (1987) On the optimality and a generalization of Rao–Robson's statistic. Commun Stat Theory Meth 16, 3255–3273
Spruill MC (1976) A comparison of chi-square goodness-of-fit tests based on approximate Bahadur slope. Ann Stat 2:237–284
Stigler SM (2008) Karl Pearson's theoretical errors and the advances they inspired. Stat Sci 23:261–171
Voinov V (2006) On optimality of the Rao–Robson–Nikulin test. Ind Lab 72:65–70
Voinov V (2010) A decomposition of Pearson–Fisher and Dzhaparidze–Nikulin statistics and some ideas for a more powerful test construction. Commun Stat Theory Meth 39(4):667–677
Voinov V, Pya N (2010) A note on vector-valued goodness-of-fit tests. Commun Stat 39(3):452–459
Voinov V, Nikulin MS, Pya N (2008) Independently distributed in the limit components of some chi-squared tests. In: Skiadas CH (ed) Recent advances in stochastic modelling and data analysis. World Scientific, New Jersey
Voinov V, Pya N, Alloyarova R (2009) A comparative study of some modified chi-squared tests. Commun Stat Simulat Comput 38:355–367

Chi-Square Test: Analysis of Contingency Tables

David C. Howell
Professor Emeritus
University of Vermont, Burlington, VT, USA

The term "chi-square" refers both to a statistical distribution and to a hypothesis testing procedure that produces a statistic that is approximately distributed as the ▶chi-square distribution. In this entry the term is used in its second sense.

Pearson's Chi-Square

The original chi-square test, often known as Pearson's chi-square, dates from papers by Karl Pearson in the earlier 1900s. The test serves both as a "goodness-of-fit" test, where the data are categorized along one dimension, and as a test for the more common "contingency table," in which categorization is across two or more dimensions. Voinov and Nikulin, this volume, discuss the controversy over the correct form for the goodness of fit test. This entry will focus on the lack of agreement about tests on contingency tables.

In 2000 the Vermont State legislature approved a bill authorizing civil unions. The vote can be broken down by gender to produce the following table, with the expected frequencies given in parentheses. The expected frequencies are computed as $R_i \times C_j/N$, where R_i and C_j represent row and column marginal totals and N is the grand total.

	Vote		
	Yes	No	Total
Women	35 (28.83)	9 (15.17)	44
Men	60 (66.17)	41 (34.83)	101
Total	95	50	145

The standard Pearson chi-square statistic is defined as

$$\chi^2 = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(35 - 28.83)^2}{28.83} + \cdots + \frac{(41 - 34.83)^2}{34.83} = 5.50$$

where i and j index the rows and columns of the table. (For the goodness-of-fit test we simply drop the subscript j.) The resulting test statistic from the formula on the left is approximately distributed as χ^2 on $(r-1)(c-1)$ degrees of freedom. The probability of $\chi^2 \geq 5.50$ on 1 df = 0.019, so we can reject the null hypothesis that voting behavior is independent of gender. (Pearson originally misidentified the degrees of freedom, Fisher corrected him, though Pearson long refused to recognize the error, and Pearson and Fisher were enemies for the rest of their lives.)

Likelihood Ratio Chi-Square

Pearson's chi-square statistic is not the only chi-square test that we have. The likelihood ratio chi-square builds on the likelihood of the data under the null hypothesis relative to the maximum likelihood. It is defined as

$$G^2 = 2\sum O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right) = 2\left[35\ln\left(\frac{35}{28.83}\right) + 9\ln\left(\frac{9}{15.17}\right) + 60\ln\left(\frac{60}{66.17}\right) + 41\ln\left(\frac{41}{34.83}\right)\right] = 5.81$$

This result is slightly larger than the Pearson chi-square of 5.50. One advantage of the likelihood ratio chi-square is that G^2 for a large dimensional table can be neatly decomposed into smaller components. This cannot be done exactly with Pearson's chi-square, and G^2 is the usual statistic for log-linear analyses. As sample sizes increase the two chi-square statistics converge.

Small Expected Frequencies

Probably no one would object to the use of the Pearson or likelihood ratio chi-square tests for our example. However, the chi-square statistic is only approximated by the chi-square distribution, and that approximation worsens with small expected frequencies. When we have very small expected frequencies, the possible values of the chi-square statistic are quite discrete. For example, for a table with only four observations in each row and column, the only possible values of chi-square are 8, 2, and 0. It should be clear that a continuous chi-square distribution is not a good match for a discrete distribution having only three values. The general rule is that the smallest expected frequency should be at least five. However Cochran (1952), who is generally considered the source of this rule, acknowledged that the number "5" seems to be chosen arbitrarily.

Yates proposed a correction to the formula for chi-square to bring it more in line with the true probability. However, given modern computing alternatives, Yates' correction is much less necessary and should be replaced by more exact methods.

For situations in which we do not satisfy Cochran's rule about small expected frequencies, the obvious question concerns what we should do instead. This is an issue over which there has been considerable debate. One of the most common alternatives is Fisher's Exact Test (see below), but even that is controversial for many designs.

Alternative Research Designs

There are at least four different research designs that will lead to data forming a contingency table. One design assumes that all marginal totals are fixed. Fisher's famous "tea-tasting" study had four cups of tea with milk added first and four with milk added second (row totals are fixed). The taster had to assign four cups to each guessed order of pouring, fixing the column totals. The underlying probability model is hypergeometric, and Fisher's exact test (1934) is ideally suited to this design and gives an exact probability. This test is reported by most software for 2×2 tables, though it is not restricted to the 2×2 case.

Alternatively we could fix only one set of marginals, as in our earlier example. Every replication of that experiment would include 44 women and 101 men, although the vote totals could vary. This design is exactly equivalent to comparing the proportion of "yes" votes for men and women, and it is based on the ▶binomial distribution. The square of a z-test on proportions would be exactly equal to the resulting chi-square statistic. One alternative analysis for this design would be to generate all possible tables with those row marginals and compute the percentage of obtained chi-square statistics that are as extreme as the statistic obtained from the actual data. Alternatively, some authorities recommend the use of a mid-p value, which sums the probability of all tables less likely than the one we obtained, plus half of the probability of the table we actually obtained.

For a different design, suppose that we had asked 145 Vermont citizens to record their opinion on civil unions. In this case neither the Gender nor Vote totals would be fixed, only the total sample size. The underlying probability model would be multinomial. Pearson's chi-square test would be appropriate, but a more exact test would be obtained by taking all possible tables (or, more likely, a very large number of randomly generated tables) with 145 observations and calculating chi-square for each. Again the probability value would be the proportion of tables with more extreme outcomes than the actual table. And, again, we could compute a mid-p probability instead.

Finally, suppose that we went into college classrooms and asked the students to vote. In this case not even the total sample size is fixed. The underlying probability model here is Poisson.

Computer scripts written in R are available for each model with a fixed total sample size at http://www.uvm.edu/~dhowell/StatPages/chi-square-alternatives.html

Summary

Based on a large number of studies of the analysis of contingency tables, the current recommendation would be to continue to use the standard Pearson chi-square test whenever the expected cell frequencies are sufficiently large. There seems to be no problem defining large as "at least 5." With small expected frequencies ▶Fisher's Exact Test seems to perform well regardless of the sampling plan, but ▶randomization tests adapted for the actual research design, as described above, will give a somewhat more exact solution. Recently Campbell (2007) carried out a very large sampling study on 2×2 tables comparing different chi-square statistics under different sample sizes and different underlying designs. He found that across all sampling designs, a statistic suggested by Karl Pearson's son Egon Pearson worked best in most situations. The statistic is defined as $\chi^2 \frac{N}{N-1}$. (For the justification for that adjustment see Campbell's paper.) Campbell found that as

long as the smallest expected frequency was at least one, the adjusted chi-square held the Type I error rate at very nearly α. When the smallest expected frequency fell below 1, Fisher's Exact Test did best.

About the Author

David Howell is a Professor Emeritus (since 2002), and former chair of the Psychology department at the University of Vermont (1987–1992) and (2000–2002). Professor Howell's primary area of research is in statistics and experimental methods. He has authored well known texts: *Statistical Methods for Psychology* (Wadsworth Publishing, 7*th* ed., 2009), *Fundamental Statistics for Behavioral Sciences* (Wadsworth Publishing, 7*th* ed., 2010), and is a coauthor (with Brian Everitt) of a four volume *Encyclopedia of Statistics in Behavior Science* (Wiley & Sons, 2005).

Cross References

►Chi-Square Distribution
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Tests

References and Further Reading

Campbell I (2007) Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. Stat Med 26:3661–3675

Cochran WG (1952) The χ^2 test of goodness of fit. Ann Math Stat 25:315–345

Fisher RA (1934) The logic of inductive inference. J R Stat Soc 98: 39–54

Chi-Square Tests

Karl L. Wuensch
Professor
East Carolina University, Greenville, NC, USA

The χ^2 statistic was developed by Karl Pearson (1900) as a means to compare an obtained distribution of scores with a theoretical distribution of scores. While it is sometimes still employed as a univariate goodness of fit test, other statistics, such as the ►Kolmogorov–Smirnov test and, where the theoretical distribution is normal, the Shapiro–Wilk test, are now more often used for that purpose.

The chi-square statistic on n degrees of freedom is defined as

$$\chi_n^2 = \sum_{i=1}^{n} z_i^2 = \sum \frac{(Y-\mu)^2}{\sigma^2},$$

where z_i is normally distributed with mean zero and standard deviation one (Winkler and Hayes 1975, pp. 375–380). If one were repeatedly to draw samples of one Y score from a normally distributed population, transform that score to a standard z score, and then square that z score, the resulting distribution of squared z scores would be a χ^2 distribution on one degree of freedom. If one were repeatedly to draw samples of three scores, standardize, square, and sum them, the resulting distribution would be χ^2 on three degrees of freedom. Because the χ^2 statistic is so closely related to the normal distribution, it is also closely related to other statistics that are related to the normal distribution, such as t and F.

One simple application of the χ^2 statistic is to test the null hypothesis that the variance of a population has a specified value (Winkler and Hayes 1975, pp. 453–455; Wuensch 2009). From the definition of the sample variance, $s^2 = \frac{\sum(Y-M)^2}{N-1}$, where Y is a score, M is the sample mean, and N is the sample size, the corrected sum of squares $\sum(Y-M)^2 = (N-1)s^2$. Substituting this expression for $\sum(Y-\mu)^2$ in the defining formula yields $\chi^2 = \frac{(N-1)s^2}{\sigma^2}$. To test the hypothesis that an observed sample came from a population with a particular variance, one simply divides the sample sum of squares, $(N-1)s^2$, by the hypothesized variance. The resulting χ^2 is evaluated on $N-1$ degrees of freedom, with a two-tailed p value for nondirectional hypotheses and a one-tailed p for directional hypotheses.

One can also compute a confidence interval for the population variance (Winkler and Hayes 1975, pp. 383–385; Wuensch 2009). For a $100(1-\alpha)\%$ confidence interval for the population variance, compute:

$$\frac{(N-1)s^2}{b} \quad \text{and} \quad \frac{(N-1)s^2}{a}$$

where a and b are the $\alpha/2$ and $(1-\alpha/2)$ fractiles of the chi square distribution on $(N-1)df$. It should be noted that these procedures are not very robust to their assumption that the population is normally distributed.

When one states that he or she has conducted a "chi-square test," that test is most often a "one-way chi-square test" or a "two-way chi-square test" (Howell 2010, pp. 141–151). The one-way test is a univariate goodness of fit test. For each of k groups one has an observed frequency (O) and a theoretical frequency (E), the latter being derived from the theoretical model being tested. The test

statistic is $\chi^2 = \sum \frac{(O-E)^2}{E}$ on $k-1$ degrees of freedom. The appropriate p value is one-tailed, upper-tailed, for nondirectional hypotheses. When $k = 2$, one should make a "correction for continuity":

$$\chi^2 = \sum \frac{(|O-E| - .5)^2}{E}.$$

The two-way chi-square test is employed to test the null hypothesis that two categorical variables are independent of one another. The data may be represented as an $r \times c$ contingency table, where r is the number of rows (levels of one categorical variable) and c is the number of columns (levels of the other categorical variable). For each cell in this table two frequencies are obtained, the observed frequency (O) and the expected frequency (E). The expected frequencies are those which would be expected given the marginal frequencies if the row variable and the column variable were independent of each other. These expected frequencies are easily calculated from the multiplication rule of probability under the assumption of independence. For each cell, the expected frequency is $(R_i C_j / N)$, where R_i is the marginal total for all cells in the same row, C_j is the marginal total for all cells in the same column, and N is the total sample size. The χ^2 is computed exactly as with the one-way chi-square and is evaluated on $(r-1)(c-1)$ degrees of freedom, with an upper-tailed p value for nondirectional hypotheses. Although statistical software often provides a χ^2 with a correction for continuity when there are only two rows and two columns, almost always the uncorrected χ^2 is more appropriate (Camilli and Hopkins 1978).

It is not unusual to see the two-way chi-square inappropriately employed (Howell 2010, pp. 152–153). Most often this is a result of having counted some observations more than once or having not counted some observations at all. Each case should be counted once and only one. Statistical software will often provide a warning if one or more of the cells has a low expected frequency. The primary consequence of low expected frequencies is low power. Even with quite small expected frequencies, actual Type I error rates do not deviate much from the nominal level of alpha (Camilli and Hopkins 1978).

The results of a two-way chi-square test are commonly accompanied by an estimate of the magnitude of the association between the two categorical variables. When the contingency table is 2 × 2, an odds ratio and/or the phi coefficient (Pearson r between the two dichotomous variables) may be useful. With larger contingency tables Cramer's phi statistic may be useful.

The chi-square statistic is also employed in many other statistical procedures, only a few of which will be mentioned here. The Cochran-Mantel-Haenszel χ^2 is employed to test the hypothesis that there is no relationship between rows and columns when you average across two or more levels of a third variable. The Breslow-Day χ^2 is employed to test the hypothesis that the odds ratios do not differ across levels of a third variable. Likelihood ratio chi-square is employed in the log-linear analysis of multidimensional contingency tables, where it can be employed to test the difference between two models, where one is nested within the other. Likewise, in ►logistic regression, chi-square can be employed to test the effect of removing one or more of the predictors from the model. In discriminant function analysis, chi-square may be employed to approximate the p value associated with the obtained value of Wilks' lambda. A chi-square statistic can be employed to test the null hypothesis that k Pearson correlation coefficients are identical. Chi-square is also used to approximate the p value in the Kruskal-Wallis ANOVA and the Friedman ANOVA. Many more uses of the chi-square statistic could be cited.

About the Author

Dr. Karl L. Wuensch is a Professor in the Department or Psychology, East Carolina University, Greenville, NC, U.S.A. He has authored or coauthored 77 scholarly articles and chapters in books. Professor Wuensch has received several teaching awards, including the Board of Governors Award for Excellence in Teaching, the most prestigious teaching award in the University of North Carolina system.

Cross References

►Chi-Square Distribution
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Test: Analysis of Contingency Tables

References and Further Reading

Camilli G, Hopkins KD (1978) Applicability of chi-square to 2 × 2 contingency tables with small expected cell frequencies. Psychol Bull 85:163–167

Howell DC (2010) Statistical methods for psychology, 7th edn. Cengage Wadsworth, Belmont

Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh and Dublin Phil Mag J Sci Ser 5 50:157–175. Retrieved from www.economics.soton.ac.uk/staff/aldrich/1900.pdf

Winkler RL, Hays WL (1975) Statistics: probability, inference, and decision, 2nd edn. Holt Rinehart & Winston, New York

Wuensch KL (2009) Common univariate and bivariate applications of the chi-square distribution. Retrieved from http://core.ecu.edu/psyc/wuenschk/docs30/Chi-square.doc

Clinical Trials, History of

Toshimitsu Hamasaki
Associate Professor
Osaka University Graduate School of Medicine, Osaka, Japan

In recent years, in addition to advances in methodology, the number of clinical trials conducted and published has greatly increased. Clinical trials, in particular, *blinded, randomized, controlled* comparative clinical trials, are widely recognized as the most scientific and reliable method for evaluating the effectiveness of therapies and promoting a culture of evidence-based medicine (Tukey 1977; Byar et al. 1976; Zelen 1979; Cowan 1981; Byar 1991; Royall 1991; Smith 1998).

The first modern clinical trial is generally considered to be the treatment of pulmonary tuberculosis with streptomycin conducted by the UK Medical Research Council (MRC) and published in *British Medical Journal* in 1948 (MRC 1948; Pocock 1984; Ederer 2005; Day 2006). However, there is still some controversy surrounding this claim as some authors refer to the study with the common cold vaccine conducted by Diehl et al. (1938) as the first modern trial (Hart 1972, 1996; Gill 1996). The design of the streptomycin trial included blinding, ►randomization, and control groups as fundamental elements of the clinical trial. The trial included a total of 107 patients from seven centers, who were assigned to either "streptomycin and bed-rest" (S case) or "bed-rest" (C case) groups, by a process involving a statistical series based on random sampling numbers drawn up for each sex and each center and sealed envelopes. The efficacy of streptomycin was evaluated based upon the examination of patient X-ray films by three experts consisting of one clinician and two radiologists. The decision of whether or not the treatment was effective was made by the majority based on independently reached conclusions by each expert, who were also blinded as to which treatment the patient had received. The streptomycin trial also included Sir Austin Bradford Hill who served as the trial statistician. Hill was recognized as the world's leading medical statistician and popularized the use of statistical methods in clinical trials, and who also attempted to improve the quality of their implementation and evaluation by publishing a series of 17 articles in *The Lancet* in 1937 (Hill et al. 2000).

With the success of the streptomycin trial, the MRC and Hill continued with further blinded, randomized, controlled comparative clinical trials (Ederer 2005; Days 2006): for example, chemotherapy of pulmonary tuberculosis in young adults (MRC 1952), an antihistaminic drug in the prevention and treatment of the common cold (MRC 1950), the use of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis (MRC 1954, 1955), and an anticoagulant to treat cerebrovascular disease (Hill et al. 1960). In United States, the first randomized controlled trail started in 1951 and was the US National Institute of Health study of the adrenocorticotropic hormone, cortisone and aspirin in the treatment of rheumatic heart disease in children (Rheumatics Fever Working Party 1960). Presently, a huge number of randomized controlled clinical trials are being conducted worldwide, with the number of clinical trials steadily increasing each year.

Although now commonplace, the fundamental elements of clinical trials, such as blinding, randomization, and control groups, did not just suddenly appear in the second quarter of the twentieth century. Evidence exists that a comparative concept for evaluating therapeutic efficacy with control groups has been known since ancient times (Ederer 2005; Day 2006). For example, Lilienfeld (1949) and Slotki (1951) cited the description of a nutritional experiment using a control group in the Book of Daniel from the Old Testament:

> **1.6**: Among these were some from Judah: Daniel, Hananiah, Mishael and Azariah. . . **1.8**: But Daniel resolved not to defile himself with the royal food and wine, and he asked the chief official for permission not to defile himself this way. . . . **1.11**: Daniel then said to the guard whom the chief official had appointed over Daniel, Hananiah, Mishael and Azariah. **1.12**: Please test your servants for ten days; Give us nothing but vegetables to eat and water to drink. **1.13**: Then compare our appearance with that of the young men who eat the royal food, and treat your servants in accordance with what you see. **1.14**: So he agreed to this and tested them for ten days. **1.15**: At the end of the ten days they looked healthier and better nourished than any of the young men who ate the royal food. **1.16** So the guard took away their choice food and the wine they were to drink and gave them vegetables instead.

The above description is part of a story dating from approximately 800 BC when Daniel was taken captive by the ruler of Babylonia, Nebuchadnezzar. In order to refrain from eating royal meals containing meat (perhaps pork) and wine offered by Nebuchadnezzar, Daniel proposed a comparative evaluation and was rewarded when his test group fared better than the royal food group. Although it is difficult to confirm the accuracy of the account, it is clear that the comparative concept already existed when the Book of Daniel was written around 150 BC. In particular, it is

remarkable that the passage from the Book of Daniel mentioned not only the *choice of a control group* but the *use of a concurrent control group*. Unfortunately, this fundamental concept was not widely practiced until the latter half of the twentieth century (Ederer 2005; Day 2006).

Much later than the Book of Daniel, in the eighteenth and nineteenth centuries, there were some epoch-making clinical researches that formed the basis of the methodology used in current clinical trials. Before the modern clinical trial of the treatment of pulmonary tuberculosis with streptomycin mentioned above (Pocock 1984; Ederer 2005; Day 2006), the most famous historical example of a planned, controlled clinical trial involved six dietary treatments for scurvy on board a British ship. The trial was conducted by the ship's surgeon, James Lind, who was appalled by the ravages of scurvy which had claimed the lives of three quarters of the crew during the circumnavigation of the world by British admiral, George Anson (Lind 1753; Bull 1959; Pocock 1984; Mosteller 1981; Ederer 2005; Day 2006). In 1947, Lind conducted a comparative trial to establish the most promising "cure" for patients with scurvy using twelve individuals who had very similar symptoms on board the Salisbury. In addition to one common dietary supplement given to all of the patients, he assigned each of six pairs one of the following six dietary supplements:

1. Six spoonfuls of vinegar
2. A half-pint of sea water
3. A quart of cider
4. Seventy-five drops of vitriol elixir
5. Two oranges and one lemon
6. Nutmeg

Those patients who received the two oranges and one lemon were cured within approximately 6 days and were able to help nurse the other patients. Apart from the patients who improved somewhat after receiving the cider, Lind observed that the other remedies were ineffective. The reason for the success of Lind's trial was likely due to his knowledge of previous work by James Lancaster (Purchas 1625), who had served three teaspoons of lemon juice each day to sailors suffering from scurvy during the first expedition to India sent by the East India Company in 1601 (Mosteller 1981). Unfortunately, however, the British Navy did not supply lemon juice to its sailors until 1975, although conclusive results concerning the efficacy of such treatment had already been obtained much earlier (Bull 1959; Mosteller 1981).

The use of statistical concepts in clinical trials was also advocated earlier than the streptomycin trials. For example, Pierre Simon Laplace, a French mathematician and astronomer, mentioned the use of probability theory to determine the best treatment for the cure of a disease (Laplace 1814; Hill et al. 2000). Also, Pierre-Charles-Alexandre Louis, a French physician and pathologist, discussed the use of a "numerical method" for the assessment of treatments by constructing comparable groups of patients with similar degrees of a disease, i.e., to compare "like with like" (Louis 1837; Ederer 2005; Day 2006). Unfortunately, these suggestions were not earnestly acted upon until the streptomycin trial because in the eighteenth and nineteenth centuries, the investigators were more involved with the practice of medicine and less versed in the use of probability theory since saving patients' life was considered more important rather than collecting data from the aspect of ethics (Bull 1959; Hill et al. 2000).

Here, the history and development of clinical trials was very briefly traced. More detailed aspects of the history of clinical trials can be found in articles by Bull (1959), Armitage (1972, 1991), Lilienfeld (1982), Pocock (1984), Meinert (1986), Gail (1996), Ederer (2005) and Day (2006).

About the Author

Toshimitsu Hamasaki is Associate Professor of Department of Biomedical Statistics, Osaka University Graduate School of Medicine. He is the Elected member of International Statistical Institute. He has over 50 peer-reviewed publications roughly evenly split between applications and methods. He was awarded the Best Paper Prize (the Japanese Society of Computational Statistics, 1997) and Hida-Mizuno Prize (the Behaviormetric Society of Japan, 2003). He is the Editor-in-Chief of the *Journal of the Japanese Society of Computational Statistics* (2007–2010).

Cross References

►Biostatistics
►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest
►Design of Experiments: A Pattern of Progress
►Medical Research, Statistics in
►Medical Statistics

References and Further Reading

Armitage P (1972) History of randomised controlled trials. Lancet 299:1388

Armitage P (1991) Interim analysis in clinical trials. Stat Med 10: 925–937

Bayar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellen-berg JH et al (1976) Randomized clinical trials: perspective on some recent idea. New Engl J Med 295:74–80

Bull JP (1959) The historical development of the clinical trials. J Chron Dis 10:218–248

Byar DP (1991) Comment on "Ethics and statistics in randomized clinical trials" by R.M. Royall. Stat Sci 6:65–68
Cowan DH (1981) The ethics of trials of ineffective therapy. IRB: Rev Human Subjects Res 3:10–11
Day S (2006) The development of clinical trials. In: Machin D, Day S, Green S (eds) Textbook of clinical trials, 2nd edn. Wiley, Chichester, pp 5–11
Diehl HS, Baker AB, Cowan DW (1938) Cold vaccines: an evaluation based on a controlled study. J Am Med Assoc 11: 1168–1173
Ederer F (2005) Clinical trials, history of. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, 2nd edn. Wiley, Chichester, pp 864–874
Gail MH (1996) Statistics in action. J Am Stat Assoc 91:1–13
Gill DBEC (1996) Early controlled trials. Brit Med J 312:1298
Hart PD (1972) History of randomised controlled trials. Lancet 299:965
Hart PD (1996) Early controlled clinical trials. Brit Med J 312: 378–379
Hill AB, Marshall J, Shaw DA (1960) A controlled clinical trial of long-term anticoagulant therapy in cerebrovascular disease. Q J Med 29:597–608
Hill G, Forbes W, Kozak J, MacNeil I (2000) Likelihood and clinical trials. J Clin Epidemiol 53:223–227
Laplace PS (1814) Théori analytique des probabilités. Courcier, Paris
Lilienfeld AM (1949) Ceteris paribus: the evaluation of the clinical trial. Bull Hist Med 56:1–18
Lind J (1753) A treatise of the scurvy. Sands Murray & Cochran, Edinburgh
Louis PCA (1837) The applicability of statistics to the practice of medicine. London Medical Gazette 20:488–491
Medical Research Council (1948) Streptomycin treatment of pulmonary tuberculosis. Brit Med J 2:769–782
Medical Research Council (1950) Clinical trials of antihistaminic drugs in the prevention and treatment of the common cold. Brit Med J 2:425–431
Medical Research Council (1952) Chemotherapy of pulmonary tuberculosis in young adults. Brit Med J 1:1162–1168
Medical Research Council (1954) A comparison of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis I. Brit Med J 1:1223–1227
Medical Research Council (1955) A comparison of cortisone and aspirin in the treatment of early cases of rheumatoid arthritis II. Brit Med J 2:695–700
Meinert CL (1986) Clinical trials: design, conduct and analysis. Oxford University, New York
Mosteller F (1981) Innovation and evaluation. Science 211: 881–886
Pocock SJ (1984) Clinical trials: a practical approach. Wiley, Chichester
Purchas S (1625) Hakluytus posthumus or purchas his pilgrimes: contayning a histrory of the world in sea voyages and lande travells by englishmen and others. James MacLehose & Sons, Glasgow (reprinted, 1905)
Rheumatics Fever Working Party (1960) The evaluation of rheumatic heart disease in children: five years report f a co-operative clinical trials of ACTH, cortisone, and aspirin. Circulation 22:505–515
Royall RM (1991) Ethics and statics in randomized clinical trials. Stat Sci 6:52–88
Slotki JJ (1951) Daniel, Ezra, Nehemiah, Hebrew: text and english translation with introductions and commentary. Soncino Press, London
Smith R (1998) Fifty years of randomized controlled trials. Brit Med J 317:1166
Tukey JW (1977) Some thoughts on clinical trials, especially problems of multiplicity. Science 198:679–684
Zelen M (1979) A new design for randomized clinical trials. N Engl J Med 300:1242–1246

Clinical Trials: An Overview

Hiroyuki Uesaka
Osaka University, Osaka, Japan

A clinical trial is one type of clinical research where a procedure or drug is intentionally administered outside the realm of standard medical practice to human subjects with the aim of studying its effect on the human body. This includes medications, operations, psychotherapy, physiotherapy, rehabilitation, nursing, restricted diets, and the use of medical devices. The comparative study of two or more treatments, involving the random assignment of treatments to patients, is considered a clinical trial even if the study includes approved drugs or medical devices. This means that a clinical trial is an experiment which includes human subjects. It is necessary to distinguish clinical trials from observational studies which collect outcomes when executing a study treatment as an ordinary treatment.

Since clinical trials include human subjects, the ethical aspects, i.e., the rights, safety and well-being of individual research subjects, should take precedence over all other interests at all stages, from the planning of clinical trials to the reporting of results. Such ethical principles for clinical research are in accordance with the Declaration of Helsinki, "Ethical Principles for Medical Research Involving Human Subjects," issued by the World Medical Association (1964). In conducting a clinical trial, the study protocol should clearly describe the plan and content of the trial. The protocol must also be reviewed and approved by the ethics committee. Furthermore, the Declaration of Helsinki states: The protocol should contain a statement of the ethical considerations involved and indicate how the principles in the above declaration have been addressed. To protect the safety, well-being and rights of the human subjects participating in the trial, the Declaration indicates that potential subjects must be adequately informed of all

relevant aspects of the study which include aims, methods, the anticipated benefits and potential risks of the study and any discomfort that participation may entail. And it states: The potential subjects must be informed of their right to refuse to participate in the study or to withdraw consent to participate at any time without reprisal. The voluntary agreement of a subject to participate after sufficient details have been provided is called informed consent. It is also recommended that the clinical trial be registered in a publicly accessible database, and the results from the trial should be made publicly available, regardless of whether the results are positive or negative.

Clinical trials for a new drug application, when they are conducted in the EU, Japan and/or the United States of America, must meet the requirements of the "Good Clinical Practice" (GCP) guideline (ICH Steering Committee 1996). The GCP guideline is a unified standard provided by the Europe, Japan and the United States in the framework of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use [http://www.ich.org/]. The GCP guideline provides protection for the safety, well-being and human rights of subjects in clinical trials in accordance with the Declaration of Helsinki. The GCP guideline also requires that people appointed by the sponsor, the so-called monitors, verify that the rights and well-being of all human subjects are being protected, that the reported trial data are accurate, complete, and verifiable from source documents, and that the conduct of the trial is in compliance with the currently approved protocol/amendment(s), with the GCP, and with the applicable regulatory requirement(s). This is referred to as trial monitoring in the GCP.

In planning a clinical trial, a protocol must be prepared, including descriptions of the trial justification, trial objectives, study treatments, the population to be studied as defined by the study inclusion and exclusion criteria, test treatments and treatment procedures, observed variables and observation procedures, specification of variables to assess treatment effect, collection of safety information, prohibited concomitant medications, discontinuation criteria for individual subjects, the number of subjects planned to be enrolled and the justification for such, statistical methods to be employed, data collection, quality control and quality assurance (ICH Steering Committee 1997). A case report form (CRF) should be prepared as well. The CRF is a document designed to record all of the required information to be reported according to the protocol. After a trial, a so-called clinical study report is prepared (ICH Steering Committee 1995). This is a document which contains clinical and statistical descriptions of the methods, rationale, results and analyzes of a specific clinical trial fully integrated into a single report. Clinical trials are conducted as collaborative activities involving many specialists, such as investigators, nurses, diagnostic testing specialists and other collaborators. Furthermore, regulators are involved in new drug applications. Therefore, the protocol, CRF and clinical study report should be clearly and accurately documented to be easily understood by those involved in the trial and by those who will make use of the trial results.

Clinical trials can be classified into several types depending on various features (ICH Steering Committee 1998; ICH Steering Committee 2000). First, a trail can be controlled or uncontrolled, this being determined by the presence of a control group. A controlled trial is a trial to compare the study treatment(s) with a control treatment that is either the current standard treatment, best supportive care, placebo, or some other treatment; an uncontrolled trial involves giving the same treatment to all of the subjects participating in the trial. The second feature involves the objective of a trial, either exploratory or confirmatory. A clinical trial that aims to generate or identify a research topic, or provide information to determine the specifics of a trial method is called an exploratory trial. A confirmatory trial is defined as an adequately controlled trial where hypotheses which were derived from earlier research or theoretical considerations are stated in advance and evaluated. Furthermore, a confirmatory trial generally includes three types of comparisons: a superiority trial, a non-inferiority trial, and an equivalence trial. A superiority trial is used to show the superiority of a test treatment over a control. A non-inferiority trial is designed to show that the efficacy or safety of the study treatment is no worse than that of the control. An equivalence trial serves to demonstrate that the test treatment is neither better nor worse than the control. The third aspect involves distinguishing between a pragmatic and an explanatory trial (Gent and Sackett 1979; Schwartz and Lellouch 1967). The objective of a pragmatic trial is to confirm effectiveness of the test treatment for those subjects who are assigned to the test treatment. An explanatory trial serves to establish a biological action for the treatment. Finally, the fourth characteristic focuses on the difference between a single- and a multi-center trial. The single-center trial is conducted by a single investigator, and the multi-center trial is co-conducted by multiple investigators at multiple study sites. Recently, many multi-center trials have been planned and conducted across not only a single country but also two or more countries. Such a multi-center trial is called a multinational trial.

The clinical development of a new drug advances in stages (ICH Steering Committee 1997). A safety trial is executed first to determine the maximum dose that can be safely administered to a subject. In most safety trials of the first use of a new drug in humans, the subjects are healthy volunteers. Administration of the study treatment begins from a dosage expected to be safe enough for normal healthy volunteers, and then the dosage is increased in stages. The pharmacokinetic profile is usually examined in the same trial. Pharmacokinetics investigates the process of drug disposition which usually consists of absorption, distribution, metabolism and excretion. This stage is called Phase I. The next stage is to determine the dosage range that can be safely administered to patients and at which sufficient effectiveness can be expected. The dosage that will be used in clinical treatment as well as the dose intervals are also clarified at this stage. This is called the Phase II. In the third stage, the efficacy and safety of the study treatment is confirmed in the target patient population. This stage is referred to as Phase III. The dose and dosage regimen which are confirmed to be efficacious and safe in phase III are then submitted to the regulatory authority to obtain marketing authorization of the new drug. After marketing authorization is obtained, the drug becomes widely used for clinical treatment. This stage is called Phase IV. During the phase III trials many restrictions are imposed to ensure the safety of the participating subjects. These include the necessity of physical examinations, collection of patient anamneses, regulation of concomitant medications, and clearly defined test treatment administration periods. However, in phase IV such restrictions are relaxed and the approved study treatment can be used by various patients under diverse conditions. Therefore, because the number of patients who are administered the newly approved drug increases rapidly, with patients often using the drug for very long times according to their disease condition, there is a real concern about harmful effects that have not been anticipated. Therefore, an investigation to clarify the safety and effectiveness of the treatment in daily life, an observational study, a large-scale trial, or a long-term trial, is conducted. Moreover, a clinical trial to compare the newly approved drug with other medicines that have been approved for the same indication may also be conducted.

The result of the trial should be scientifically valid. Clinical trial results are intended to be applied to a target population defined by inclusion and exclusion criteria for a given trial. The enrolled subjects should be a random sample from the target population so that the trial results can be applied to the target population. However a trial is conducted in a limited number of medical sites, and not all candidate subjects give informed consent. Therefore, whether or not the trial result can be generalized to the target population will depend on the study protocol and the actual execution procedure. Accordingly, it is preferable to execute the trial in a variety of medical institutions with a wide range of patients corresponding to the diversity of the target population to improve the possibility of generalizing to the target population. A controlled trial usually estimates the difference in response to treatments between treatment groups. As described above, the clinical trial participants are not a random sample of the target population. Therefore the true mean difference in the study population (all subjects who participate in the trial) will be estimated. This is accomplished by dividing the study population into two or more treatment groups which are assigned to different treatments, and then comparing the means of response to treatment between groups. The estimated mean difference is usually different from this true value. When random allocation of treatment to subjects is used, it is assumed that the departure from the true difference is probabilistic or random error. However, there is the possibility of systematic error due to the execution procedure of the trial. This systematic error is called bias (ICH Steering Committee 1998). The execution of treatment, evaluation of results, and/or subjects' reactions can be influenced if the people involved in a trial, such as investigators, relevant clinical staff or subjects, are aware of which treatment is assigned to subjects. Therefore, masking (blinding) and randomization are used to prevent participants from knowing which treatment is being allocated to which subjects. There are several levels of blinding: double-blind, single-blind, observer-blind and open-label. In a double-blind study neither the subjects, nor the investigator, nor any of the relevant clinical trial staff know who belongs to which treatment group. In a single-blind study only treatment assignments are unknown to the subjects or investigator and relevant clinical staff. In an observer-blind study treatment assignments are unknown to the observers who assess the subjects' conditions. In an open-label study treatment assignments are known to both investigators and subjects.

One of the typical methods of treatment assignment is to assign only one treatment to each subject, and then to compare the effects of the treatments between subject groups. This method is referred to as parallel group design. The other typical method is the cross-over design where one subject receives two or more treatments and an intra-subject comparison of treatments is done. It is necessary to select an appropriate design because bias can be caused by the design itself.

A clinical trial is an experiment with human beings as subjects. It is preferable that the number of subjects be as small as possible to protect the rights, health and welfare

of the subjects included in the trial. However, if the objective of the trial is not achieved, the reason for executing the trial is lost. Therefore, based on the estimated difference between the treatments, the trial should be designed to have sufficient precision to either detect such a difference if it truly exists, or to conclude that the difference is below a definite value based on concrete evidence. For this purpose, it is necessary to maintain high accuracy and precision in trials. To ensure the precision of a trial, it is important to consider the stratification of the study population, to make precise observations, and to secure a sufficient number of subjects.

The objective of these trials is to estimate beneficial and adverse effects, and to confirm a hypothesis about the effect of the study treatment. Even if the effect size of the test treatment is assumed to be of a given size, the true effect size may be less than assumed. When the gap between the actual and the assumed value is large, the planned number of subjects might be insufficient and, in some cases, many more subjects than originally planned will be needed. In such cases, a sequential design (Jennison and Turnbull 2000) and a more advanced adaptive design (Bretz et al. 2009) would be proposed.

About the Author

Dr. Hiroyuki Uesaka is a specially appointed Professor of The Center for Advanced Medical Engineering and Informatics, Osaka University, Suita, Japan. He has been working for pharmaceutical companies for about 40 years as a statistical expert of clinical trials. He authored more than 20 original statistical papers, and authored, coauthored and contributed to 6 books written in Japanese (including *Design and Analysis of Clinical Trials for Clinical Drug Development* (Asakura-Shoten, 2006) and *Handbook of Clinical Trials* (jointly edited with Dr. Toshiro Tango, Asakura-Shoten, 2006).

Cross References

►Biopharmaceutical Research, Statistics in
►Biostatistics
►Causation and Causal Inference
►Clinical Trials, History of
►Clinical Trials: Some Aspects of Public Interest
►Design of Experiments: A Pattern of Progress
►Equivalence Testing
►Hazard Regression Models
►Medical Research, Statistics in
►Medical Statistics
►Randomization
►Statistics Targeted Clinical Trials Stratified and Personalized Medicines
►Statistics: Controversies in Practice
►Statistics: Nelder's view

References and Further Reading

Bretz F, Koenig F, Brannath W, Glimm E, Posch M (2009) Adaptive designs for confirmatory clinical trials. Stat Med 28:1181–1217

Gent M, Sackett DL (1979) The qualification and disqualification of patients and events in long-term cardiovascular clinical trials. Thromb Hemosta 41:123–134

ICH Steering Committee (1995) ICH harmonised tripartite guideline structure and content of clinical study reports. Recommended for adoption at Step 4 of the ICH Process on 30 November 1995

ICH Steering Committee (1996) ICH Harmonized tripartite guideline. Guideline for good clinical practice. Recommended for adoption at step 4 of the ICH process on 1 May 1996

ICH Steering Committee (1997) ICH Harmonized tripartite guideline. General considerations for clinical trials. Recommended for adoption at step 4 of the ICH process on 17 July 1997

ICH Steering Committee (1998) ICH Harmonized tripartite guideline. Statistical principles for clinical trials. Recommended for adoption at step 4 of the ICH process on 5 February 1998

ICH Steering Committee (2000) ICH Harmonized tripartite guideline. Recommended for adoption at step 4 of the ICH process on 20 July 2000

Jennison C, Turnbull BW (2000) Group sequential methods with applications to clinical trials. Chapman & Hall/CRC, Boca Raton

Schwartz D, Lellouch J (1967) Explanatory and pragmatic attitudes in therapeutical trials. J Chron Dis 20: 637–648

World Medical Association (1964) Declaration of Helsinki. Ethical principles for medical research involving human subjects 1964. Amended by the 59th WMA General Assembly, Seoul, Korea 2008, http://www.wma.net/

Clinical Trials: Some Aspects of Public Interest

Jagdish N. Srivastava
CNS Research Professor Emeritus
Colorado State University, Fort Collins, CO, USA

Medical experiments, often called "clinical trials," are obviously extremely important for the human race. Here, we shall briefly talk, in layman's language, about some important aspects of the same which are of great public interest.

Side Effect of Drugs

The side effects of allopathic drugs are notorious; death is often included in the same. For degenerative diseases (as opposed to infectious diseases, as in epidemics) it is not clear to the author whether any serious effort is being made by the pharmaceutical companies to develop drugs

which actually cure diseases; the trend seems to be at best to maintain people on drugs for a long time (even for the whole life). Mostly, people live under varying forms of painkiller-surgery regimes.

However, in many institutions (for example, departments doing research on nutrition) there are people who are genuinely interested in finding cures, though often they do not possess the resources they need. Many things, considered true by the public, are not quite so. Consider preservatives and other food additives that are legal. They are found in varying quantities in most foods, and many people do not pay attention to this at all, and consume unknown amounts each day. The thought that they have no side effects is based on relatively (time-wise) small experiments and extrapolations there from. It is probably true that if some food with a particular preservative or a (combination of the same with others) is consumed, it may not have any noticeable effect within a short period. But, the worry that many thinkers have is whether consuming food (all the time ignoring preservatives that it may contain) will have a disastrous effect (like producing cancer, heart attack, diabetes, stroke, etc.) 20, 30, 40, or 50 years earlier than it would have been expected for an additives-free diet. (The fact that, now, teenagers and young ones in their twenties are developing such diseases which, in an earlier age, were found mainly among seniors only, is alarming.) A full scale clinical trial (to study the long term effect of preservatives etc.) will take more than a century, and has not been done. Thus, extrapolations proclaiming that such additives are safe are based on guess work only, and are not necessarily scientifically sound. We live in an age when shelf life has become more important than human life.

It is not even clear whether the damage done from side effects and the painkiller-surgery policies is limited to the increase in the periods of sickness of people, the intensities of such sickness, and the reduction in the age at death. The bigger question is whether there is an effect on the progeny, and for how many generations. We recall that in the processes of natural selection in the theory of evolution, only the fittest may survive. Clearly, for the human race, only the policy that promotes the good of the general public corresponds to being fit for survival.

Contradictory Statements by Opposing Camps of Medical Researchers

Often, seemingly good scientists are found to be contradicting each other. For example, there may be a substance (say, an extract from some herbs) which may be claimed by some nature-cure scientists (based on their experiments) to positively affect some disease (relative to a placebo). However, some pharmaceutical scientists may claim that their experiments show that the drug is no better than the placebo. This is to say that, often in such cases, a close look may reveal that the two sets of experiments are not referring to the same situation. To illustrate, the subjects (people, on whom an experiment is done) in the first group may be people who just contracted the disease, these people being randomly assigned the drug or the placebo. In the second case, the subjects may be people who have had the disease for some time and have been taking painkillers. Now, the herbal drug may be quite effective on a body which is in a more primeval and natural state, and yet not work well in a body which has been corrupted by the chemicals in the painkiller. Clearly, that would explain the discrepancy and support the use of the herbal drug soon after the disease begins, simultaneously discouraging the use of painkillers etc. whose primary effect is to temporarily fool the mind into thinking that one is feeling better. Thus, it is necessary to examine a clinical trial closely rather than take its results on face value.

Large Clinical Trials: Meta-analysis

"Large" clinical trials are often touted as being very "informative." To illustrate, take the simple case of comparing two drugs *A* and *B* with respect to a placebo *C*. Now, how effective a drug is for a person may depend upon his or her constitution. On some people, *A* may be the best, on some *B*, and on others, all the three may be essentially useless. For me, even though I may not know the reality, suppose the reality is that *B* would be very effective with little negative side effect, *A* would be only somewhat effective but with a large negative side effect, and the effect of *C* would be small (being somewhat positive or somewhat negative depending on environmental factors). Suppose a trial is done in Arizona, involving 6,000 patients randomly divided into three equal groups, the result being that *A* is effective in 45% (cases in its group), *B* in 35%, and *C* in 5% cases. Clearly, here, the drug *A* wins. But, for me, what is the value of this information? I really need to know which drug would be best for me.

Now suppose a similar trial is done in Idaho and in California, the result for *A*, *B*, and *C* being 33%, 42%, 7%, and 54%, 52%, and 30% respectively in the two states. Does this help me in some way or does it simply add to the confusion? The drugs manufacturer, Mr. Gaines, would like "meta–analysis" (whose purpose is to combine the results in a legitimate and meaningful way), because his interest is in seeing the overall picture so that he can formulate an appropriate manufacturing policy for his company. However, the interest of the general public is different from that of Gaines, because each individual needs to know what is good for him or her personally. The individual's interest, in a sense, runs counter to ►meta–analysis; he or she would be more interested in knowing what aspects of a person's

health make him or her more receptive to *A* or *B*. Instead of combining the data, more delineation needs to be done. In other words, one needs to connect the results with various features of the subjects and other related factors. Then we may gain knowledge not only on what proportion of subjects are positively affected by a drug, but what bodily features of people (or the food that they eat, or their lifestyle, or the environment around them, etc.) lead to this positive effect.

For example, in the above (artificial) data, it seems that *A* is better in a warm climate and *B* in cold. Where the climate is mild, all of them do well, and many people may recover without much of drugs. If I have been more used to a cold climate, *B* may be more effective on me. With this knowledge, even though *A* may turn out to be much better than *B* in the area where I live, *B* may be better for me individually.

(This leads us to the *philosophy* of statistical inference. Not only do we need to plan our experiments or investigations properly, we need to be careful in drawing inferences from the data obtained from them. According to the author, trying to find what a bunch of data "says" must involve in a relevant way the space of applications where such a finding will be made use of. Many scholars believe that, given a set of data, the "information" that the data contains is a fixed attribute of the data, and the purpose of inference is to bring out this attribute accurately. The author believes that the reason why the inference is sought (in particular, to what use or application the inference will be put) is also important, and should have a bearing on the inference drawn. This policy will give insight into the kind of information we need, what should receive more emphasis, etc. Clinical trials would really gain from this approach.)

Reducing Side Effects of Drugs

Studies are usually done using a "loss function" which tells how much "loss" shall we incur by adopting each of a set of policies. For example, we may have many drugs, several possible doses of a drug per day, many possible durations of time over which a drug is to be continued, etc. For each combination of these factors, the "loss" may be "the total time of absence from work," or "the total financial loss incurred because of sickness," or "the amount of fever," or "the blood pressure," etc. If the loss function involves only one variable (like "blood pressure"), it is "uni-dimensional." But if, many variables are involved simultaneously (like "blood pressure," "fever," "financial loss"), then it is called multi-dimensional. Usually, only one dimension is used or emphasized (like, "intensity of fever"). More theory needs to be developed on how to work with multi-dimensional loss functions.

Besides theory, we also need to develop good quantitative criteria for measuring "healthfulness." There can be various sectors. For example, we can have one criterion for the sector of upper digestive track, one for the middle, one for the colon, one for the respiratory system, one for bone diseases, one for the joints, one for nerves, one for cancerous growth, and so on. For each sector, the corresponding criterion will provide a measure of how healthy that sector is. Suppose we decided to have 25 such sectors. Then the loss function will be 25-dimensional. The drugs will be evaluated in each dimension, and the results will also be combined in various ways. The side effect of a drug in a particular sector will be caught more easily. When the drug is marketed, an assessment for each sector can be provided. Drugs with large effect in any sector can be rejected.

Experiments with Many Factors: Interactions

We make some technical remarks here. A large part of the field of statistical design of multi-factorial scientific experiments is concerned with the simplistic situation when there are either no interactions or the set of non-negligible interactions is essentially known (though the values of these interactions and the main effects are not known). However, in medical experiments, we can have interactions of even very high orders. Thus, for the field of multifactor clinical trials, we have to go beyond Plackett–Burman designs, and orthogonal arrays of small strength (such as 2). There is work available on search theory by the author and others, which would help. However, further work is needed in that field. Indeed, for vigorous full fledged research on how to cure diseases, the statistical theory of the design and analysis of multifactor multi-response experiments need to be developed much further beyond its current levels. However, the basics of the same are available in books such as Roy et al. (1970). For the reader who wishes to go deeper in the field of this article, some further references are provided below.

About the Author

Jagdish N. Srivastava is now CNS Research Professor Emeritus at Colorado State University, where he worked during 1966–2004 as Full Professor, and where he also once held a joint appointment in Philosophy. He was born in Lucknow, India on 20 June 1933. He did his Ph.D. in 1961 (at the University of North Carolina, Chapel Hill, NC, USA). He was President (Indian Society of Agricultural Statistics, 1977 Session; International Indian Statistical Association (IISA) (1993–1997); Forum for Interdisciplinary Mathematics (1994–1996). He visited institutions all over the world. His honors include the Fellowship of the American Statistical Association, Institute of Mathematical Statistics,

Institute of Combinatorial Mathematics and its Applications, IISA (Honorary), International Statistical Institute, and TWAS (The World Academy of Science for developing countries). He is the founder (1975) and Editor-in-Chief of the monthly *Journal of Statistical Planning and Inference*. Two volumes containing papers by experts from all over the world were published in honor of his 65th birthday. He discovered, among other results, the "Srivastava Codes," the "Srivastava Estimators," and the Bose-Srivastava Algebras. He has been interested in, and has contributed to, Science and Spirituality, and is a major researcher in Consciousness and the "Foundations of Reality" (including, in particular, "Quantum Reality").

Cross References

►Clinical Trials, History of
►Clinical Trials: An Overview
►Medical Research, Statistics in
►Medical Statistics
►Statistics Targeted Clinical Trials Stratified and Personalized Medicines

References and Further Reading

Roy SN, Gnanadesikan R, Srivastava JN (1970) Analysis and design of quantitative multiresponse experiments. Pergamon, Oxford, England

Srivastava JN (1990) Modern factorial design theory for experimenters. In: Ghosh S (ed) Statistical design and analysis of industrial experiments. Marcel Dekker, New York, pp 311–406

Srivastava JN (1996) A critique of some aspects of experimental designs. In: Ghosh S, Rao CR (eds) Handbook of statistics, vol 13. Elsevier, Amsterdam, pp 309–341

Srivastava JN (2006) Foods, drugs and clinical trials: some outrageous issues. J Combinatorics Inf Syst Sci 31(1–4):365–378

Srivastava JN (2008) Some statistical issues concerning allopathic drugs for degenerative diseases. J Indian Soc Agric Stat 62: 120–125

Cluster Analysis: An Introduction

H. Charles Romesburg
Professor
Utah State University, Logan, UT, USA

Cluster analysis is the generic name for a variety of mathematical methods for appraising similarities among a set of objects, where each object is described by measurements made on its attributes. The input to a cluster analysis is a *data matrix* having t columns, one for each object, and n rows, one for each attribute. The (i, j)th element of the data matrix is the measurement of the ith attribute for the jth object. The output from a cluster analysis identifies groups of similar objects called *clusters*. A cluster may contain as few as one object, because an object is similar to itself.

Applications of cluster analysis are widespread because the need to assess similarities and dissimilarities among objects is basic to fields as diverse as agriculture, geology, market research, medicine, sociology, and zoology. For example, a hydrologist considers as the objects a set of streams, and for attributes describes each stream with a list of water quality measures. A cluster analysis of the data matrix identifies clusters of streams. The streams within a given cluster are similar, and any stream in one cluster is dissimilar to any stream in another cluster.

There are two types of cluster analysis. *Hierarchical cluster analysis* is the name of the collection of methods that produce a hierarchy of clusters in the form of a *tree*. The other type, *nonhierarchical cluster analysis*, is the name of the collection of methods that produce the number of clusters that the user specifies. For both types, computer software packages containing programs for the methods are available.

Let us illustrate the main features of hierarchical cluster analysis with an example where the calculations can be done by hand because the data matrix is small, five objects and two attributes, consisting of made-up data:

Data matrix						
		Object				
		1	2	3	4	5
Attribute	1	10	20	30	30	5
	2	5	20	10	15	10

To perform a hierarchical cluster analysis, we must specify: (1) a coefficient for assessing the similarity between any two objects, j and k; and (2) a clustering method for forming clusters.

For the first, let us choose the "Euclidean distance coefficient," e_{jk}. The smaller its value is, the more similar objects j and k are. If the value is zero, they are identical, i.e., maximally similar. For our example with $n = 2$ attributes, e_{jk} is the distance between object j and object k computed with the Pythagorean theorem. And for the clustering method, let us choose the "UPGMA method," standing for "unweighted pair-group method using arithmetic averages."

At the start of the cluster analysis, each object is considered to be a separate cluster. Thus with five objects, there are five clusters. For the five we compute the $5(5-1)/2 = 10$

values of e_{jk}. To demonstrate the calculation of one of these values, consider object 1 and object 5. The Euclidean distance e_{15} is

$$e_{15} = [(10-5)^2 + (5-10)^2]^{1/2} = 7.07.$$

In this manner, we compute the other values and put the ten in a list, from smallest, indicating the most similar pair of clusters (objects), to largest, indicating the least similar pair: $e_{34} = 5.0$, $e_{15} = 7.07$, $e_{24} = 11.2$, $e_{23} = 14.1$, $e_{12} = 18.0$, $e_{25} = 18.0$, $e_{13} = 20.6$, $e_{14} = 22.4$, $e_{35} = 25.0$, $e_{45} = 25.5$.

The two most similar clusters, 3 and 4, head the list, as the Euclidean distance between them is the smallest. Therefore,

Step 1 Merge clusters 3 and 4, giving 1, 2, (34), and 5 at the value of $e_{34} = 5.0$.

Next, for the four clusters – 1, 2, (34), and 5 – we obtain the $4(4-1)/2 = 6$ values of e_{jk}. Three of these values are unchanged by the clustering at step 1 and can be transcribed from the above list. The other three have to be computed according to the guiding principle of the UPGMA clustering method. It requires that we average the values of e_{jk} between clusters, like this:

$$e_{1(34)} = \tfrac{1}{2}(e_{13} + e_{14}) = \tfrac{1}{2}(20.6 + 22.4) = 21.5;$$
$$e_{2(34)} = \tfrac{1}{2}(e_{23} + e_{24}) = \tfrac{1}{2}(14.1 + 11.2) = 12.7;$$
$$e_{5(34)} = \tfrac{1}{2}(e_{35} + e_{45}) = \tfrac{1}{2}(25.0 + 25.5) = 25.3.$$

So, the six e_{jk} values listed in order of increasing distance are: $e_{15} = 7.07$, $e_{2(34)} = 12.7$, $e_{12} = 18.0$, $e_{25} = 18.0$, $e_{1(34)} = 21.5$, $e_{5(34)} = 25.3$. It follows that the two most similar clusters are 1 and 5, since the Euclidean distance between them is the smallest. Therefore,

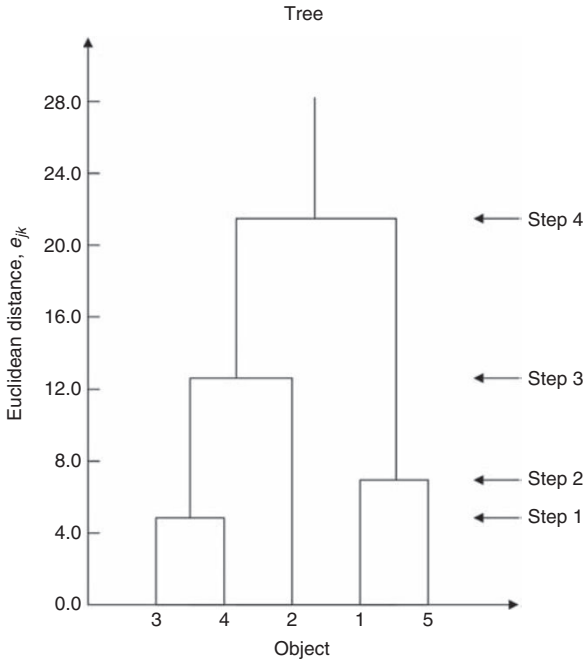
Step 2 Merge clusters 1 and 5, giving 2, (34), and (15) at the value of $e_{15} = 7.07$.

Before going to the next clustering step, we note that step 2 left the distance between clusters 2 and (34) unchanged at $e_{2(34)} = 12.7$. The two remaining distances are calculated according to the UPGMA clustering method by averaging the values of e_{jk} as follows:

$$e_{2(15)} = \tfrac{1}{2}(e_{12} + e_{25}) = \tfrac{1}{2}(18.0 + 18.0) = 18.0;$$
$$e_{(15)(34)} = \tfrac{1}{4}(e_{13} + e_{14} + e_{35} + e_{45}) = \tfrac{1}{4}(20.6 + 22.4 + 25.0 + 25.5) = 23.4.$$

The list of e_{jk} in increasing distance is now: $e_{2(34)} = 12.7$, $e_{2(15)} = 18.0$, $e_{(15)(34)} = 23.4$. The two most similar clusters, 2 and (34) head the list. Therefore,

Step 3 Merge clusters 2 and (34), giving (15) and (234) at the value of $e_{2(34)} = 12.7$.



Cluster Analysis: An Introduction. Fig. 1 Tree showing the hierarchy of similarities between the five objects specified by the data matrix in the text

At this point there are two clusters: (15) and (234). The average Euclidean distance between them is:

$$e_{(15)(234)} = \tfrac{1}{6}(e_{12} + e_{13} + e_{14} + e_{25} + e_{35} + e_{45}) = \tfrac{1}{6}(18.0 + 20.6 + 22.4 + 18.0 + 25.0 + 25.5) = 21.6.$$

The list of e_{jk} has only one value: $e_{(15)(245)} = 21.6$. Therefore,

Step 4 Merge clusters (15) and (234), giving (12345) at the value of $e_{(15)(234)} = 21.6$.

The calculations are finished. With each step, the tree (Fig. 1) has been growing. It summarizes the clustering steps, e.g., showing that the branches containing cluster (34) and cluster 2 join at an Euclidean distance value of 12.7.

The tree is a hierarchical ordering of similarities that begins at the tree's bottom where each object is separate, its own cluster. As we move to higher levels of e_{jk}, we become more tolerant and allow clusters to hold more than one object. When we reach the tree's top we are completely tolerant of the differences between objects, and all objects are considered as one cluster.

Suppose we took the five objects in the data matrix and plotted them on a graph with attribute 1 as one axis and

attribute 2 as the other. We would see that the distances between the objects suggest clusters that nearly match those in the tree. However, real applications typically have many attributes, often more than a hundred. In such cases, the researcher cannot grasp the inter-object similarities by plotting the objects in the high-dimension attribute space. Yet cluster analysis will produce a tree that approximates the inter-object similarities.

A tree is an old and intuitive way of showing a hierarchy. Witness the tree of life forms, the Linnaean classification system. At its bottom is the level of Species, at the next higher hierarchical level is the Genus, consecutively followed by levels of Order, Class, Phylum, and Kingdom.

A widely practiced way of creating a classification of objects is to perform a hierarchical cluster analysis of the objects. On the tree, draw a line perpendicular across the tree's axis, cutting it into branches, i.e., clusters. The objects in the clusters define the classes. Details may be found in Romesburg (2004) and in Sneath and Sokal (1973).

There are several general points to note about hierarchical cluster analysis:

1. There are various coefficients that can be used to assess the similarity between clusters. Of these, there are two types: dissimilarity coefficients and similarity coefficients. With a dissimilarity coefficient (as the Euclidean distance coefficient is), the smaller its value is, the more similar the two clusters are. Whereas with a similarity coefficient, the larger its value is, the more similar the two clusters are. An example of a similarity coefficient is the Pearson product moment correlation coefficient. But whether a dissimilarity coefficient or a similarity coefficient is used, a clustering method at each step merges the two clusters that are most similar.
2. Although the UPGMA clustering method (also called "average linkage clustering method") is perhaps most often used in practice, there are other clustering methods. UPGMA forms clusters based on the average value of similarity between the two clusters being merged. Another is the SLINK clustering method, short for "single linkage" clustering method, and sometimes called "nearest neighbor" clustering method. When two clusters are joined by it, their similarity is that of their most similar pair of objects, one in each cluster. Another is the CLINK clustering method, short for "complete linkage" clustering method, and sometimes called "furthest neighbor" clustering method. When two clusters are joined by it, their similarity is that of the most dissimilar pair of objects, one in each cluster. Another is Ward's clustering method, which assigns objects to clusters in such a way that a sum-of-squares index is minimized.
3. The data in the data matrix may be measured on a continuous scale (e.g., temperature), an ordinal scale (e.g., people's ranked preference for products), or on a nominal scale for unordered classes (e.g., people's sex coded as 1 = female, 0 = male).

   For an illustration of nominal scale measurement, suppose a military researcher takes a set of aircraft as the objects, and for their attributes records whether or not an aircraft can perform various functions. If the jth aircraft is able to perform the ith function, the (i, j)th element of the data matrix is coded with a "1"; if it is unable to perform the ith function, it is coded with a "0." In this way, the data matrix consists of zeroes and ones. A similarity coefficient, such as the one named "the simple matching coefficient," gives a numerical value for the similarity between any two aircraft. The cluster analysis produces a tree which shows which of the aircraft are functionally similar (belong to the same cluster) and which are functionally dissimilar (belong to different clusters).
4. Whenever the attributes of the data matrix are measured on a continuous scale, it is sometimes desired to standardize the data matrix. Standardizing recasts the units of measurement of the attributes as dimensionless units. Then the cluster analysis is performed on the standardized data matrix rather than on the data matrix. There are several alternative ways of standardizing (Romesburg 2004).
5. Commercial software packages for performing hierarchical cluster analysis include SPSS, SAS, CLUSTAN, and STATISTICA. Of these, SPSS is representative, allowing the user a choice of about 35 similarity/dissimilarity coefficients and seven clustering methods.
6. In the literature of cluster analysis, certain terms have synonyms. Among other names for the objects to be clustered are "cases," "individuals," "subjects," "entities," "observations," "data units," and "OTU's" (for "operational taxonomic units"). Among other names for the attributes are "variables," "features," "descriptors," "characters," "characteristics," and "properties." And among other names for the tree are the "dendrogram" and the "phenogram."

In contrast to hierarchical cluster analysis, nonhierarchical cluster analysis includes those clustering methods that do not produce a tree. The software packages mentioned above have programs for nonhierarchical cluster analysis. Perhaps the most-used nonhierarchical method is *K-means cluster analysis*. For it, the user specifies k, the number of clusters wanted, where k is an integer less than t, the number of objects. Software programs for K-means

cluster analysis usually differ a bit in their details, but they execute an iterative process to find clusters, which typically goes like this:

To begin the first iteration, the program selects k objects from the data matrix and uses them as k cluster seeds. The selection is made so that the Euclidean distances between the cluster seeds is large, which helps insure that the seeds cover all regions of the attribute space in which the objects reside.

Next, the program forms tentative clusters by sequentially assigning each remaining object to whichever cluster seed it is nearest to. As objects are assigned, the cluster seeds are recomputed and made to have the attribute values that are the average of those of the objects in the clusters. Hence, cluster seeds generally change as objects are tentatively assigned to clusters.

When the first iteration is finished, the resulting cluster seeds are taken as the k initial seeds to start the second iteration. Then the process is repeated, sequentially assigning the objects to their nearest cluster seed, and updating the seeds as the process moves along.

Finally, after a number of iterations, when the change in the cluster seeds is tolerably small from one iteration to the next, the program terminates. The k final clusters are composed of the objects associated with the k cluster seeds from the final iteration.

We now turn to the question, "Which is the better method for finding clusters – hierarchical cluster analysis or nonhierarchical cluster analysis?" The answer depends. Broadly speaking, researchers like having a choice of a large variety of similarity/dissimilarity coefficients, and like having the similarities among clusters displayed as a hierarchy in the form of a tree – two features that hierarchical methods offer but nonhierarchical methods do not. However, for hierarchical methods the amount of computation increases exponentially with the number of objects. Whereas for nonhierarchical methods the amount of computation increases less than exponentially because the methods do not require the calculation of similarities between all pairs of objects. In any event, all of the software packages mentioned above can handle very large data matrices for hierarchical methods and for nonhierarchical methods. For instance, according to the literature that accompanies CLUSTAN's hierarchical cluster analysis program, the limit to the size of a data matrix that at present can be processed in a reasonable time on a basic PC is in the neighborhood of 16,000 objects and 1,000 attributes. For more objects than that, CLUSTAN's nonhierarchical K-means program can handle as many as a million objects.

Books that provide detailed accounts of hierarchical cluster analysis and nonhierarchical cluster analysis include those by Aldenderfer and Blashfield (1984), Everitt (1993), and Romesburg (2004).

About the Author

Dr. H. Charles Romesburg is Professor of Environment and Society at Utah State University and holds degrees in Operations Research and Biostatistics, Nuclear Engineering, and Mechanical Engineering. He is the author of four books, including *Cluster Analysis for Researchers* (North Carolina: Lulu, 2004) and *Best Research Practices* (North Carolina: Lulu, 2009). He is an active and prolific researcher with numerous scientific articles to his credit. His publications in which he is the sole author have been cited more than 1,000 times. The Wildlife Society has awarded him its Wildlife Publication Award for his article "Wildlife Science: Gaining Reliable Knowledge."

Cross References

►Data Analysis
►Distance Measures
►Fuzzy Logic in Statistical Data Analysis
►Hierarchical Clustering
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Random Permutations and Partition Models

References and Further Reading

Aldenderfer MS, Blashfield RK (1984) Cluster analysis. Sage, Beverly Hills

Everitt B (1993) Cluster analysis. E. Arnold, London

Romesburg HC (2004) Cluster analysis for researchers. Lulu.com, North Carolina

Sneath PHA, Sokal RR (1973) Numerical taxonomy: the principles and practice of numerical classification. W. H. Freeman, San Francisco

Cluster Sampling

Janusz Wywial
Professor
Katowice University of Economics, Katowice, Poland

The quality of statistical inference is dependent not only on, for example, estimator construction but on the structure of a population and a sampling scheme too. For example, let the estimation of total wheat production in a population of farms be considered. The population of farms is divided into clusters corresponding to villages. This estimation can be based on the ordinary simple sample or on the cluster sample. Population units can be selected to the sample by means of

several sampling schemes. The units (i.e., farms) can be selected to the ordinary sample, or clusters of the units (i.e., villages) can be drawn to the cluster sample. The accuracy of the estimation depends on the sampling scheme and on the intraclass spread of a variable under study (wheat production). When should the ordinary simple sample be used and when should the cluster one?

Let us consider a fixed and finite population case. The fixed and finite population of the size N is denoted by $\Omega = \{\omega_1, \ldots, \omega_N\}$, where ω_k is an element (unit) of the population U. Let us assume that Ω is divided into G mutually disjoint clusters Ω_k $(k = 1, \ldots, G)$ such that $\bigcup_{k=1}^{G} \Omega_k = \Omega$. The size of a cluster Ω_k is denoted by N_k. So, $0 \le N_k \le N$ and $\bigcup_{k=1}^{G} N_k = N$. Let $U = \{\Omega_1, \ldots, \Omega_G\}$ be the set of all clusters. The clusters are called units (elements) of the set U. The cluster sample is selected from the set U and it is denoted by $s = \{\Omega_{i_1}, \ldots, \Omega_{i_n}\}$. The size of s is denoted by n, where $0 \le n \le G$. Let $\mathbf{S}$ be the set (space) of samples. The cluster sample is a random one if it is selected from U according to some sampling design denoted by $P(s)$, where $P(s) \ge 0$ for $s \in \mathbf{S}$ and $\sum_{s \in S} P(s) = 1$. Let the inclusion probability of the first order be denoted by $\pi_k = \sum_{\{s: k \in s\}} P(s)$, $k = 1, \ldots, G$. A random sample is selected from a population by means of the so-called sampling scheme, which fulfills the appropriate sampling design. It is well known that a sample can be selected according to previously determined inclusion probabilities of the first order without any explicit definition of the sampling design. This inference simplifies our practical research. Frequently, the inclusion probabilities are determined as proportional to cluster sizes, so $\pi_k \propto N_k$ for $k = 1, \ldots, G$. In general, $\pi_k \propto x_k$, where $x_k > 0$ is the value of an auxiliary variable.

Let us note that it is possible to show that the well-known systematic sampling design is a particular case of the cluster sampling design. Moreover, the cluster sampling design is a particular case of two (or more) stage sampling designs.

In general, all known sampling designs and schemes can be applied to the cluster case. The examples of sampling designs and schemes are as follows: the simple cluster sample of fixed size n, drawn without replacement, is selected according to the following sampling design: $P(s) = 1/\binom{G}{g}$ for $s \in \mathbf{S}$. The inclusion probability of the first order is $\pi_k = \frac{g}{G}$. The sampling scheme fulfilling that sampling design is as follows: The first element (unit) of the set U is selected to the sample with the probability $1/G$, the next one with the probability $1/(G-1)$, the kth element with the probability $1/(G - k + 1)$, and so on until the nth element of the sample.

The sampling scheme selecting with replacement units to the sample of fixed size n is as follows: Each element of U is selected with probabilities equal to p_k, where, for example, $p_k = x_k / \sum_{i \in U} x_i$. So, elements are independently drawn to the sample n times. In this case, the sampling design is defined in a straightforward manner. Particularly, if $p_k = 1/G$ for all $k = 1, \ldots, G$, the simple cluster sample drawn with replacement is selected according to the sampling design $P(s) = 1/G^n$. In this case, each element of U is selected with the probability $1/G$ to the sample of size n.

The so-called Poisson without replacement sampling scheme is as follows: The kth unit is selected with the probability p_k, $0 < p_k \le 1$, $k = 1, \ldots, G$. In this case, the sample size is not fixed because $0 \le n \le G$. There exists the so-called conditional without replacement sampling design of a fixed sample size, but unfortunately its sampling schemes are complicated, see, for example, Tillé (2006). Additionally, let us note that the cluster sample can be useful in the case of estimating the population mean.

It is well known that the precision of a population mean estimation, performed on the basis of the simple cluster sample, depends on the so-called intraclass (intracluster) correlation coefficient, see, for example, Hansen et al. (1953) or Cochran (1963).

Let us assume that sizes of clusters are the same and equal to M and $N = GM$. The ordinary variance of the variable is defined by $v = \frac{1}{N} \sum_{k=1}^{G} \sum_{j \in \Omega_k} (y_{kj} - \bar{y})^2$, where $\bar{y}_k = \frac{1}{N} \sum_{k=1}^{G} \sum_{j \in \Omega_k} y_{kj}$ is the cluster sample. The intraclass and the betweenclass variances are given by the expressions: $v_w = \frac{1}{G(M-1)} \sum_{k=1}^{G} \sum_{j \in \Omega_k} (y_{kj} - \bar{y}_k)^2$ and $v_b = \frac{1}{G-1} \sum_{k=1}^{G} (\bar{y}_k - \bar{y})^2$, respectively, where $\bar{y}_k = \frac{1}{M} \sum_{j \in \Omega_k} y_{kj}$. The intraclass correlation coefficient is defined by the following expression: $r_I = \frac{2}{N^2 v} \sum_{k=1}^{G} \sum_{i \ne j \in \Omega_k} \sum (y_{ki} - \bar{y})(y_{kj} - \bar{y})$. The coefficient ρ takes its value from the closed interval $[-1/(M-1), 1]$. The coefficient r_I can be rewritten in the following forms: $r_I = (v_b - v_w/M)/v$, $r_I = 1 - v_w/v$ or $r_I = (Mv_m/v - 1)/(M-1)$. The expressions lead to the conclusion that the intraclaas correlation coefficient is negative (positive) when the ordinary variance is smaller (larger) than the intraclass variance or equivalent if the ordinary variance is larger (smaller) than the betweenclass variance divided by M.

Let us note that it is well known that the simple *cluster sample* mean is a more accurate estimator of the population mean than the simple sample mean when the

intraclass correlation coefficient is negative. So, this leads to the conclusion that, if only possible, a population should be clustered in such a way that the intraclass correlation coefficient takes the smallest negative value. A more complicated case of unequal cluster sizes was considered, for example, by Konijn (1973). In this case, Särndal et al. (1992) considered the so-called homogeneity coefficient, which is the function of the intraclass variance. On the basis of the cluster sample, not only the estimation of population parameters is performed but also testing statistical hypothesis, see, for example, Rao and Scott (1981).

Cross References

►Adaptive Sampling
►Intraclass Correlation Coefficient
►Multistage Sampling
►Sample Survey Methods
►Sampling From Finite Populations

References and Further Reading

Cochran WG (1963) Sampling techniques. Wiley, New York
Hansen MH, Hurvitz WN, Madow WG (1953) Sample survey methods and theory, vols I and II. Wiley, New York
Konijn HS (1973) Statistical theory of sample survey and analysis. North-Holland, Amsterdam
Rao JNK, Scott A (1981) The analysis of categorical data from complex sample surveys: chi-square tests of goodnees of fit and independence in two-way tables. J Am Stat Assoc 76(374): 221–230
Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York/Berlin/Heidelberg/London/Paris/Tokyo/Hong Kong/Barcelona/Budapest
Tillé Y (2006) Sampling algorithms. Springer, New York

Coefficient of Variation

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

Coefficient of variation is a relative measure of dispersion and it may be considered in three different contexts: in probability, in a data set or in a sample.

In the first context it refers to distribution of a random variable X and is defined by the ratio

$$v = \frac{\sigma}{\mu} \tag{1}$$

where $\mu = EX$ and $\sigma = \sqrt{E(X - EX)^2}$. It is well defined if $EX > 0$. Moreover it is scale-invariant in the sense that cX has the same v for all positive c.

Data series $x = (x_1, \ldots, x_n)$ corresponds to distribution of a random variable X taking values x_i with probabilities $p_i = \frac{k_i}{n}$, for $i = 1, \ldots, n$, where k_i is the number of appearance of x_i in the series. In this case the formula (1) remains valid if we replace μ by $\overline{x} = \frac{1}{n}\sum_i x_i$ and σ by $\sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$.

If $x = (x_1, \ldots, x_n)$ is a sample from a population, then the coefficient may be treated as a potential estimator of the coefficient of variation v in the whole population. Since $\frac{1}{n}\sum(x_i - \overline{x})^2$ is biased estimator of σ^2 in order to eliminate this bias we use the sample coefficient of variance in the form

$$\mathrm{v} = \frac{s}{\overline{x}}, \tag{2}$$

where $s = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{x})^2}$.

One can ask whether v is normalized, i.e., whether it takes values in the interval $[0, 1]$.

In spite of that v is well defined providing $\overline{x} > 0$ it seems reasonable to restrict oneself to the nonnegative samples x, i.e., satisfying the condition $x_i \geq 0$ for all i and $\sum_i x_i > 0$. Under this assumption the sample coefficient (2) of variance in the sample takes values in the interval $[0, \sqrt{n}]$ and the lower and upper bound is attained. Therefore it is not normalized.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Semi-Variance in Finance
►Standard Deviation
►Variance

References and Further Reading

Stępniak C (2007) An effective characterization of Schur-convex functions with applications. J Convex Anal 14:103–108

Collapsibility

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

Collapsibility in Contingency Tables

Consider the I by J by K contingency table representing the joint distribution of three discrete variables X, Y, Z, the I by J marginal table representing the joint distribution of X and Y, and the set of conditional I by J subtables (strata) representing the joint distributions of X and Y within levels

Collapsibility. Table 1 Trivariate distribution with (a) strict collapsibility of $Y|X$ risk differences over Z, (b) collapsibility of $Y|X$ risk ratios when standardized over the Z margin, and (c) noncollapsibility of $Y|X$ odds ratios over Z. Table entries are cell probabilities

	$Z = 1$		$Z = 0$		Collapsed over Z	
	$X = 1$	$X = 0$	$X = 1$	$X = 0$	$X = 1$	$X = 0$
$Y = 1$	0.20	0.15	0.10	0.05	0.30	0.20
$Y = 0$	0.05	0.10	0.15	0.20	0.20	0.30
Risks[a]	0.80	0.60	0.40	0.20	0.60	0.40
Differences	0.20		0.20		0.20	
Ratios	1.33		2.00		1.50	
Odds ratios	2.67		2.67		2.25	

[a]Probabilities of $Y = 1$ in column

of Z. A measure of association of X and Y is *strictly collapsible* across Z if it is constant across the strata (subtables) and this constant value equals the value obtained from the marginal table.

Noncollapsibility (violation of collapsibility) is sometimes referred to as ►Simpson's paradox, after a celebrated article by Simpson (1951). This phenomenon had however been discussed by earlier authors, including Yule (1903); see also Cohen and Nagel (1934). Some statisticians reserve the term Simpson's paradox to refer to the special case of noncollapsibility in which the conditional and marginal associations are in opposite directions, as in Simpson's numerical examples. Simpson's algebra and discussion, however, dealt with the general case of inequality. The term "collapsibility" seems to have arisen in later work; see Bishop et al. (1975).

Table 1 provides some simple examples. The difference of probabilities that $Y = 1$ (the risk difference) is strictly collapsible. Nonetheless, the ratio of probabilities that $Y = 1$ (the risk ratio) is not strictly collapsible because the risk ratio varies across the Z strata, and the odds ratio is not at all collapsible because its marginal value does not equal the constant conditional (stratum-specific) value. Thus, collapsibility depends on the chosen measure of association.

Now suppose that a measure is not constant across the strata, but that a particular summary of the conditional measures does equal the marginal measure. This summary is then said to be *collapsible* across Z. As an example, in Table 1 the ratio of risks averaged over (standardized to) the marginal distribution of Z is

$$\Sigma_z P(Y = 1|X = 1, Z = z)P(Z = z)/\Sigma_z P(Y = 1|X = 0, Z = z)$$
$$P(Z = z) = \{-0.8(0.5) + 0.4(0.5)\}/\{-0.6(0.5) + 0.2(0.5)\} = 1.50,$$

which is equal to the marginal (crude) risk ratio. Thus, the risk ratio in Table 1 is collapsible under this particular weighting (standardization) scheme for the risks.

Various tests of collapsibility and strict collapsibility have been developed (e.g., Whittemore 1978; Asmussen and Edwards 1983; Ducharme and LePage 1986; Greenland and Mickey 1988; Geng 1989) as well as generalizations to partial collapsibility. The literature on graphical probability models distinguishes other types of collapsibility; see Frydenberg (1990), Whittaker (1990, Sect. 12.5) and Lauritzen (1996, Sect. 46.1) for examples. Both definitions given above are special cases of parametric collapsibility (Whittaker 1990).

Collapsibility in Regression Models

The above definition of strict collapsibility extends to regression contexts. Consider a generalized linear model (see ►Generalized Linear Models) for the regression of Y on three vectors w, x, z:

$$g[E(Y|w, x, z)] = \alpha + w\beta + x\gamma + z\delta.$$

This regression is said to be collapsible for β over z if $\beta^* = \beta$ in the regression omitting z,

$$g[E(Y|w, x)] = \alpha^* + w\beta^* + x\gamma^*$$

and is noncollapsible otherwise. Thus, if the regression is collapsible for β over Z and β is the parameter of interest, Z need not be measured to estimate β. If Z is measured, however, tests of $\beta^* = \beta$ can be constructed (Hausman 1978; Clogg et al. 1995).

The preceding definition generalizes the original contingency-table definition to arbitrary variables. There is a technical problem with the above regression definition,

however: If the first (full) model is correct, it is unlikely that the second (reduced) regression will follow the given form; that is, most families of regression models are not closed under deletion of Z. For example, suppose Y is Bernoulli with mean p and g is the logit link function $\ln[p/(1-p)]$, so that the full regression is first-order logistic. Then the reduced regression will not follow a first-order logistic model except in special cases. One way around this dilemma (and the fact that neither of the models is likely to be exactly correct) is to define the model parameters as the asymptotic means of the maximum-likelihood estimators. These means are well-defined and interpretable even if the models are not correct (White 1994).

If the full model is correct, $\delta = 0$ implies collapsibility for β and γ over Z. Nonetheless, if neither β nor δ is zero, marginal independence of the regressors does not ensure collapsibility for β over Z except when g is the identity or log link (Gail et al. 1984; Gail 1986). Conversely, collapsibility can occur even if the regressors are associated (Whittemore 1978; Greenland et al. 1999). Thus, it is not generally correct to equate collapsibility over Z with simple independence conditions, although useful results are available for the important special cases of linear, log-linear, and logistic models (e.g., see Gail 1986; Wermuth 1987, 1989; Robinson and Jewell 1991; Geng 1992; Guo and Geng 1995).

Confounding Versus Noncollapsibility

Much of the statistics literature does not distinguish between the concept of confounding as a bias in effect estimation and the concept of noncollapsibility; for example, Becher (1992) defines confounding as $\beta^* \neq \beta$ in the regression models given above, in which case the elements of Z are called confounders. Similarly, Guo and Geng (1995) define Z to be a nonconfounder if $\beta^* = \beta$. Nonetheless, confounding as defined in the causal-modeling literature (*See* ►Confounding) may occur with or without noncollapsibility, and noncollapsibility may occur with or without confounding; see Greenland (1987, 1996) and Greenland et al. (1999) for examples. Mathematically identical conclusions have been reached by other authors, albeit with different terminology in which noncollapsibility is called "bias" and confounding corresponds to "covariate imbalance" (Gail 1986; Hauck et al. 1991).

About the Author

For biography *see* the entry ►Confounding and Confounder Control.

Cross References

►Confounding and Confounder Control
►Simpson's Paradox

References and Further Reading

Asmussen S, Edwards D (1983) Collapsibility and response variables in contingency tables. Biometrika 70:567–578

Becher H (1992) The concept of residual confounding in regression models and some applications. Stat Med 11:1747–1758

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge

Clogg CC, Petkova E, Haritou A (1995) Statistical methods for comparing regression coefficients between models (with discussion). Am J Sociol 100:1261–1305

Cohen MR, Nagel E (1934) An introduction to logic and the scientific method. Harcourt Brace, New York

Ducharme GR, LePage Y (1986) Testing collapsibility in contingency tables. J R Stat Soc Ser B 48:197–205

Frydenberg M (1990) Marginalization and collapsibility in graphical statistical models. Ann Stat 18:790–805

Gail MH (1986) Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In: Moolgavkar SH, Prentice RL (eds) Modern statistical methods in chronic disease epidemiology. Wiley, New York, pp 3–18

Gail MH, Wieand S, Piantadosi S (1984) Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika 71:431–444

Geng Z (1989) Algorithm AS 299. Decomposability and collapsibility for log-linear models. J R Stat Soc Ser C 38:189–197

Geng Z (1992) Collapsibility of relative risk in contingency tables with a response variable. J R Stat Soc Ser B 54:585–593

Greenland S (1987) Interpretation and choice of effect measures in epidemiologic analyses. Am J Epidemiol 125:761–768

Greenland S (1996) Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology 7:498–501

Greenland S, Mickey RM (1988) Closed-form and dually consistent methods for inference on collapsibility in $2 \times 2 \times K$ and $2 \times J \times K$ tables. J R Stat Soc Ser C 37:335–343

Greenland S, Robins J, Pearl J (1999) Confounding and collapsibility in causal inference. Stat Sci 14:29–46

Guo J, Geng Z (1995) Collapsibility of logistic regression coefficients. J R Stat Soc Ser B 57:263–267

Hauck WW, Neuhas JM, Kalbfleisch JD, Anderson S (1991) A consequence of omitted covariates when estimating odds ratios. J Clin Epidemiol 44:77–81

Hausman J (1978) Specification tests in econometrics. Econometrica 46:1251–1271

Lauritzen SL (1996) Graphical models. Clarendon, Oxford

Neuhaus JM, Kalbfleisch JD, Hauck WW (1991) A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. Int Stat Rev 59:25–35

Robinson LD, Jewell NP (1991) Some surprising results about covariate adjustment in logistic regression. Int Stat Rev 59:227–240

Simpson EH (1951) The interpretation of interaction in contingency tables. J R Stat Soc Ser B 13:238–241

Wermuth N (1987) Parametric collapsibility and lack of moderating effects in contingency tables with a dichotomous response variable. J R Stat Soc Ser B 49:353–364

Wermuth N (1989) Moderating effects of subgroups in linear models. Biometrika 76:81–92
White HA (1994) Estimation, inference, and specification analysis. Cambridge University Press, New York
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, New York
Whittemore AS (1978) Collapsing multidimensional contingency tables. J R Stat Soc Ser B 40:328–340
Yule GU (1903) Notes on the theory of association of attributes in statistics. Biometrika 2:121–134

Comparability of Statistics

Peter Hackl
Professor
Vienna University of Business and Economics, Vienna, Austria

Comparison is the cognitive function that is basis for any measuring process and a frequent activity in everyday human life. Nowadays, comparisons of statistics over time and, even more demanding, cross-national and multilateral comparisons are a central element of economic and social analyzes. For politics and administration, for business and media, and for each citizen, comparisons are means to understand and assess the political, economic, and social processes of this world. This situation raises questions: Under which conditions are statistics comparable? Under which conditions is a comparison valid and leads to reliable results? What conclusions may be drawn and what are the risks implied by such conclusions?

Comparability: Definition and Assessment

Rathsmann-Sponsel and Sponsel (2001) describe "comparison" as a 7-digit relation $f(P, S, Z, K, V, A, B)$; P represents the comparing person, S and Z describe the comparing situation and the purpose of comparison, respectively; vector K stands for a number of criteria, V for a number of procedures, and A and B represent the characteristics of the objects to be compared, respectively. This rather formal view of psychologists indicates the complexity of the interaction between the individual and the objects to be compared. More visible becomes this complexity when the definition is applied to real situations, e.g., comparing the employment rates of two countries.

The employment rate is the number of persons aged 15–64 in employment as the share of the total population of the same age group. Obviously, the definition of "being in employment" and the exact meaning of "persons aged 15–64" are crucial for the value that is obtained for the employment rate. In addition, the statistical value is affected by the sampling design and other aspects of the data collection.

In general, statistics are based on concepts and definitions, and the value of a statistic is the result of a complex measurement process; *comparability is affected* by all these factors and, consequently, a wide range of facts must be considered. Moreover, the relevance of these facts depends on the purpose of comparison, the comparing situation, and other aspects of the comparison process. E.g., if the comparison of the employment rates of two countries is the basis for a decision on the allocation of subsidies for structural development, comparability is a more serious issue than in the case where the result of the comparisons does not have such consequences. These – and many other – characteristics must be taken into consideration when assessing differences between two employment rates.

Assessment of comparability has to take into account the multi-dimensionality of the conditions for comparability. Many aspects to be considered are qualitative, so that the corresponding dimensions cannot be measured on a metric scale. Moreover, important characteristics of the statistical products or the underlying measurement processes are often not available or uncertain.

Hence, in general, it is not feasible to give a comprehensive picture of comparability by means of a few metric measures. Alternatives are

- An indicator in form of a number between zero and one that indicates the degree of comparability, a one indicating perfect comparability.
- A rating of comparisons on an ordinal rating scale with a small number of points, a high value representing good comparability.

An example for a rating scale is the three point scale that is used for the "Overall assessment of accuracy and comparability" of indicators – such as the employment rate – within the Eurostat Quality Profiles; see Jouhette and Sproge (2005). This overall assessment is rated from "A" to "C". Grad "A" indicates that

- Data is collected from reliable sources applying high standards with regard to methodology/accuracy and is well documented in line with Eurostat metadata standard.
- The underlying data is collected on the basis of a common methodology for the European Union and,

where applicable, data for US and Japan can be considered comparable; major differences being assessed and documented.
- Data are comparable over time; impact of procedural or conceptual changes being documented.

This example illustrates:

- That the rating process reduces a high-dimensional information to a single digit.
- Where the characteristics of the statistics to be compared, the underlying measurement processes, and also conditions of the comparison process are crucial input elements for the rating of comparability.
- That the rating outcome has only the character of a label which the user might trust but which only reflects – perhaps vaguely – the result of a complex and subjective assessment process.
- That the rating outcome may miss to give appropriate weight to aspects that are important for a certain user.

A rating of the comparability on an ordinal rating scale has the advantage that it allows an easy communication about comparability.

The professional assessment of comparability requires:

- An adequate *competence of the scrutinizer*
- A careful *documentation of all characteristics* of the statistical products that are relevant for assessing the comparability

Generally, the scrutinizer will be different from the producer of a statistical product. This certainly will be the case in respect of cross-national comparisons. The producer has to provide a comprehensive set of metadata that documents all characteristics which are relevant for assessing the comparability. The outcome of this exercise might be an indicator of the types that are described above.

For the non-expert user, the assessment of the comparability of statistical products is hardly possible even when a well-designed set of all metadata is available that are relevant for assessing the comparability. Most users of the statistical products will have to rely on the professional assessment of the experts.

Comparability in Official Statistics

The integration of societies and the globalization of economies have the consequence that not only comparisons over time but especially cross-regional comparisons of statistical products are of increasing interest and importance. Political planning and decisions of supranational bodies need information that encompasses all involved nations. Multi-national enterprises and globally acting companies face the same problem.

Of even higher relevance for the need of comparability is the fact that statistical products are more and more used for *operational purposes*. Within the European Union, the process of integration of the member states into a community of states requires political measures in many areas. National statistical indicators are the basis for allocating a part of the common budget to member states, for administrating the regional and structural funds, for assessing the national performances with respect to the pact for stability and growth, and for various other purposes. It is in particular the European version of the System of National Accounts (SNA) ESA that plays such an operational role in various administrative processes of the European Union. The Millennium Development Goals and the Kyoto Protocol are other examples for the use of statistical indicators in defining political aims and assessing the corresponding progress. In all these cases, the comparability of the relevant statistical products is a core issue.

In the cross-national context, the responsibility for harmonizing cross-national concepts, definitions, and methodological aspects must be assigned to an authority with *supra-national competence*. Organizations like the UN, OECD, and Eurostat are engaged in the compilation of standards and the editing of recommendations, guidelines, handbooks, and training manuals, important means to harmonize statistical products and improve their comparability. Examples of standards are the Statistical Classifications of Economic Activities (ISIC) and the International Statistical Classification of Diseases (ICD). Principles and Recommendations for Population and Housing Censuses adopted by the Statistical Commission of the UN is an example for a standard methodology. Examples of standards on the European level are the NACE and CPA.

Within the European Union, standards and methods are laid down in regulations which are *legally binding* for the member states. E.g., the ESA 95 was approved as a Council Regulation in June 1996 and stipulates the member states to apply the SNA in a very concrete form. In working groups, experts from the member states are dealing with the preparation and implementation of such regulations; the harmonization of national statistical products is a central concern of these activities.

The important role that is attributed to statistical comparability within the ESS is stressed by the fact that the European Statistics Code of Practice (2005) contains Coherence and Comparability as one of its 15 principles. The corresponding indicators refer mainly to national aspects but also to the European dimension.

To assess the comparability of statistical products, national reports are essential that provide *metadata* for all related aspects of the statistical product. Standard formats for the documentation of metadata have been suggested by the International Monetary Fund in form of the General Data Dissemination Standard (GDDS) and the Special Data Dissemination Standard (SDDS).

It should be mentioned that standardizing concepts, definitions, and methods also has unfavorable effects; *comparability has a price*. An important means for improving harmonization are standards; however, they are never perfect and tend to get outdated over time. In particular the adaptation of standards to methodological progress might be a time-consuming task. Generally, standardization reduces flexibility and makes adaptations to new developments, especially of methodological alternatives, more difficult. This is especially true if standards are put into the form of regulations. It is even truer if such standards are implemented in order to ease the use for operational purposes, as it is the case in the ESS.

Conclusions

Lack of comparability may lead to erroneous results when statistical products are compared. The need for cross-national comparability is even more pronounced if statistical results are used for operational purposes as it is the case, e.g., in the European Union. Hence, comparability is an important quality aspect of statistical products. It is affected by the involved concepts and definitions, the measurement processes, and comparability may also depend on conditions of the comparison. The producer of a statistical product has to care that the conditions of comparability are fulfilled to the highest extent possible. In the cross-national context, international organizations like ►Eurostat are fostering the compilation of standards for concepts and definitions and of principles and standards for methods and processes in order to harmonize statistical products and improve their cross-national comparability.

For the assessment of comparability, a wide range of information is needed, as many aspects of the statistics to be compared but also of the purpose and conditions of the comparison have to be taken into account. No general rules are available that guarantee a valid assessment of comparability; only experts with profound knowledge can be expected to give a reliable assessment. For such an assessment, metadata which document all relevant characteristics are essential and have to be provided by the producer of the statistical product. For cross-national purposes, organizations like Eurostat have to care that the relevant metadata are provided by the respective producers. The user, e.g., the consumer of an economic or social analysis, has to trust that the analysts and experts made use of the involved statistics responsively.

About the Author

Dr. Peter Hackl was born 1942 in Linz, Austria. He is a Professor (since 1981) at the Department of Statistics, Vienna University of Business and Economics (Wirtschaftsuniversität); Head of the Division of Economic Statistics (since 1991). During the academic year 1988/1989 and during the summer term 1992, he was visiting professor at the University of Iowa, Iowa City. He was President of the Austrian Statistical Society (Österreichische Statistische Gesellschaft), (1995–2000). He was Vice-Dean of Studies (1995–2000) at the Wirtschaftsuniversität, member and deputy chairman (1999–2004) of the Statistikrat, the advisory board of Statistics Austria, the Austrian National Statistical Office; Chairman of the Committee for Quality Assurance. He was Director General of Statistics Austria (2005–2009). Professor Hackl is Elected member of the International Statistical Institute (since 1981). He has authored/coauthored about 100 articles in refereed journals in statistical theory and methods and applications of statistics. He was also editor of two books and has authored/coauthored 5 books. He was awarded an Honorary Doctorate, Stockholm University of Economics (1996).

Cross References

►Economic Statistics
►Eurostat
►National Account Statistics

References and Further Reading

European Commission (2005) Communication from the Commission to the European Parliament and to the Council on the independence, integrity and accountability of the national and Community statistical authorities, Recommendation on the independence, integrity and accountability of the national and Community statistical authorities, COM (2005) 217

European Statistics Code of Practice (2005) http://epp.eurostat.ec.europa.eu/portall/page/portal/quality/documents/code_practice.pdf

Eurostat (2009) ESS handbook for quality reports. European Communities, Luxembourg

Jouhette S, Sproge L (2005) Quality profiles for structural indicators EU LFS based indicators. In: 29th CEIES seminar, Expert meeting statistics "Structural indicators", Luxembourg, European Communities, pp 107–129

Rathsmann-Sponsel I, Sponsel R (2001) Allgemeine Theorie und Praxis des Vergleichens und der Vergleichbarkeit. Internet

Publikation für allgemeine und integrative Psychotherapie, www.sgipt.org/wisms/verglbk0.htm
Nederpelt V, Peter WM (2009) Checklist quality of statistical output. Statistics Netherlands, Den Haag/Heerlen

Complier-Average Causal Effect (CACE) Estimation

Graham Dunn
Professor of Biomedical Statistics and Head of the Health Methodology Research Group
University of Manchester, Manchester, UK

Imagine a simple randomized controlled trial evaluating a psychotherapeutic intervention for depression. Participants are randomized to one of two treatment groups. The first (the control condition) comprises treatment and management as usual (TAU). Participants in the second group are offered a course of individual cognitive behavior therapy (CBT) *in addition* to TAU. The outcome of the trial is evaluated by assessing the severity of depression in all of the participants 6 months after ▶randomization. For simplicity, we assume there are no missing outcomes. We find a difference between the average outcomes for the two groups to be about four points on the depression rating scale, a difference that is of only borderline clinical significance. This difference of four points is the well-known intention-to-treat (ITT) effect – it is the estimated effect treatment *allocation* (i.e., offering the treatment). This we will call ITT_{ALL}.

Participants in the control (TAU) condition did not get any access to CBT but we now discover that only about half of those offered psychotherapy actually took up the offer. Only 50% of the treatment group actually received CBT. So, it might be reasonable to now ask "What was the effect of receiving CBT?" or "What was the treatment effect in those who complied with the trial protocol (i.e., treatment allocation)?" Traditionally, trialists have been tempted to carry out what is called a "Per Protocol" analysis. This involves dropping the non-compliers from the treatment arm and comparing the average outcomes in the compliers with the average outcome in all of the controls. But this is not comparing like with like. There are likely to be selection effects (confounding) – those with a better (or worse) treatment-free prognosis might be more likely to turn up for their psychotherapy. The same criticism also applies to the abandonment of randomization altogether and comparing the average outcomes in those who received treatment with those who did not (a mixture of controls and non-compliers) in a so-called "As treated" analysis.

To obtain a valid estimate of the *receipt* of treatment we need to be able to compare the average of the outcomes in those who received CBT with the average of the outcomes of the control participants who *would have received* CBT had they been allocated to the treatment group. This is the Complier-Average Causal Effect (CACE) of treatment. How do we do this? The simplest approach is based on the realization that the ITT effect is attenuated estimate of the CACE, and that the amount of attenuation is simply the proportion of compliers (or would-be compliers) in the trial. The proportion of compliers (P_C) is simply estimated by the proportion of those allocated to the treatment group who actually receive CBT. We postulate that we have two (possibly hidden) classes of participant: Compliers and Non-compliers. Non-compliers receive no CBT irrespective of their treatment allocation. The intention-to-treat effects in the Compliers and Non-compliers are ITT_C (≡ CACE) and ITT_{NC}, respectively. It should be clear to the reader that $ITT_{ALL} = P_C ITT_C$ and $(1 - P_C)ITT_{NC}$.

To make use of this simple relationship, let's now assume that treatment allocation in the Non-compliers has no impact on their outcome (i.e., does not affect the severity of their depression). This assumption is often referred to as an exclusion restriction. It follows that $ITT_{ALL} = P_C ITT_C$ and that

$$CACE = ITT_C = ITT_{ALL}/P_C$$

So with 50% compliance, and estimated overall ITT effect of 4 units on the depression scale, the CACE estimate is 8 units – a result with much more promise to our clinicians. To get a standard error estimate we might apply a simple bootstrap (see ▶Bootstrap Methods). Note that CACE estimation is not a means of conjuring up a treatment effect from nowhere – if the overall ITT effect is zero so will the CACE be. If the overall ITT effect is not statistically-significant, the CACE will not be statistically-significant.

One last point: in a conventional treatment trial aiming to demonstrate efficacy, the ITT estimate will be conservative, but in a trial designed to demonstrate equivalence (or non-inferiority) it is the CACE estimate that will be the choice of the cautious analyst (we do not wish to confuse attenuation arising from non-compliance with differences in efficacy).

Here we have illustrated the ideas with the simplest of examples. And here we have also made the derivation of the CACE estimate as simple as possible without any detailed reference to required assumptions. An analogous procedure was first used by Bloom (1984) but its

formal properties were derived and compared with instrumental variable estimators of treatment effects by Angrist et al. (1996). Non-compliance usually has an implication for missing data – those that do not comply (or would-be Non-compliers) with their allocated treatment are also those who are less likely to turn up to have their outcome assessed. The links between CACE estimation and missing data models (assuming latent ignorability) are discussed by Frangakis and Rubin (1999). Generalizations of CACE methodology to estimation of treatment effects through the use of Principal Stratification have also been introduced by Frangakis and Rubin (2002).

About the Author

For biography *see* the entry ►Psychiatry, Statistics in

Cross References

►Instrumental Variables
►Principles Underlying Econometric Estimators for Identifying Causal Effects

References and Further Reading

Angrist JD, Imbens GW, Rubin DB (1996) Indentification of causal effects using instrumental variables (with discussion). J Am Stat Assoc 91:444–472

Bloom HS (1984) Accounting for no-shows in experimental evaluation designs. Evaluation Rev 8:225–246

Frangakis CE, Rubin DB (1999) Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. Biometrika 86:365–379

Frangakis CE, Rubin DB (2002) Principal stratification in causal inference. Biometrics 58:21–29

Components of Statistics

Clive William John Granger†
Professor Emeritus
University of California-San Diego, San Diego, CA, USA

The two obvious subdivisions of statistics are: (a) Theoretical Statistics and (b) Practical Statistics.

1. The theoretical side is largely based on a mathematical development of probability theory (*see* ►Probability Theory: An Outline) applicable to data, particularly the asymptotic properties of estimates (*see* ►Properties of Estimators) which lead to powerful theorems such as the ►Central Limit Theorem. The aim is to put many practical approaches to data analysis (*see also* ►Categorical Data Analysis; ►Multivariate Data Analysis: An Overview; ►Exploratory Data Analysis; ►Functional Data Analysis) on a sound theoretical foundation and to develop theorems about the properties of these approaches. The theories are usually based on a number of assumptions that may or may not hold in practice.
2. Practical statistics considers the analysis of data, how the data can be summarized in useful fashions, and how relationships between sets of data from different variables can be described and interpreted. The amount and the quality of the data (*see* ►Data Quality) that is available are essential features in this area. On occasions data may be badly constructed or terms may be missing which makes analysis more complicated.

Descriptive statistics include means, variances, histograms, correlations, and estimates of quantiles, for example. There are now various types of statistics depending on the area of application. General statistics arose from considerations of gambling (*see* ►Statistics and Gambling), agriculture (*see* ►Agriculture, Statistics in; ►Analysis of Multivariate Agricultural Data), and health topics (*see* ►Medical research, Statistics in; ►Medical Statistics) but eventually a number of specialized areas arose when it was realized that these areas contained special types of data which required their own methods of analysis. Examples are:

1. Biometrics (*see* ►Biostatistics), from biological data which required different forms of measurement and associated tests.
2. ►Econometrics, for which ►Variables may or may not be related with a time gap; data can be in the form of ►Time Series (particularly in economies and finance) or in large panels (*see* ►Panel Data) in various parts of economics. The techniques developed over a wide range and the ideas have spread into other parts of statistics.
3. Psychometrics, where methods are required for the analysis of results from very specific types of experiments used in the area of psychology (*see* ►Psychology, Statistics in).

Other major areas of application such as engineering, marketing, and meteorology generally use techniques derived from methods in the areas mentioned above, but all have developed some area-specific methods.

About the Author

In 2003 Professor Granger was awarded the Nobel Memorial Prize in Economic Sciences (with Professor Robert

Engle) for methods of analyzing economic time series with common trends (cointegration). Granger was knighted in 2005.

Professor Granger had sent his contributed entry on June 2 2008, excusing himself for not writing a bit longer, "Lexicon" kind of paper: "I have never written anything for a 'Lexicon' before and so have failed in my attempt to be helpful, but I do attach a page or so. I wish you good luck with your effort." We are immensely thankful for his unselfish contribution to the prosperity of this project.

Cross References

►Statistics: An Overview
►Statistics: Nelder's view

Composite Indicators

Jože Rovan
Associate Professor, Faculty of Economics
University of Ljubljana, Ljubljana, Slovenia

Definition: A composite indicator is formed when individual indicators are compiled into a single index, on the basis of an underlying model of the multidimensional concept that is being measured (OECD, Glossary of Statistical Terms).

Multidimensional concepts like welfare, well-being, human development, environmental sustainability, industrial competitiveness, etc., cannot be adequately represented by individual indicators. For that reason, composite indicators are becoming increasingly acknowledged as a tool for summarizing complex and multidimensional issues.

Composite indicators primarily arise in the following areas: economy, society, globalization, environment, innovation, and technology. A comprehensive list of indicators can be found at the following address: http:// composite-indicators.jrc.ec.europa.eu/FAQ.htm#List_of_Composite_Indicators_

The *Handbook on Constructing Composite Indicators: Methodology and User Guide* (OECD 2008) recommends a ten-step process of constructing composite indicators:

- **Theoretical framework** is the starting point of the process of constructing composite indicators, defining individual indicators (e.g., variables) and their appropriate weights, reflecting the structure of the investigated multidimensional concept. This step is crucial in construction process because it has the greatest impact on the relevance of the indicator of the investigated phenomena. For that reason, the constructors team should include, besides the statisticians, who play the major role, the experts and stakeholders from the topic of the composite indicator.

Due to the fact that many new multidimensional concepts do not have a generally agreed theoretical framework, transparency is essential in constructing credible indicators.

- **Data selection** should acquire analytically sound relevant indicators, having in mind their avaliability (country coverage, time, appropriate scale of measurement, etc.). Engagement of experts and stakeholders is recommended.
- **Imputation of missing data** (see ►Imputation) provides a complete dataset (single or. multiple). Inspection of presence of ►outliers in the dataset should not be omitted.
- **Multivariate analysis** reveals the structure of the considered dataset from two aspects: (a) units and (b) available individual indicators, using appropriate multivariate methods, e.g., ►principal component analysis, factor analysis (see ►Factor Analysis and Latent Variable Modelling), Cronbach coefficient alpha, cluster analysis (see ►Cluster Analysis: An Introduction), ►correspondence analysis, etc. These methods are able to reveal internally homogenous groups of countries or groups of indicators and interpret the results.
- **Normalization procedures** are used to achieve comparability of variables of the considered dataset, taking into account theoretical framework and the properties of the data. The robustness of normalization methods against possible ►outliers must be considered.
- **Weighting and aggregation** should take into account the theoretical framework and the properties of the data. The most frequently used aggregation form is *a weighted linear aggregation rule* applied to a set of variables (OECD 2003). Weights shoud reflect the relative importance of individual indicators in a construction of the particular composite indicator.
- **Uncertainty and ►sensitivity analysis** are necessary to evaluate robustness of composite indicators and to improve transparency, having in mind selection of indicators, data quality, imputation of missing data, data normalization, weighting, aggregation methods, etc.
- **Back to the original data**, to (a) reconsider the relationships between composite indicator and the original

data set and to identify the most influential indicators and (b) compare profiled performance of the considered units to reveal what is driving the composite indicator results, and in particular whether the composite indicator is overly dominated by a small number of indicators.

- **Links to other indicators** identify the relationships between the composite indicator (or its dimensions) and other individual or composite indicators.
- **Visualization of results** should attract audience, presenting composite indicators in a clear and accurate way.

Following the above-mentioned guidelines, the constructors of composite indicators should never forget *that composite indicators should never be seen as a goal per se. They should be seen, instead, as a starting point for initiating discussion and attracting public interest and concern* (Nardo et al. 2005).

However, there is now general agreement about the usefulness of composite indicators: There is a strong belief among the constructors of composite indicators that such summary measures are meaningful and that they can capture the main characteristic of the investigated phenomena. On the other side, there is a scepticism among the critics of this approach, who believe that there is no need to go beyond an appropriate set of individual indicators. Their criticism is focused on the "arbitrary nature of the weighting process" (Sharpe 2004) in construction of the composite indicators.

Cross References

►Aggregation Schemes
►Imputation
►Multiple Imputation
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Scales of Measurement
►Sensitivity Analysis

References and Further Reading

An information server on composite indicators and ranking systems (methods, case studies, events) http://composite-indicators.jrc.ec.europa.eu/FAQ.htm#List_of_Composite_Indicators

Freudenberg M (2003) Composite indicators of country performance: a critical assessment, OECD science, technology and industry working papers, OECD Publishing, 2003/16

OECD, European Commission, Joint Research Centre (2008) Handbook on constructing composite indicators: methodology and user guide. OECD Publishing

OECD, Glossary of statistical terms (http://stats.oecd.org/glossary/index.htm)

OECD (2003) Composite indicators of country performance: a critical assessment, DST/IND(2003)5, Paris

Munda G, Nardo M (2005) Constructing consistent composite indicators: the issue of weights, EUR 21834 EN. Joint Research Centre, Ispra

Nardo M, Saisana M, Saltelli A, Tarantola S (2005) Tools for composite indicators building. european commission, EUR 21682 EN. Joint Research Centre, Ispra, Italy

Saltelli A (2007) Composite indicators between analysis and advocacy. Soc Indic Res 81:65–77

Sharpe A (2004) Literature review of frameworks for macro-indicators. Centre for the Study of Living Standards, Ottawa, Canada

Computational Statistics

Colin Rose
Director
Theoretical Research Institute, Sydney, NSW, Australia

What Is Computational Statistics?

We define *computational statistics* to be: ... 'statistical methods/results that are enabled by using computational methods'. Having set forth a definition, it should be stressed, first, that names such as *computational statistics* and *statistical computing* are essentially semantic constructs that do not have any absolute or rigourous structure within the profession; second, that there are any number of competing definitions on offer. Some are unsatisfactory because they focus purely on data or graphical methods and exclude symbolic/exact methods; others are unsatisfactory because they place undue emphasis on 'computationally-intensive methods' or brute force, almost as if to exclude well-written efficient and elegant algorithms that might be computationally quite simple. Sometimes, the difficulty is not in the execution of an algorithm, but in writing the algorithm itself.

Computational statistics can enable one:

- To work with arbitrary functional forms/distributions, rather than being restricted to traditional known textbook distributions.
- To simulate distributional properties of estimators and test statistics, even if closed-form solutions do not exist (*computational inference* rather than *asymptotic inference*).
- To compare statistical methods under different alternatives.
- To solve problems numerically, even if closed-form solutions are not possible or cannot be derived.
- To derive symbolic solutions to probability, moments, and distributional problems that may never have been solved before, and to do so essentially in real-time.

- To explore multiple different models, rather than just one model.
- To explore potentially good or bad ideas via simulation in just a few seconds.
- To choose methods that are theoretically appropriate, rather than because they are mathematically tractable.
- To check symbolic/exact solutions using numerical methods.
- To bring to life theoretical models that previously were too complicated to evaluate . . .

Journals and Societies

Important journals in the field include:

- Combinatorics, Probability & Computing
- Communications in Statistics – Simulation and Computation
- Computational Statistics
- Computational Statistics and Data Analysis
- Journal of Computational and Graphical Statistics
- Journal of the Japanese Society of Computational Statistics
- Journal of Statistical Computation and Simulation
- Journal of Statistical Software
- SIAM Journal on Scientific Computing
- Statistics and Computing

Societies include: the International Association for Statistical Computing (IASC – a subsection of the ISI), the American Statistical Association (Statistical Computing Section), the Royal Statistical Society (Statistical Computing Section), and the Japanese Society of Computational Statistics (JSCS) . . .

Computational statistics consists of three main areas, namely numerical, graphical and symbolic methods . . .

Numerical Methods

The numerical approach is discussed in texts such as Gentle (2009), Givens and Hoeting (2005), and Martinez and Martinez (2007); for Bayesian methods, see Bolstad (2009). Numerical methods include: Monte Carlo studies to explore asymptotic properties or finite sample properties, pseudo-random number generation and sampling, parametric density estimation, non-parametric density estimation, ▶bootstrap methods, statistical approaches to software errors, information retrieval, statistics of databases, high-dimensional data, temporal and spatial modeling, ▶data mining, model mining, statistical learning, computational learning theory and optimisation etc. . . . While optimisation itself is an absolutely essential tool in the field, it is very much a field in its own right.

Graphical Methods

Graphical methods are primarily concerned with either (a) viewing theoretical models and/or (b) viewing data/fitted models.

In the case of *theoretical* models, one typically seeks to provide understanding by viewing one, two or three variables, or indeed even four dimensions (using 3-dimensional plots animated over time, translucent graphics etc.).

Visualizing *data* is essential to data analysis and assessing fit; see, for instance, Chen et al. (2008). Special interest topics include smoothing techniques, kernel density estimation, multi-dimensional data visualization, clustering, exploratory data analysis, and a huge range of special statistical plot types. Modern computing power makes handling and interacting with large data sets with millions of values feasible . . . including live interactive manipulations.

Symbolic/Exact Methods

The 21st century has brought with it a conceptually entirely new methodology: symbolic/exact methods. Recent texts applying the symbolic framework include Andrews and Stafford (2000), Rose and Smith (2002), and Drew et al. (2008).

Traditional 20th century computer packages are based on numerical methods that tend to be designed much like a cookbook. That is, they consist of hundreds or even thousands of numerical recipes designed for specific cases. One function is written for one aspect of the Normal distribution, another for the LogNormal, etc. This works very well provided one stays within the constraints of the known common distributions, but unfortunately, it breaks down as soon as one moves outside the catered framework. It might work perfectly for random variable X, but not for X^2, nor $\exp(X)$, nor mixtures, nor truncations, nor reflections, nor folding, nor censoring, nor products, nor sums, nor . . .

By contrast, symbolic/exact methods are built on top of computer algebra systems . . . programs such as *Mathematica* and Maple that understand algebra and mathematics. Accordingly, symbolic algorithms can provide exact general solutions . . . not just for specific distributions/models. Symbolic computational statistical packages include mathStatica (2002–2010, based on top of *Mathematica*) and APPL (based on top of Maple).

Symbolic methods include: automated expectations for arbitrary distributions, probability, combinatorial probability, transformations of random variables, products of random variables, sums and differences of random variables, generating functions, inversion theorems, maxima/minima of random variables, symbolic and numerical maximum likelihood estimation (using exact methods),

curve fitting (using exact methods), non-parametric kernel density estimation (for arbitrary kernels), moment conversion formulae, component-mix and parameter-mix distributions, copulae, pseudo-random number generation for arbitrary distributions, decision theory, asymptotic expansions, ▶order statistics (for identical and non-identical parents), unbiased estimators (h-statistics, k-statistics, polykays), moments of moments, etc.

The Changing Notion of What is Computational Statistics

Just 10 or 20 years ago, it was quite common for people working in computational statistics to write up their own code for almost everything they did. For example, the *Handbook of Statistics 9: Computational Statistics* (see Rao 1993) starts out Chapter 1 by describing algorithms for sorting data. Today, of course, one would expect to find sorting functionality built into any software package one uses . . . indeed even into a word processor. And, of course, the 'bar' keeps on moving and evolving. Even in recent texts such as Gentle (2009), about half of the text (almost all of Part 1) is devoted to computing techniques such as fixed- and floating-point, numerical quadrature, numerical linear algebra, solving non-linear equations, optimisation etc., . . . techniques that Gentle et al. (2004, p. 5) call "*statistical computing*" but which are really just *computing*. Such methods lie firmly within the domain of computational science and/or computational mathematics . . . they are now built into any decent modern statistical/mathematical software package . . . they take years of work to develop into a decent modern product, and they require tens of thousands of lines of code to be done properly . . . all of which means that it is extremely unlikely that any individual would write their own in today's world. Today, one does not tend to build an airplane simply in order to take a flight. And yet many current texts are still firmly based in the older world of 'roll your own', devoting substantial space to routines that are (a) general mathematical tools such as numerical optimisation and (b) which are now standard functionality in modern packages used for computational statistics. While it is, of course, valuable to understand how such methods work (in particular so that one is aware of their limitations), and while such tools are absolutely imperative to carrying out the discipline of computational statistics (indeed, as a computer itself is) – these tools are now general mathematical tools and the days of building one's own are essentially long gone.

Future Directions

It is both interesting and tempting to suggest likely future directions.

(a) *Software packages*: At the present time, the computational statistics software market is catered for from two polar extremes. On the one hand, there are major general mathematical/computational languages such as *Mathematica* and Maple which provide best of breed general computational/numerical/graphical tools, and hundreds of high-level functional programming constructs to expand on same, but they are less than comprehensive in field-specific functionality. It seems likely such packages will further evolve by developing and growing tentacles into specific fields (such as statistics, combinatorics, finance, econometrics, biometrics etc.). At the other extreme, there exist narrow field-specific packages such as S-Plus, Gauss and R which provide considerable depth in field-specific functionality; in order to grow, these packages will likely need to broaden out to develop more general methods/general mathematical functions, up to the standard offered by the major packages. The software industry is nascent and evolving, and it will be interesting to see if the long-run equilibrium allows for both extremes to co-exist. Perhaps, all that is required is for a critical number of users to be reached in order for each eco-system to become self-sustaining.

(b) *Methodology*: It seems likely that the field will see a continuing shift or growth from *statistical inference* to *structural inference*, . . . from *data mining* to *model mining*, . . . from *asymptotic inference* to *computational inference*.

(c) *Parallel computing*: Multicore processors have already become mainstream, while, at the same time, the growth in CPU speeds appears to be stagnating. It seems likely then that parallel computing will become vastly more important in evolving computational statistics into the future. Future computational statistical software may also take advantage of GPUs (graphical processing units), though it should be cautioned that the latter are constrained in serious statistical work by the extremely poor numerical precision of current GPUs.

(d) *Symbolic methods*: Symbolic methods are still somewhat in their infancy and show great promise as knowledge engines i.e., algorithms that can derive exact theoretical results for arbitrary random variables.

(e) *On knowledge and proof*: Symbolic algorithms can derive solutions to problems that have never been posed before – they place enormous technological power into the hands of end-users. Of course, it is possible (though rare) that an error may occur (say in integration, or by mis-entering a model). In a sense,

this is no different to traditional reference texts and journal papers which are also not infallible, and which are often surprisingly peppered with typographical or other errors.

In this regard, the computational approach offers both greater exposure to danger, as well as the tools to avoid it. The "danger" is that it has become extremely easy to generate output in real-time. The sheer scale and volume of calculation has magnified, so that the average user is more likely to encounter an error, just as someone who drives a lot is more likely to encounter an accident. *Proving* that the computer's output is actually correct can be very tricky, or impractical, or indeed impossible for the average practitioner to do, just as the very same practitioner will tend to accept a journal result at face value, without properly checking it, even if they could do so. The philosopher, Karl Popper, argued that the aim of science should not be to prove things, but to seek to refute them. Indeed, the advantage of the computational statistical approach is that it is often possible to check one's work using two completely different methods: both numerical and symbolic. Here, numerical methods take on a new role of checking symbolic results. One can throw in some numbers in place of symbolic parameters, and one can then check if the solution obtained using symbolic methods (the exact theoretical solution) matches the solution obtained using numerical methods (typically, ►numerical integration or Monte Carlo methods, see ►Monte Carlo Methods in Statistics). If the numerical and symbolic solutions do *not* match, there is an obvious problem and we can generally immediately reject the theoretical solution (*a la* Popper). On the other hand, if the two approaches match up, we still do not have a proof of correctness . . . all we have is just one point of agreement in parameter space. We can repeat and repeat and repeat the exercise with different parameter values . . . and as we do so, we effectively build up, not an absolute proof in the traditional sense, but, appropriately for the statistics profession, an ever increasing degree of confidence . . . effectively a proof by probabilistic induction . . . that the theoretical solution is indeed correct. This is an extremely valuable (though time-consuming) skill to develop, not only when working with computers, but equally with textbooks and journal papers.

About the Author

For biography *see* the entry ►Bivariate distributions.

Cross References

►Bootstrap Asymptotics
►Bootstrap Methods
►Data Mining
►Monte Carlo Methods in Statistics
►Nonparametric Density Estimation
►Non-Uniform Random Variate Generations
►Numerical Integration
►Numerical Methods for Stochastic Differential Equations
►R Language
►Statistical Software: An Overview
►Uniform Random Number Generators

References and Further Reading

Andrews DF, Stafford JEH (2000) Symbolic computation for statistical inference. Oxford University Press, New York
Bolstad WM (2009) Understanding computational Bayesian statistics. Wiley, USA
Chen C, Härdle W, Unwin A (2008) Handbook of data visualization. Springer, Berlin
Drew JH, Evans DL, Glen AG, Leemis LM (2008) Computational probability. Springer, New York
Gentle JE (2009) Computational statistics. Springer, New York
Gentle JE, Härdle W, Mori Y (eds) (2004) Handbook of computational statistics: concepts and methods. Springer, Berlin
Givens GH, Hoeting JA (2005) Computational statistics. Wiley, New Jersey
Martinez WL, Martinez AR (2007) Computational statistics handbook with MATLAB, 2nd edn. Chapman & Hall, New York
mathStatica (2002–2010), www.mathStatica.com
Rao CR (1993) Handbook of statistics 9: computational statistics. Elsevier, Amsterdam
Rose C, Smith MD (2002) Mathematical statistics with Mathematica. Springer, New York

Conditional Expectation and Probability

Takis Konstantopoulos
Professor
Heriot-Watt University, Edinburgh, UK

In its most elementary form, the conditional probability $P(A|B)$ of an event A given an event B is defined by

$$P(A|B) = \frac{B(A \cap B)}{P(B)},$$

provided that $P(B) \neq 0$. This is a well-motivated definition, compatible both with the frequency interpretation of probability as well as with elementary probability on countable spaces. An immediate consequence of the definition is ▶*Bayes' theorem*: if $A_1, A_2, \ldots, A_n$ are mutually disjoint events whose union has probability one, then $P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$.

Suppose now that X, Y are random variables taking values in finite sets. We define the conditional distribution of X given Y by

$$P(X = x|Y = y) = \begin{cases} \frac{P(X=x,Y=y)}{P(Y=y)}, & \text{if } P(Y = y) \neq 0 \\ 0, & \text{if } P(Y = y) = 0. \end{cases}$$

The latter choice, i.e., 0/0 interpreted as 0, is both physically motivated and mathematically desirable. The object $P(X = x|Y = y)$ is a probability in x (i.e., it sums up to 1 over x) and a function of y. If X takes values in a set of real numbers then we can define the conditional expectation of X given $Y = y$ by

$$E(X|Y = y) = \sum_x xP(X = x|Y = y), \tag{1}$$

where the summation extends over all possible values x of X. This is a function of y, say $h(y) = E(X|Y = y)$. We can then talk about the conditional expectation of X given Y as the *random variable* $h(Y)$ obtained by substituting y by the random variable Y in the argument of $h(y)$. From this definition the following important property of $E(X|Y)$ is easily derived:

$$E\big[(X - E(X|Y)) \cdot g(Y)\big] = 0, \tag{2}$$

for any random variable $g(Y)$ which is a (deterministic) function of Y.

One can easily generalise the above to countably-valued random variables. However, defining conditional probability and expectation for general random variables cannot be done in the previous naive manner. One can mimic the definitions for random variables possessing density but this has two drawbacks: first, it is not easy to rigorously reconcile with the previous definitions; second, it is not easy to generalize. Instead, we resort to an axiomatic definition of conditional expectation, stemming directly from the fundamental property (2). It can be easily verified that, in the earlier setup, there is only one function $h(y)$ satisfying (2) for all functions $g(y)$, and this $h(y)$ is defined by (1).

The last observation leads us to the following definition: Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a positive random variable (i.e., a measurable function $X : \Omega \to \mathbb{R}_+$). Let $\mathcal{G} \subset \mathcal{F}$ be another sigma-algebra. We say that $E(X|\mathcal{G})$ is the conditional expectation of X given $\mathcal{G}$ if (a) $E(X|\mathcal{G})$ is $\mathcal{G}$-measurable and (b) for all bounded $\mathcal{G}$-measurable random variables G, we have

$$E[XG] = E[E(X|\mathcal{G})G]. \tag{3}$$

Such an object exists and is *almost surely* unique. The latter means that if two random variables, H_1, H_2, say, satisfy $E[XG] = E[H_iG]$, $i = 1, 2$, for all G then $P(H_1 = H_2) = 1$. (Such H_i are called *versions* of the conditional expectation.) Existence is immediate by the ▶Radon–Nikodým theorem. Consider two measures on $(\Omega, \mathcal{G})$: the first one is P; the second one is $E[X\mathbb{1}_G]$, $G \in \mathcal{G}$ (where $\mathbb{1}_G$ is defined as 1 on G and 0 on $\Omega\backslash G$). When $P(G) = 0$ we have $E[X\mathbb{1}_G] = 0$ and therefore the second measure is *absolutely continuous* with respect to the first. The Radon–Nikodým theorem ensures that the derivative (density) of the second measure with respect to the first exists and that it satisfies (3). This observation and string of arguments is due to Kolmogorov (1933), and it is through this that modern Probability Theory was established as a mathematical discipline having a natural connection with Measure Theory.

Having defined $E[X|\mathcal{G}]$ for positive X we can define it for negative X by reversing signs and for general X via the formula $E[X|\mathcal{G}] = E[\max(X, 0)|\mathcal{G}] + E[\min(X, 0)|\mathcal{G}]$, provided that wither $E[\max(X, 0)] < \infty$ or $E[\min(X, 0)] > -\infty$.

Given then two random variables X, Y (the first of which is real-valued, but the second may take values in fairly arbitrary spaces (such as a space of functions), we can define $E[X|Y]$ as $E[X|\sigma(Y)]$ where $\sigma(Y)$ is the σ-algebra generated by Y. It can be seen that this is entirely compatible with the initial definition (1).

Passing on to conditional probability, observe that if A is an event, the expectation of $\mathbb{1}_A$ is precisely $P(A)$. By analogy, we define

$$P(A|\mathcal{G}) = E[\mathbb{1}_A|\mathcal{G}].$$

For each event $A \in \mathcal{F}$, this is a random variable, i.e., a measurable function of $\omega \in \Omega$ which is defined almost surely uniquely (see explanation after formula (3)). For a real-valued random variable X we define the conditional distribution function $P(X \leq x|\mathcal{G})$ as $E[\mathbb{1}_{X \leq x}|\mathcal{G}]$. We would like this to be a right-continuous non-decreasing function of x. Since, for each x, $P(X \leq x|\mathcal{G})$ is defined only almost surely, we need to show that we can, for each x, pick a version H_x of $P(X \leq x|\mathcal{G})$ in a way that the probability of the event $\{H_x \leq H_y \text{ if } x \leq y \text{ and } \lim_{\epsilon \downarrow 0} H_{x+\epsilon} = H_x\}$ is one. This can be done and $\{H_x\}_{x \in \mathbb{R}}$ is referred to as a *regular conditional distribution function* of X given $\mathcal{G}$. Informally (and in practice) it is denoted as $P(X \leq x|\mathcal{G})$. Regular conditional

probabilities exist not only for real random variables X but also for random elements X taking values in a Borel space Kallenberg (2002).

From this general viewpoint we can now go down again and verify everything we wish to have defined in an intuitive or informal manner. For instance, if (X, Y) is a pair of real-valued random variables having joint density $f(x,y)$ then, letting $f_Y(y) := \int_{\mathbb{R}} f(x,y)dx$, define

$$h(x|y) := \begin{cases} \frac{f(x,y)}{f_Y(y)}, & \text{if } f_Y(y) \neq 0 \\ 0, & \text{if } f_Y(y) = 0 \end{cases}.$$

Then the function $\int_{-\infty}^{x} h(s|Y)ds$ serves as a the regular conditional distribution of X given Y. We can also verify that $E(\max(X,0)|Y) = \int_0^\infty P(X > x|Y)dx$ (in the sense that the right-hand side is a version of the left) and several other elementary formulae.

It is important to mention an interpretation of $E(X|Y)$ as a projection. First recall the definition of projection in Euclidean space: Let x be a point (vector) in the space $\mathbb{R}^n$ and Π a hyperplane. We define the projection $\widehat{x}$ of x onto Π as the unique element of Π which has minimal distance from x. Equivalently, the angle between $x - \widehat{x}$ and any other vector g of Π must be 90°: this can be written as $\langle x - \widehat{x}, g\rangle = 0$, i.e., the standard inner product of $x - \widehat{x}$ and g is equal to 0. Next suppose that $E[X^2] < \infty$. Then it can be seen that

$$E[(X - E(X|\mathcal{G}))^2] = \min_G E[(X - G)^2],$$

where the minimum is taken over all $\mathcal{G}$-measurable random variables G with $E[G^2] < \infty$. The defining property (3) then says that the inner product between $X - E(X|\mathcal{G})$ and any G is zero, just as in Euclidean space. Keeping the geometric meaning in mind, we can devise (prove and interpret) several properties of the conditional expectation. We mention one below.

The *tower property*: If $\mathcal{G}_2 \subset \mathcal{G}_1$ are two sigma-algebras then

$$E[E(X|\mathcal{G}_1)|\mathcal{G}_2] = E[X|\mathcal{G}_2].$$

The geometric meaning is as follows: If Π_1 is a hyperplane (e.g., a plane in three dimensions) and Π_2 a hyperplane contained in Π_1 (e.g., a line on the plane) then we can find the projection onto Π_2 by first projecting onto Π_1 and then projecting the projection. The tower property holds for general random variables as long as conditional expectation can be defined, i.e., it does not require $E[X^2] < \infty$. Another interpretation of it is as follows: if $\mathcal{G}_1, \mathcal{G}_2$ represent states of knowledge (information, say) and $\mathcal{G}_1$ is wider than $\mathcal{G}_2$ (in the sense that $\mathcal{G}_2$ can be obtained from $\mathcal{G}_1$) then, in finding the conditional expectation of X given $\mathcal{G}_2$, the additional knowledge contained in $\mathcal{G}_1$ can be ignored. A particular form of this property is in the relation

$$E[E(X|\mathcal{G})] = E[X].$$

Another important property is that $E[GX|\mathcal{G}] = GE[X|\mathcal{G}]$ if G is $\mathcal{G}$-measurable. On the other hand, if Z is independent of (X, Y) then $E[X|Y,Z] = E[X|Y]$. For further properties, see Williams (1989). In particular, if X and Y are independent then $E[X|Y] = E[X]$, i.e., it is a constant.

For *normal* random variables, the geometric picture completely characterizes what we can do with them. Recall that a random variable X is centred normal if it has finite variance σ^2 and if for all constants a, b there is a constant c such that cX has the same distribution as $aX' + bX''$ where X', X'' are independent copies of X. It follows that $a^2 + b^2 = c^2$ and that X has density proportional to $e^{-x^2/2\sigma^2}$. We say that X is normal if $X - E[X]$ is centred normal. We say that a collection of random variables $\{X_t\}_{t\in T}$, with T being an arbitrary set, is (jointly) normal if for any $t_1, \ldots, t_n$, and any constants $c_1, \ldots, c_n$, the random variable $c_1X_{t_1} + \cdots + c_nX_{t_n}$ is normal. It follows that if $\{X, Y_1, \ldots, Y_k\}$ are jointly normal then $E[X|Y_1, \ldots, Y_k] = E[X|\sigma(Y_1, \ldots, Y_k)] = a_1Y_1 + \cdots + a_kY_k + b$ where the constants can be easily computed by (3). The *Kalman filter property* says that if $\{X, Y_1, Y_2\}$ are centred jointly normal such that Y_1 and Y_2 are independent then $E[X|Y_1, Y_2] = E[X|Y_1] + E[X|Y_2]$. The geometric interpretation of this is: to project a vector onto a plane defined by two orthogonal lines, we project to each line and then add the projections. The Kalman filter is one of the important applications of Probability to the fields of Signal Processing, Control, Estimation, and Inference (Catlin 1989).

By the term *conditioning* in Probability we often mean an effective application of the tower property in order to define a probability measure or to compute the expectation of a functional. For example, if $X_1, X_2, \ldots$ are i.i.d. positive random variables and an N is a geometric random variable, say $P(N = n) = \alpha^{n-1}(1 - \alpha)$, $n = 1, 2, \ldots$, then $E[\theta^{X_1+\cdots+X_N}] = E[E(\theta^{X_1+\cdots+X_N}|N)]$. But $E[\theta^{X_1+\cdots+X_N}|N = n] = E[\theta^{X_1+\cdots+X_n}] = (E[\theta^{X_1}])^n$, by independence. Hence $E[\theta^{X_1+\cdots+X_N}|N] = (E[\theta^{X_1}])^N$ and so $E[\theta^{X_1+\cdots+X_N}] = E\big[(E[\theta^{X_1}])^N\big] = (1 - \alpha)/(1 - \alpha E[\theta^{X_1}])$. Conditional expectation and probability are used in defining various classes of ▶stochastic processes such as ▶*martingales* and ▶*Markov chains* (Williams 1989). Conditional probability is a fundamental object in ▶*Bayesian statistics* (Williams 2001). Other applications are in the field of Financial Mathematics where the operation of taking conditional expectation of a future random variable with respect to the sigma-algebra of all events prior to the

current time t plays a fundamental role. In fact, it can be said that the notion of conditioning, along with that of independence and coupling, are the cornerstones of modern probability theory and its widespread applications.

About the Author

For biography *see* the entry ►Radon–Nikodým Theorem.

Cross References

►Bayes' Theorem
►Bayesian Statistics
►Bivariate Distributions
►Expected Value
►Radon–Nikodým Theorem

References and Further Reading

Catlin D (1989) Estimation, control, and the discrete Kalman filter. Springer, New York
Kallenberg O (2002) Foundations of modern probability, 2nd edn. Springer, New York
Kolmogorov A (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. Julius Springer, Berlin (English translation by Chelsea, New York, 1956)
Williams D (1989) Probability with Martingales. Cambridge University Press, Cambridge
Williams D (2001) Weighing the odds: a course in probability and statistics. Cambridge University Press, Cambridge

Confidence Distributions

William E. Strawderman
Professor
Rutgers University, Newark, NJ, USA

A Confidence Distribution (CD), $H(X,\theta)$, for a parameter is a function of the data, X, and the parameter in question, θ, such that: (a) for each data value X, $H(X,.)$ is a (continuous) cumulative distribution function for the parameter, and (b) for the true parameter value, θ_0, $H(.,\theta_0)$ has a uniform distribution (see ►Uniform Distribution in Statistics) on the interval (0,1). The concept of CD has its historic roots in Fisher's fiducial distribution (Fisher 1930), although, in its modern version, it is a strictly frequentist construct (Schweder and Hjort 2002, 2003; Singh et al. 2005, 2007, and see also Efron 1993). The CD carries a great deal of information pertinent to a variety of frequentist inferences and may be used for the construction of confidence intervals, tests of hypotheses and point estimation.

For instance, the αth quantile of $H(X,.)$, is the upper end of a $100(1-\alpha)$ percent one sided confidence interval for θ, and also the interval formed by the sth and tth quantiles $(s<t)$ is a $100(t-s)$ percent confidence interval. These properties indicate that a confidence distribution is, in a sense, a direct frequentist version of Fisher's fiducial distribution.

Similarly, a level α test of the one-sided hypothesis $K_0: \theta \le \theta_0$ versus $K_1: \theta > \theta_0$ is given by rejecting K_0 when $H(X,\theta_0) \le \alpha$, and an analogous result holds for testing $K_0: \theta \ge \theta_0$ versus $K_1: \theta < \theta_0$. Additionally, for testing the two sided hypothesis $K_0: \theta = \theta_0$ versus $K_1: \theta \ne \theta_0$, the rejection region $2\{\min(H(X,\theta_0), 1-H(X,\theta_0))\} \ge \alpha$ gives an α level test.

The CD may also be used in a natural ways to construct a point estimate of θ. Perhaps the most straightforward estimator is the median of $H(X,.)$, which is median unbiased, and under mild conditions, consistent. Another obvious estimator, $\tilde{\theta} = \int \theta(\partial \mathrm{H}(\mathrm{X},\theta)/\partial\theta)\mathrm{d}\theta$ is also consistent under weak conditions.

One particularly simple way to construct a CD is via a pivotal quantity, $\psi(X,\theta)$, a function of X and θ whose cumulative distribution function, $G(.)$ under the true θ does not depend on θ. Then $G(\Psi(X,\theta))$ is a CD provided $\Psi(X,\theta)$ is increasing in θ. Such quantities are easy to construct in invariant models such as location or scale models. Here is a prototypical example in a normal location model. Suppose $X_i \sim N(\theta,1)$, for $i = 1,\cdots,n$, are iid. Then $\Psi(X_1,\cdots,X_n,\theta) = (\overline{X}-\theta) \sim N(0,1/n)$ so that $H(X_1,\cdots,X_n,\theta) = \Phi^{-1}\left(\sqrt{n}(\theta-\overline{X})\right)$ is a CD.

Another common construction is based on a series of one sided α-level tests of $K_0: \theta \le \theta_0$ versus $K_1: \theta > \theta_0$. If the function $P[\theta_0, X]$ is a p-value for each value of θ_0, then typically $P[\theta_0,.]$ has a uniform distribution for each value of θ_0, and hence $H(X,\theta) = P[\theta, X]$ is a CD.

The above discussion can be extended naturally to include the notion of an asymptotic CD by replacing (b) above, with the requirement that $H(.,\theta_0)$ approaches a uniform distribution on $(0,1)$ weakly as the sample size approaches infinity, and dropping the continuity requirement in (a). Profile likelihoods (see, e.g., Efron 1993; Schweder and Hjort 2002; Singh et al. 2007), and Bootstrap Distributions (see Efron 1998; Singh et al. 2005, 2007) are asymptotic CD's under weak conditions.

It can also be extended to include nuisance parameters. For example, in the case of a sample from a normal population with unknown mean and variance, the usual t-pivot can be used to construct a CD for the mean, while the usual chi-square pivot can be used to construct a CD for the variance.

See Schweder and Hjort (2002, 2003) or Singh et al. (2005, 2007), for more detailed discussion on construction,

properties and uses of *CD*'s. In particular Singh et al. (2005) discusses the combination of information from independent sources via *CD*'s.

About the Author

William E. Strawderman is Professor of Statistics and Biostatistics at Rutgers University, and past chair of the Department of Statistics. He is a Fellow of IMS and ASA and an elected member if ISI and has served on the councils of IMS and ISBA, as Chair of the ASA Section on Bayesian Statistics. He had been on the editorial boards of the Annals of Statistics, JASA, and the IMS Lecture Notes series. He has won the Distinguished Alumni Award in Science of the Graduate School at Rutgers University, and the ASA's Youden Award in Interlaboratory Studies.

Cross References

►Bootstrap Methods
►Confidence Interval
►Data Depth
►Fiducial Inference

References and Further Reading

Efron B (1993) Bayes and likelihood calculations from confidence intervals. Biometrika 80:3–26

Efron B (1993) Empirical Bayes methods for combining likelihoods (with discussion), (1998). J Am Stat Assoc 91:538–565

Fisher RA (1930) Inverse Probability. Proc Camb Philos Soc 26: 528–535

Schweder T, Hjort NL (2002) Confidence and likelihood. Scand J Stat 29:309–332

Schweder T, Hjort NL (2003) Frequentist analogues of priors and posteriors. In: Econometrics and the philosophy of economics. Princeton University Press, Princeton, NJ, pp 285–317

Singh K, Xie M, Strawderman WE (2005) Combining information from independent sources through confidence distributions. Ann Stat 33:159–183

Singh K, Xie M, Strawderman WE (2007) Confidence distribution (CD)-distribution estimator of a parameter. In: Regina Liu, William Strawderman, and Cun-Hui Zhang (eds) Complex datasets and inverse problems. IMS Lecture Notes-Monograph Series, 54, pp 132–150

Confidence Interval

Michael Smithson
Professor
The Australian National University, Canberra, ACT, Australia

A $100(1-\alpha)\%$ confidence interval is an interval estimate around a population parameter θ that, under repeated random samples of size N, is expected to include θ's true value $100(1-\alpha)\%$ of the time. The confidence interval thereby indicates the precision with which a population parameter is estimated by a sample statistic, given N and α. For many statistics there are also methods of constructing *confidence regions,* which are multivariate versions of simultaneous confidence intervals.

The *confidence level,* $100(1-\alpha)\%$, is chosen a priori. A *two-sided* confidence interval uses a lower limit L and upper limit U that each contain θ's true value $100(1-\alpha/2)\%$ of the time, so that together they contain θ's true value $100(1-\alpha)\%$ of the time. This interval often is written as $[L, U]$, and sometimes writers combine a confidence level and interval by writing $\Pr(L \leq \theta \leq U) = 1-\alpha$. In some applications, a *one-sided* confidence interval is used, primarily when only one limit has a sensible meaning or when interest is limited to bounding a parameter estimate from one side only.

The confidence interval is said to be an inversion of its corresponding significance test because the $100(1-\alpha)\%$ confidence interval includes all hypothetical values of the population parameter that cannot be rejected by its associated significance test using a Type I error-rate criterion of α. In this respect, it provides more information than a significance test does. Confidence intervals become narrower with larger sample size and/or lower confidence levels. Narrower confidence intervals imply greater statistical power for the corresponding significance test, but the converse does not always hold.

The limits L and U are derived from a sample statistic (often the sample estimate of θ) and a sampling distribution specifying a probability for each value that the sample statistic can take. Thus L and U also are sample statistics and will vary from one sample to another. This fact underscores a crucial point of interpretation regarding a confidence interval, namely that we cannot claim that a particular interval has a $1-\alpha$ probability of containing the population parameter value.

A widespread practice regarding two-sided confidence intervals is to place L and U so that α is evenly split between the lower and upper tails. This is often a matter of convention, but can be dictated by criteria that statisticians have used for determining the "best" possible confidence interval. One such criterion is simply narrowness. It is readily apparent, for instance, that if a sampling distribution is symmetric and unimodal then for high confidence levels the shortest $100(1-\alpha)\%$ confidence interval constructed from that distribution is one that allocates $\alpha/2$ to the tails outside of the lower and upper limits.

Other criteria for evaluating confidence intervals are as follows. A $100(1-\alpha)\%$ confidence interval is *exact* if it can be expected to contain the relevant parameter's true value $100(1-\alpha)\%$ of the time. When approximate intervals

are used instead, if the rate of coverage is greater than $100(1-\alpha)\%$ then the interval is *conservative*; if the rate is less then the interval is *liberal*. The $100(1-\alpha)\%$ interval that has the smallest probability of containing values other than the true parameter value is said to be *uniformly most accurate*. A confidence interval whose probability of including any value other than the parameter's true value is less than or equal to $100(1-\alpha)\%$ is *unbiased*.

Example 1 Suppose that a standard IQ test has been administered to a random sample of $N = 25$ adults from a large population with a sample mean of 103 and standard deviation $s = 10$. We will construct a two-sided 95% confidence interval for the mean, μ. The limits U and L must have the property that, given a significance criterion of α, sample size of 25, mean of 103 and standard deviation of 10, we could reject the hypotheses that $\mu > 103 + U$ or $\mu < 103 - L$ but not $L \le \mu \le U$.

The sampling distribution of the t-statistic defined by $t = \dfrac{\overline{X} - \mu}{s_{err}}$ is a t-distribution with $df = N - 1 = 24$. When $df = 24$ the value $t_{\alpha/2} = 2.064$ standard-error units above the mean cuts $\alpha/2 = .025$ from the upper tail of this t-distribution, and likewise $-t_{\alpha/2} = -2.064$ standard-error units below the mean cuts $\alpha/2 = .025$ from the lower tail. The sample standard error is $s_{err} = s/\sqrt{N} = 4.128$. So a t-distribution around $U = 103 + (2.064)(4.128) = 107.13$ has .025 of its tail below 103, while a t-distribution around $L = 103 - (2.064)(4.128) = 98.87$ has 0.025 of its tail above 103. These limits fulfill the above required property, so the 95% confidence interval for μ is $[98.87, 107.13]$. Thus, we cannot reject hypothetical values of μ that lie between 98.87 and 107.13, using $\alpha = .05$.

Example 2 (transforming one interval to obtain another) Cohen's d for two independent samples is defined by $\delta = (\mu_1 - \mu_2)/\sigma_p$, where μ_1 and μ_2 are the means of two populations from which the samples have been drawn and σ_p is the population pooled standard deviation. This quantity has a noncentral t distribution with a noncentrality parameter $\Delta = \delta[N_1N_2/(N_1 + N_2)]^{1/2}$, where N_1 and N_2 are the sizes of the two samples. The sample t-statistic is the sample estimate of Δ. Suppose a two-condition between-subjects experiment with $N_1 = N_2 = 40$ yields $t(78) = 3.45$. Using an appropriate algorithm (Smithson 2003) we can find the 95% confidence interval for Δ, which is $[1.407, 5.473]$. Because δ and Δ are monotonically related by $\delta = \Delta/[N_1N_2/(N_1 + N_2)]^{1/2}$, we can obtain a 95% confidence interval for δ by applying this formula to the lower and upper limits of the interval for Δ. The sample estimate of δ is $d = t/[N_1N_2/(N_1 + N_2)]^{1/2} = 3.45/4.472 = 0.771$, and applying the same transformation to the limits of the interval for Δ gives an interval of $[0.315, 1.224]$ for δ.

About the Author

Michael Smithson is a Professor in the Department of Psychology at The Australian National University in Canberra, and received his Ph.D. from the University of Oregon. He is the author of *Confidence Intervals* (2003), *Statistics with Confidence* (2000), *Ignorance and Uncertainty* (1989), and *Fuzzy Set Analysis for the Behavioral and Social Sciences* (1987), coauthor of *Fuzzy Set Theory: Applications in the Social Sciences* (2006), and coeditor of *Uncertainty and Risk: Multidisciplinary Perspectives* (2008) and *Resolving Social Dilemmas: Dynamic, Structural, and Intergroup Aspects* (1999). His other publications include more than 120 refereed journal articles and book chapters. His primary research interests are in judgment and decision making under uncertainty and quantitative methods for the social sciences.

Cross References

►Confidence Distributions
►Decision Trees for the Teaching of Statistical Estimation
►Effect Size
►Fuzzy Logic in Statistical Data Analysis
►Margin of Error
►Sample Size Determination
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Inference
►Statistical Inference: An Overview

References and Further Reading

Altman DG, Machin D, Bryant TN, Gardner MJ (2000) Statistics with confidence: confidence intervals and statistical guidelines, 2nd edn. British Medical Journal Books, London

Smithson M (2003) Confidence intervals. Sage University Papers on Quantitative Applications in the Social Sciences, 139. Sage, Thousand Oaks

Confounding and Confounder Control

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

Introduction

The word *confounding* has been used to refer to at least three distinct concepts. In the oldest and most widespread usage, confounding is a source of bias in estimating causal

effects. This bias is sometimes informally described as a mixing of effects of extraneous factors (called confounders) with the effect of interest, and important in causal inference (*see* ▶Causation and Causal Inference). This usage predominates in nonexperimental research, especially in epidemiology and sociology. In a second and more recent usage originating in statistics, confounding is a synonym for a change in an effect measure upon stratification or adjustment for extraneous factors (a phenomenon called *noncollapsibility* or *Simpson's paradox; see* ▶Simpson's Paradox; Collapsibility). In a third usage, originating in the experimental-design literature, confounding refers to inseparability of main effects and interactions under a particular design (*see* ▶Interaction).

The three concepts are closely related and are not always distinguished from one another. In particular, the concepts of confounding as a bias in effect estimation and as noncollapsibility are often treated as equivalent, even though they are not. Only the first usage, confounding as a bias, will be described here; for more detailed coverage and comparisons of concepts see, Greenland et al. (1999a), Pearl (2009), and Greenland et al. (2008).

Confounding as a Bias in Effect Estimation

In the first half of the nineteenth century, John Stuart Mill described the problem of confounding in causal inference; he acknowledged the seventeenth century scientist Francis Bacon as a forerunner in dealing with these issues (Mill 1843, Chap. III). Mill listed a key requirement for an experiment intended to determine causal relations:

> "...none of the circumstances [of the experiment] that we do know shall have effects susceptible of being *confounded* with those of the agents whose properties we wish to study" (emphasis added) (Mill 1843, Chap. X).

In Mill's time the word "experiment" referred to an observation in which some circumstances were under the control of the observer, as it still is used in ordinary English, rather than to the notion of a comparative trial. Nonetheless, Mill's requirement suggests that a comparison is to be made between the outcome of our "experiment" (which is, essentially, an uncontrolled trial) and what we would expect the outcome to be if the agents we wish to study had been absent. If the outcome is not as one would expect in the absence of the study agents, then Mill's requirement ensures that the unexpected outcome was not brought about by extraneous "circumstances" (factors). If, however, those circumstances do bring about the unexpected outcome, and that outcome is mistakenly attributed to effects of the study agents, then the mistake is one of confounding (or confusion) of the extraneous effects with the agent effects.

Much of the modern literature follows the same informal conceptualization given by Mill. Terminology is now more specific, with "treatment" used to refer to an agent administered by the investigator and "exposure" often used to denote an unmanipulated agent. The chief development beyond Mill is that the expectation for the outcome in the absence of the study exposure is now almost always explicitly derived from observation of a control group that is untreated or unexposed.

Confounding typically occurs when natural or social forces or personal preferences affect whether a person ends up in the treated or control group, and these forces or preferences also affect the outcome variable. While such confounding is common in observational studies, it can also occur in randomized experiments when there are systematic improprieties in treatment allocation, administration, and compliance. A further and somewhat controversial point is that confounding (as per Mill's original definition) can also occur in perfect randomized trials due to *random* differences between comparison groups (Fisher 1935; Rothman 1977); this problem will be discussed further below.

The Potential-Outcome Model of Confounding

Various models of confounding have been proposed for use in statistical analyses. Perhaps the one closest to Mill's concept is based on the *potential-outcome* or counterfactual model for causal effects (*see* ▶Causation and Causal Inference). Suppose we wish to consider how a health-status (outcome) measure of a population would change in response to an intervention (population treatment). More precisely, suppose our objective is to determine the effect that applying a treatment x_1 had or would have on an outcome measure μ relative to applying treatment x_0 to a specific target population A. For example, this population could be a cohort of breast-cancer patients, treatment x_1 could be a new hormone therapy, x_0 could be a placebo therapy, and the measure μ could be the 5-year survival probability. The treatment x_1 is sometimes called the *index* treatment; and x_0 is sometimes called the *control* or *reference* treatment (which is often a standard or placebo treatment).

The potential-outcome model posits that, in population A, μ will equal μ_{A1} if x_1 is applied, μ_{A0} if x_0 is applied; the causal effect of x_1 relative to x_0 is defined as the change from μ_{A0} to μ_{A1}, which might be measured as $\mu_{A1} - \mu_{A0}$, or if μ is strictly positive, μ_{A1}/μ_{A0}. If A is given treatment x_1,

then μ will equal μ_{A1} and μ_{A1} will be observable, but μ_{A0} will be unobserved.

Suppose now that μ_{B0} is the value of the outcome μ observed or estimated for a population B that was administered treatment x_0. If this population is used as a substitute for the unobserved experience of population A under treatment x_0, it is called the control or reference population. *Confounding* is said to be present if $\mu_{A0} \neq \mu_{B0}$, for then there must be some difference between populations A and B *other than treatment difference* that is affecting μ.

If confounding is present, a naive (crude) association measure obtained by substituting μ_{B0} for μ_{A0} in an effect measure will not equal the effect measure, and the association measure is said to be *confounded.* Consider $\mu_{A1} - \mu_{B0}$, which measures the association of treatments with outcomes *across* the populations. If $\mu_{A0} \neq \mu_{B0}$, then $\mu_{A1} - \mu_{B0}$ is said to be *confounded* for $\mu_{A1} - \mu_{A0}$, which measures the effect of treatment x_1 on population A. Thus, to say an association measure $\mu_{A1} - \mu_{B0}$ is confounded for an effect measure $\mu_{A1} - \mu_{A0}$ is to say these two measures are not equal.

Dependence of Confounding on the Outcome Measure and the Population

A noteworthy aspect of the potential-outcome model is that confounding depends on the outcome measure. For example, suppose populations A and B have a different 5-year survival probability μ under placebo treatment x_0; that is, suppose $\mu_{B0} \neq \mu_{A0}$ so that $\mu_{A1} - \mu_{B0}$ is confounded for the actual effect $\mu_{A1} - \mu_{A0}$ of treatment on 5-year survival. It is then still possible that 10-year survival, v, under the placebo would be identical in both populations; that is v_{A0} could still equal v_{B0}, so that $v_{A1} - v_{B0}$ is not confounded for the actual effect of treatment on 10-year survival. Lest one think this situation unlikely, note that we should generally expect no confounding for 200-year survival, since no known treatment is likely to raise the 200-year survival probability of human patients above zero.

Even though the presence of confounding is dependent on the chosen outcome measure, as defined above its presence does not depend on how the outcome is contrasted between treatment levels. For example, if Y is binary so that $\mu = E(Y)$ is the Bernoulli parameter or risk $\Pr(Y = 1)$, then the risk difference $\mu_{A1} - \mu_{B0}$, risk ratio μ_{A1}/μ_{B0}, and odds ratio $\{\mu_{A1}/(1-\mu_{A1})\}/\{\mu_{B0}/(1-\mu_{B0})\}$ are all confounded under exactly the same circumstances. In particular, and somewhat paradoxically, confounding may be absent even if the odds ratio changes upon covariate adjustment, i.e., even if the odds ratio is noncollapsible (Greenland and Robins 1986; Greenland et al. 1999a, 2008; *see* ►Collapsibility).

A second noteworthy point is that confounding depends on the target population. The preceding example, with A as the target, had different 5-year survivals μ_{A0} and μ_{B0} for A and B under placebo therapy, and hence $\mu_{A1} - \mu_{B0}$ was confounded for the effect $\mu_{A1} - \mu_{A0}$ of treatment on population A. A lawyer or ethicist may also be interested in what effect the hormone treatment would have had on population B. Writing μ_{B1} for the (unobserved) outcome under treatment, this effect on B may be measured by $\mu_{B1} - \mu_{B0}$. Substituting μ_{A1} for the unobserved μ_{B1} yields $\mu_{A1} - \mu_{B0}$. This measure of association is confounded for $\mu_{B1} - \mu_{B0}$ (the effect of treatment x_1 on 5-year survival in population B) if and only if $\mu_{A1} \neq \mu_{B1}$. Thus, the same measure of association, $\mu_{A1} - \mu_{B0}$, may be confounded for the effect of treatment on neither, one, or both of populations A and B, and may or may not be confounded for the effect of treatment on other targets such as the combined population $A \cup B$.

Confounders (Confounding Factors) and Covariate Imbalance

The potential-outcome model is that it invokes no explicit differences (imbalances) between populations A and B with respect to circumstances or covariates that might influence μ. (Greenland and Robins 1986, 2009). Clearly, if μ_{A0} and μ_{B0} differ, then A and B must differ with respect to factors that influence μ. This observation has led some authors to define confounding as the presence of such covariate differences between the compared populations (Stone 1993). This is incorrect, however, because confounding is only a consequence of these covariate differences. In fact, A and B may differ profoundly with respect to covariates that influence μ, and yet confounding may be absent. In other words, a covariate difference between A and B is a necessary but not sufficient condition for confounding, as can be seen when the impact of covariate differences may balance each other out, leaving no confounding.

Suppose now that populations A and B differ with respect to certain covariates, and that these differences have led to confounding of an association measure for the effect measure of interest. The responsible covariates are then termed *confounders* of the association measure. In the above example, with $\mu_{A1} - \mu_{B0}$ confounded for the effect $\mu_{A1} - \mu_{A0}$, the factors responsible for the confounding (i.e., the factors that led to $\mu_{A0} \neq \mu_{B0}$) are the confounders.

It can be deduced that a variable cannot be a confounder unless it can affect the outcome parameter μ within treatment groups and it is distributed differently among the compared populations (e.g., see Yule 1903, who uses terms such as "fictitious association" rather than confounding). These two necessary conditions are sometimes

offered together as a definition of a confounder. Nonetheless, counterexamples show that the two conditions are not sufficient for a variable with more than two levels to be a confounder (Greenland et al. 1999a). Note that the condition of affecting the outcome parameter is a causal assertion and thus relies on background knowledge for justification (Greenland and Robins 1986; Robins 2001; Pearl 2009).

Control of Confounding

Prevention of Confounding

An obvious way to avoid confounding is estimating $\mu_{A1} - \mu_{A0}$ is to obtain a reference population B for which μ_{B0} is known to equal μ_{A0}. Such a population is sometimes said to be *comparable* to or *exchangeable* with A with respect to the outcome under the reference treatment. In practice, such a population may be difficult or impossible to find. Thus, an investigator may attempt to construct such a population, or to construct exchangeable index and reference populations. These constructions may be viewed as design-based methods for the control of confounding.

Perhaps no approach is more effective for preventing confounding by a known factor than *restriction*. For example, gender imbalances cannot confound a study restricted to women. However, there are several drawbacks: restriction on enough factors can reduce the number of available subjects to unacceptably low levels, and may greatly reduce the generalizability of results as well. *Matching* the treatment populations on confounders overcomes these drawbacks, and, if successful, can be as effective as restriction. For example, gender imbalances cannot confound a study in which the compared groups have identical proportions of women. Unfortunately, differential losses to observation may undo the initial covariate balances produced by matching.

Neither restriction nor matching prevents (although it may diminish) imbalances on unrestricted, unmatched, or unmeasured covariates. In contrast, ►*randomization* offers a means of dealing with confounding by covariates not accounted for by the design. It must be emphasized, however, that this solution is only probabilistic and subject to severe constraints in practice. Randomization is not always feasible or ethical, and many practical problems (such as differential loss and noncompliance) can lead to confounding in comparisons of the groups actually receiving treatments x_1 and x_0.

One somewhat controversial solution to noncompliance problems is *intent-to-treat analysis,* which defines the comparison groups A and B by treatment assigned rather than treatment received. Confounding may, however, affect even intent-to-treat analyses, and (contrary to widespread misperceptions) the bias in those analyses can exaggerate the apparent treatment effect (Robins 1998). For example, the assignments may not always be random, as when blinding is insufficient to prevent the treatment providers from protocol violations. And, purely by bad luck, randomization may itself produce allocations with severe covariate imbalances between the groups (and consequent confounding), especially if the study size is small (Fisher 1935; Rothman 1977). *Blocked* (matched) randomization can help ensure that random imbalances on the blocking factors will not occur, but it does not guarantee balance of unblocked factors.

Adjustment for Confounding

Design-based methods are often infeasible or insufficient to prevent confounding. Thus, there has been an enormous amount of work devoted to analytic adjustments for confounding. With a few exceptions, these methods are based on observed covariate distributions in the compared populations. Such methods can successfully control confounding only to the extent that enough confounders are adequately measured. Then, too, many methods employ parametric models at some stage, and their success may thus depend on the faithfulness of the model to reality. These issues cannot be covered in depth here, but a few basic points are worth noting. The simplest and most widely trusted methods of adjustment begin with *stratification* on confounders. A covariate cannot be responsible for confounding within internally homogeneous strata of the covariate. For example, gender imbalances cannot confound observations within a stratum composed solely of women. More generally, comparisons within strata cannot be confounded by a covariate that is unassociated with treatment within strata. This is so, whether the covariate was used to define the strata or not. Thus, one need not stratify on all confounders in order to control confounding; it suffices to stratify on a balancing score (such as a propensity score) that yields strata in which the confounders are unassociated with treatment.

If one has accurate background information on relations among the confounders, one may use this information to identify sets of covariates statistically sufficient for adjustment, for example by using causal diagrams or conditional independence conditions (Pearl 1995, 2009; Greenland et al. 1999ab; Glymour and Greenland 2008). Nonetheless, if the stratification on the confounders is too coarse (e.g., because categories are too broadly defined), stratification may fail to adjust for much of the confounding by the adjustment variables.

One of the most common adjustment approaches today is to enter suspected confounders into a model for the outcome parameter μ. For example, let μ be the mean (expectation) of an outcome variable of interest Y,

let X be the treatment variable of interest, and let Z be a suspected confounder of the $X - Y$ relation. Adjustment for Z is often made by fitting a generalized-linear model (see ▶Generalized Linear Models) $g(\mu) = g(\alpha + \beta x + \gamma z)$ or some variant, where $g(\mu)$ is a strictly increasing function such as the natural log $\ln(\mu)$, as in log-linear modeling, or the logit function $\ln\{\mu/(1 - \mu)\}$, as in ▶logistic regression; the estimate of β that results is then taken as the Z-adjusted estimate of the X effect on $g(\mu)$.

An oft-cited advantage of model-based adjustment is that it allows adjustment for more variables and in finer detail than stratification. If however the form of the fitted model cannot adapt well to the true dependence of Y on X and Z, such model-based adjustments may fail to adjust for confounding by Z. For example, suppose Z is symmetrically distributed around zero within X levels, and the true dependence is $g(\mu) = g(\alpha + \beta x + \gamma z^2)$; then using the model $g(\mu) = g(\alpha + \beta x + \gamma z)$ will produce little or no adjustment for Z. Similar failures can arise in adjustments based on models for treatment probability (propensity scores). Such failures can be minimized or avoided by using reasonably flexible models, by carefully checking each fitted model against the data, and by combining treatment-probability and outcome models to produce *doubly robust* effect estimators (Hirano et al. 2003; Bang and Robins 2005).

Finally, if (as is often done) a variable used for adjustment is not a confounder, bias may be introduced by the adjustment (Greenland and Neutra 1980; Greenland et al. 1999b; Hernán et al. 2002; Pearl 2009). The form of this bias often parallels *selection bias* familiar to epidemiologists, and tends to be especially severe if the variable is affected by both the treatment and the outcome under study, as in classic *Berksonian bias* (Greenland 2003). In some but not all cases the resulting bias is a form of confounding within strata of the covariate (Greenland et al. 1999b); adjustment for covariates affected by treatment can produce such confounding, even in randomized trials (Cox 1958, Chap. 2; Greenland 2003).

Confounded Mechanisms Versus Confounded Assignments

If the mechanism by which the observational units come to have a particular treatment is independent of the potential outcomes of the units, the mechanism is sometimes described as *unconfounded* or *unbiased* for μ (Rubin 1991; Stone 1993); otherwise the mechanism is confounded or biased. Randomization is the main practical example of such a mechanism. Graphical models (*see* ▶Causal Diagrams) provide an elegant algorithm for checking whether the graphed mechanism is unconfounded within strata of covariates (Pearl 1995, 2009; Greenland et al. 1999b; Glymour and Greenland 2008). Note however that in typical epidemiologic usage the term "confounded" refers to the result of a single assignment (the study group actually observed), not the behavior of the mechanism. Thus an unconfounded mechanism can by chance produce confounded assignments.

The latter fact resolves a controversy about adjustment for baseline (pre-treatment) covariates in randomized trials. Although Fisher asserted that randomized comparisons were "unbiased," he also pointed out that particular assignments could be confounded in the single-trial sense used in epidemiology; see Fisher (1935, p. 49). Resolution comes from noting that Fisher's use of the word "unbiased" referred to the design and corresponds to an unconfounded assignment mechanism; it was not meant to guide analysis of a given trial (which has a particular assignment). Once the trial is underway and the actual treatment allocation is completed, the unadjusted treatment-effect estimate will be biased conditional on the observed allocation if the baseline covariate is associated with treatment in the allocation and the covariate affects the outcome; this bias can be removed by adjustment for the covariate (Rothman 1977; Greenland and Robins 1986, 2009; Greenland et al. 1999a).

Confounder Selection

An essential first step in the control of confounding is to identify which variables among those measured satisfied the minimal necessary conditions to be a confounder. This implies among other things that the variables cannot be affected by exposure or outcome; it thus excludes intermediate variables and effects of exposure and disease, whose control could introduce Berksonian bias. This initial screening is primarily a subject-matter decision that requires consideration of the causal ordering of the variables. Relatively safe candidate confounders will be "pre-treatment" covariates (those occurring before treatment or exposure), which have the advantage that they cannot be intermediates or effects of exposure and outcome. Exceptions occur in which control of certain pre-treatment variables introduce bias (Pearl 1995, 2009; Greenland et al. 1999b), although the bias so introduced may be much less than the confounding removed (Greenland 2003).

Variables that pass the initial causal screening are sometimes called "potential confounders." Once these are identified, the question arises as to which must be used for adjustment. A common but unjustified strategy is to select confounders to control based on a test (usually a significance test) of each confounder's association with the treatment X (a test of imbalance) or with the outcome Y, e.g.,

using stepwise regression. Suppose Z is a pre-treatment covariate (potential confounder). The strategy of testing the Z association with X arises from a confusion of two distinct inferential problems:

1. Do the treated ($X = 1$) evince larger differences from the untreated ($X = 0$) with respect to Z than one should expect from a random (or unconfounded) assignment mechanism?
2. Should we control for Z to estimate the treatment effect?

A test of the $X - Z$ association addresses question (a), but not (b). For (b), the "large-sample" answer is that control is advisable, regardless of whether the $X - Z$ association is random. This is because an imbalance produces bias conditional on the observed imbalance, even if the imbalance derived from random variation.

The mistake of significance testing for confounding lies in thinking that one can ignore an imbalance if it is from random variation. Random assignment only guarantees valid performance of statistics over all possible treatment allocations. It does not however guarantee validity conditional on the observed Z imbalance, even though any such imbalance must be random in a randomized trial. Thus the $X - Z$ test addresses a real question (one relevant to a field methodologist studying determinants of response/treatment), but is irrelevant to the second question (b) (Greenland and Neutra 1980; Robins and Morgenstern 1987; Greenland et al. 1999a).

The case of testing the Z association with Y devolves in part to whether one trusts prior (subject-matter) knowledge that Z affects Y (or is a proxy for a cause of Y) more than the results of a significance test in one's own limited data. There are many examples in which a well-known risk factor exhibits the expected association with Y in the data, but for no more than chance reasons or sample-size limitations, that association fails to reach conventional levels of "significance" (e.g., Greenland and Neutra 1980). In such cases there is a demonstrable statistical advantage to controlling Z, thus allowing subject-matter knowledge to over-ride nonsignificance (Robins and Morgenstern 1987).

Another problematic strategy is to select a potential confounder Z for control based on how much the effect estimate changes when Z is controlled. Like the testing methods described above, it also lacks formal justification and can exhibit poor performance in practice (Maldonado and Greenland 1993). The strategy can also mislead if the treatment affects a high proportion of subjects and one uses a "noncollapsible" effect measure (one that changes upon stratification even if no confounding is present), such as an odds ratio or rate ratio (Greenland and Robins 1986; Greenland 1996; Greenland et al. 1999a).

In practice, there may be too many variables to control using conventional methods, so the issue of confounder selection may seem pressing. Nonetheless, hierarchical-Bayesian or other shrinkage methods may be applied instead. These methods adjust for all the measured confounders by estimating the confounder effects using a prior distribution for those effects. See Greenland (2000, 2008) for details. Some of these methods (e.g., the Lasso; Tibshirani 1996) may drop certain variables entirely, and thus in effect result in confounder selection; unlike significance-testing based selection, however, this selection has a justification in statistical theory.

About the Author

Dr. Greenland is Professor of Epidemiology, UCLA School of Public Health, and Professor of Statistics, UCLA College of Letters and Science. He has published over 300 scientific papers, two of which have been cited over 500 times. Professor Greenland is a coeditor (with Kenneth J.h Rothman) of the highly influential text in the field of epidemiology, *Modern Epidemiology*. This book has received the highest number of citations among all texts and papers in the field, over 8,000 (M. Porta et al. 2006). *Book citations: influence of epidemiologic thought in the academich community*, Rev Saúde Púbica, 40, p. 50; Lippincott-Raven 1998. Currently, he is an Associate editor for *Statistics in Medicine*. He was Chair, Section in Epidemiology, American Statistical Association (2005–2007). He is Chartered Statistician and Fellow, Royal Statistical Society (1993), and a Fellow American Statistical Association (1998).

Cross References

►Bias Analysis
►Causal Diagrams
►Causation and Causal Inference
►Collapsibility
►Exchangeability
►Interaction
►Simpson's Paradox
►Statistical Methods in Epidemiology

References and Further Reading

Bang H, Robins J (2005) Doubly robust estimation in missing data and causal inference models. Biometrics 61:962–972
Cox DR (1958) Planning of experiments. Wiley, New York
Fisher RA (1935) The Design of experiments. Oliver & Boyd, Edinburgh

Glymour MM, Greenland S (2008) Causal diagrams. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott, Philadelphia, pp 183–209
Greenland S (1996) Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. Epidemiology 7:498–501
Greenland S (2000) When should epidemiologic regressions use random coefficients? Biometrics 56:915–921
Greenland S (2003) Quantifying biases in causal models. Epidemiology 14:300–307
Greenland S (2008) Variable selection and shrinkage in the control of multiple confounders. Am J Epidemiol 167:523–529. Erratum 1142
Greenland S, Neutra RR (1980) Control of confounding in the assessment of medical technology. Int J Epidemiol 9: 361–367
Greenland S, Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15:413–419
Greenland S, Robins JM (2009) Identifiability, exchangeability, and confounding revisited. Epidemiol Perspect Innov (online journal) 6:4
Greenland S, Robins JM, Pearl J (1999a) Confounding and collapsibility in causal inference. Stat Sci 14:29–46
Greenland S, Pearl J, Robins JM (1999b) Causal diagrams for epidemiologic research. Epidemiology 10:37–48
Greenland S, Rothman KJ, Lash TL (2008) Measures of effect and measures of association. In: Rothman KJ, Greenland S, Lash TL (eds) Modern epidemiology, 3rd edn. Lippincott, Philadelphia, pp 51–70
Hernán M, Hernandez-Diaz S, Werler MM, Mitchell AA (2002) Causal knowledge as a prerequisite for confounding evaluation. Am J Epidemiol 155:176–184
Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71:1161–1189
Maldonado G, Greenland S (1993) A simulation study of confounder-selection strategies. Am J Epidemiol 138: 923–936
Mill JS (1843) A system of logic, ratiocinative and inductive. Reprinted by Longmans. Green & Company, London, 1956
Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–710
Pearl J (2009) Causality, 2nd edn. Cambridge University Press, New York
Robins JM (1998) Correction for non-compliance in equivalence trials. Stat Med 17:269–302
Robins JM (2001) Data, design, and background knowledge in etiologic inference. Epidemiology 12:313–320
Robins JM, Morgenstern H (1987) The foundations of confounding in epidemiology. Comput Math Appl 14:869–916
Rothman KJ (1977) Epidemiologic methods in clinical trials. Cancer 39:1771–1775
Rubin DB (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. Biometrics 47:1213–1234
Stone R (1993) The assumptions on which causal inference rest. J R Stat Soc B 55:455–466
Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc B 58:267–288
Yule GU (1903) Notes on the theory of association of attributes in statistics. Biometrika 2:121–134

Contagious Distributions

Seng Huat Ong[1], Choung Min Ng
[1]Professor
University of Malaya, Kuala Lumpur, Malaysia

The term contagious distribution was apparently first used by Neyman (1939) for a discrete distribution that exhibits clustering or contagious effect. The classical Neyman Type *A* distribution is one well-known example. However, contagious distributions are used nowadays to describe a plethora of distributions, many of which possess complicated probability distribution expressed in terms of special functions (see, for instance, Johnson et al. 2005).

It is instructive to give an account of the derivation of the Neyman Type *A* distribution as developed by Neyman (1939) in his paper "On a new class of contagious distributions applicable in entomology and bacteriology." Neyman wanted to model the distribution of larvae on plots in a field. He assumed that the number of clusters of eggs per unit area, N, followed a Poisson distribution with mean θ denoted by $Poi(\theta)$, while the number of larvae developing from the egg clusters $X_i, i = 1, 2, \ldots, N$ is distributed as another Poisson distribution $Poi\ (\phi)$. Mathematically, this may be expressed as follows:

$$S_N = X_1 + X_2 + \ldots + X_N$$

where S_N is the total number of larvae per unit area. The distribution of S_N is then a Neyman Type *A*.

The above model is known as a *true contagion* model, where the occurrence of "a favourable event depends on the occurrence of the previous favorable events" (Gurland 1959). Among other terms used for the distribution arising from this model are *generalized, clustered, stopped*, or *stopped-sum* distribution (see Douglas 1980; Johnson et al. 2005). It is convenient to represent the distribution of S_N concisely by

$$S_N \sim U_1 \vee U_2,$$

which reads S_N distribution is a U_1 distribution generalized by a U_2 distribution. As an example, the Neyman Type *A* distribution for S_N, above, is

$$\text{Neyman Type } A \sim Poi\,(\theta) \vee Poi\,(\phi).$$

In addition, the probability generating function (pgf) of S_N distribution is

$$E\left[z^{S_N}\right] = g_1(g_2(z)),$$

where $g_i(z)$ is the pgf for the corresponding $U_i, i = 1, 2$ distribution.

Next, using the same example as above but instead, the number of larvae per unit area is now considered to be distributed as a Poisson distribution, $Poi\,(k\theta)$, where due to heterogeneity, the mean number of eggs that hatched into larvae is assumed to vary with k following a Poisson distribution $Poi\,(\phi)$. The distribution for the number of larvae per unit area is again a Neyman Type A.

The model that gives rise to a distribution as in the preceding formulation is known as an *apparent contagion* model. Generally, this distribution arises when a parameter of a U_1 distribution is a random variable Ω that follows a U_2 distribution of its own. This type of distribution is also known as a *mixed, mixture*, or *compound* distribution (see Ord 1971; Johnson et al. 2005). A compound (mixture) distribution can be represented by

$$U_1 \underset{\Omega}{\wedge} U_2,$$

which means that the U_1 distribution is compounded by U_2 distribution (the distribution of Ω). U_2 is known as the compounding (mixing) distribution. Thus, the Neyman Type A distribution formulated through compounding is represented by

$$\text{Neyman Type } A \sim Poi\,(k\theta) \underset{k}{\wedge} Poi\,(\phi)$$

The pgf of a compound (mixture) distribution is

$$\int_{\omega} g_1\,(z|\,\omega)\; dF_2\,(\omega)$$

where $g_1(.)$ is the pgf for the U_1 and $F_2(.)$ is the cumulative distribution function for U_2. The class of mixed Poisson distributions is a well-known class of compound distributions that has found applications in many areas of study including biology, sociology, and medicine.

Note that both given examples of contagion models lead to the Neyman Type A distribution. The relationship between the generalized and compound distributions is given by the following theorem:

Theorem 1 (Gurland 1957) Let U_1 be a random variable with pgf $[h\,(z)]^{\theta}$, where θ is a given parameter. Suppose now θ is regarded as a random variable. Then, whatever be U_2

$$U_1 \wedge U_2 \sim U_2 \vee U_1. \tag{1}$$

This relation shows that it may not be possible to distinguish between the two types of contagion directly from the data (Gurland 1959).

Contagious distributions have been studied by many researchers including Feller (1943), Skellam (1952), Beall and Rescia (1953), Gurland (1958), Hinz and Gurland (1970), Khatri (1971), and Hill (1993), creating a rich literature in this field. The readers are referred to Ord (1972, p. 126) for a list of generalized and compound Poisson distributions such as Polya-Aeppli, negative binomial, and Hermite distributions. Other references for generalized and compound distributions can be found in Douglas (1980, Chaps. 4 and 5) and Johnson et al. (2005, Chaps. 8 and 9). These references also describe statistical inference for the contagious distributions. Recent review articles on this subject are Gupta and Ong (2005) and Karlis and Xekalaki (2005).

Cross References

►Mixture Models
►Multivariate Statistical Distributions

References and Further Reading

Beall G, Rescia RR (1953) A generalization of Neyman's contagious distributions. Biometrics 9:354–386
Douglas JB (1980) Analysis with standard contagious distributions. International Co-operative Publishing House, Burtonsville
Feller W (1943) On a general class of "contagious" distributions. Ann Math Stat 14:389–400
Gupta RC, Ong SH (2005) Analysis of long-tailed count data by Poisson mixtures. Commun Stat 34(3):557–574
Gurland J (1957) Some interrelations among compound and generalized distributions. Biometrika 44:265–268
Gurland J (1958) A generalized class of contagious distributions. Biometrics 14:229–249
Gurland J (1959) Some applications of the negative binomial and other contagious distributions. Am J Public Health 49:1388–1399
Hill DH (1991) Response and sequencing errors in surveys: a discrete contagious regression analysis. J Am Stat Assoc 88:775–781
Hinz P, Gurland J (1970) A test of fit for the negative binomial and other contagious distributions. J Am Stat Assoc 65: 887–903
Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, Hoboken
Karlis D, Xekalaki E (2005) Mixed Poisson distributions. Int Stat Rev 73:35–58
Khatri CG (1971) On multivariate contagious distributions. Sankhya 33:197–216
Neyman J (1939) On a new class of "contagious" distributions, applicable in entomology and bacteriology. Ann Math Stat 10: 35–57
Ord JK (1972) Families of frequency distributions. Griffin, London
Skellam JG (1952) Studies in statistical ecology: I. spatial pattern. Biometrika 39:346–362

Continuity Correction

Rabi Bhattacharya
Professor of Mathematics
The University of Arizona, Tucson, AZ, USA

According to the central limit theorem (*CLT*) (see ►Central Limit Theorems), the distribution function F_n of a normalized sum $n^{-1/2}(X_1 + \ldots + X_n)$ of n independent random variables $X_1, \ldots, X_n$, having a common distribution with mean zero and variance $\sigma^2 > 0$, converges to the distribution function Φ_σ of the normal distribution with mean zero and variance σ^2, as $n \to \infty$. We will write Φ for Φ_1 for the case $\sigma = 1$. The densities of Φ_σ and Φ are denoted by ϕ_σ and ϕ, respectively. In the case $X_j's$ are discrete, F_n has jumps and the normal approximation is not very good when n is not sufficiently large. This is a problem which most commonly occurs in statistical tests and estimation involving the normal approximation to the binomial and, in its multi-dimensional version, in Pearson's frequency ►chi-square tests, or in tests for association in categorical data. Applying the *CLT* to a binomial random variable T with distribution $B(n,p)$, with mean np and variance $npq(q = 1 - p)$, the normal approximation is given, for integers $0 \le a \le b \le n$, by

$$P(a \le T \le b) \approx \Phi\left((b - np)/\sqrt{npq}\right) - \Phi\left((a - np)/\sqrt{npq}\right). \quad (1)$$

Here $\approx$ indicates that the difference between its two sides goes to zero as $n \to \infty$. In particular, when $a = b$, the binomial probability $P(T = b) = C_b^n p^b q^{n-b}$ is approximated by zero. This error is substantial if n is not very large. One way to improve the approximation is to think graphically of each integer value b of T being uniformly spread over the interval $\left[b - \frac{1}{2}, b + \frac{1}{2}\right]$. This is the so called *histogram approximation*, and leads to the *continuity correction* given by replacing $\{a \le T \le b\}$ by $\left\{a - \frac{1}{2} \le T \le b + \frac{1}{2}\right\}$

$$P\left(a - \frac{1}{2} \le T \le b + \frac{1}{2}\right) \approx \Phi\left(\left(b + \frac{1}{2} - np\right)/\sqrt{npq}\right) - \Phi\left(\left(a - \frac{1}{2} - np\right)/\sqrt{npq}\right). \quad (2)$$

To give an idea of the improvement due to this correction, let $n = 20, p = 0.4$. Then $P(T \le 7) = 0.4159$, whereas the approximation (1) gives a probability $\Phi(-0.4564) = 0.3240$, and the continuity correction (2) yields $\Phi(-0.2282) = 0.4177$. Analogous continuity corrections apply to the Poisson distribution with a large mean.

For a precise mathematical justification of the continuity correction consider, in general, i.i.d. integer-valued random variables $X_1, \ldots, X_n$, with lattice span 1, mean μ, variance σ^2, and finite moments of order at least four. The distribution function $F_n(x)$ of $n^{-1/2}(X_1 + \ldots + X_n)$ may then be approximated by the ►*Edgeworth expansion* (See Bhattacharya and Ranga 1976, p. 239, or Gnedenko and Kolmogorov 1954, p. 213)

$$F_n(x) = \Phi_\sigma(x) - n^{-\frac{1}{2}} S_1\left(n\mu + n^{\frac{1}{2}}x\right)\phi_\sigma(x) + n^{-\frac{1}{2}}\mu_3/(6\sigma^3)(1 - x^2/\sigma^2)\phi_\sigma(x) + O(n^{-1}), \quad (3)$$

where $S_1(y)$ is the right continuous periodic function $y - \frac{1}{2}$ (mod 1) which vanishes when $y = \frac{1}{2}$. Thus, when a is an integer and $x = (a - n\mu)/\sqrt{n}$, replacing a by $a + \frac{1}{2}$ (or $a - \frac{1}{2}$) on the right side of (3) gets rid of the discontinuous term involving S_1.

Consider next the continuity correction for the *(Mann-Whitney-)Wilcoxon two sample test* (see ►Wilcoxon–Mann–Whitney Test). Here one wants to test nonparametrically if one distribution G is stochastically larger than another distribution F, with distribution functions $G(.), F(.)$. Then the null hypothesis is $H_0 : F(x) = G(x)$ for all x, and the alternative is $H_1 : G(x) \le F(x)$ for all x, with strict inequality for some x. The test is based on independent random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from the two unknown continuous distributions F and G, respectively. The test statistic is W_s = the sum of the ranks of the $Y_j's$ in the combined sample of $m + n$ $X_i's$ and $Y_j's$. The test rejects H_0 if $W_s \ge c$, where c is chosen such that the probability of rejection under H_0 is a given level α. It is known (see Lehmann 1975, pp. 5–18) that W_s is asymptotically normal and $E(W_s) = \frac{1}{2}n(m + n + 1)$, $Var(W_s) = mn(m + n + 1)/12$. Since W_s is integer-valued, the continuity correction yields

$$P(W_s \ge c|H_0) = P\left(W_s \ge c - \frac{1}{2}|H_0\right) \approx 1 - \Phi(z), \quad (4)$$

where $z = \left(c - \frac{1}{2} - \frac{1}{2}n(m + n + 1)\right)/\sqrt{mn(m + n + 1)/12}$.

As an example, let $m = 5$, $n = 7$, $c = 54$. Then $P(W_s \ge 54 \mid H_0) = 0.101$, and its normal approximation is $1 - \Phi(1.380) = 0.0838$. The continuity correction yields the better approximation $P(W_s \ge 54 \mid H_0) = P(W_s \ge 53.5 \mid H_0) \approx 1 - \Phi(1.299) = 0.0097$.

The continuity correction is also often used in 2×2 *contingency tables* for testing for association between two categories. It is simplest to think of this as a two-sample problem for comparing two proportions p_1, p_2 of individuals with a certain characteristic (e.g., smokers) in two populations (e.g., men and women), based on two independent random samples of sizes n_1, n_2 from the two populations, with $n = n_1 + n_2$. Let r_1, r_2 be the numbers in the samples possessing the characteristic. Suppose first that we

wish to test $H_0 : p_1 = p_2$, against $H_1 : p_1 < p_2$. Consider the test which rejects H_0, in favor of H_1, if $r_2 \geq c(r)$, where $r = r_1 + r_2$, and $c(r)$ is chosen so that the conditional probability (under H_0) of $r_2 \geq c(r)$, given $r_1 + r_2 = r$, is α. This is the uniformly most powerful unbiased (*UMPU*) test of its size (See Lehmann 1959, pp. 140–146, or Kendall and Stuart 1973, pp. 570–576). The conditional distribution of r_2, given $r_1 + r_2 = r$, is multinomial, and the test using it is called *Fisher's exact test*. On the other hand, if $n_i p_i \geq 5$ and $n_i(1 - p_i) \geq 5$ $(i = 1, 2)$, the normal approximation is generally used to reject H_0. Note that the (conditional) expectation and variance of r_2 are $n_2 r/n$ and $n_1 n_2 r(n-r)/[n^2(n-1)]$, respectively (See Lehmann 1975, p. 216). The normalized statistic t is then

$$t = [r_2 - n_2 r/n]/\sqrt{n_1 n_2 r(n-r)/[n^2(n-1)]}, \quad (5)$$

and H_0 is rejected when t exceeds $z_{1-\alpha}$, the $(1 - \alpha)th$ quantile of Φ. For the continuity correction, one subtracts $\frac{1}{2}$ from the numerator in (5), and rejects H_0 if this adjusted t exceeds $z_{1-\alpha}$. Against the two-sided alternative $H_1 : p_1 \neq p_2$, Fisher's *UMPU* test rejects H_0 if r_2 is either too large or too small. The corresponding continuity corrected t rejects H_0 if either the adjusted t, obtained by subtracting $\frac{1}{2}$ from the numerator in (5), exceeds $z_{1-\alpha/2}$, or if the t adjusted by adding $\frac{1}{2}$ to the numerator in (5) is smaller than $-z_{1-\alpha/2}$. This may be compactly expressed as

$$\text{Reject } H_0 \text{ if } V \equiv (n-1)\left[\,|\, r_1 n_2 - r_2 n_1 \,|\, - \frac{1}{2}n\right]^2 \Big/ (n_1 n_2 r(n-r)) > \chi^2_{1-\alpha}(1), \quad (6)$$

where $\chi^2_{1-\alpha}(1)$ is the $(1 - \alpha)$th quantile of the ▶chi-square distribution with one degree of freedom. This two-sided continuity correction was originally proposed by F. Yates in 1934, and it is known as *Yates' correction*. For numerical improvements due to the continuity corrections above, we refer to Kendall and Stuart (1973, pp. 575–576) and Lehmann (1975, pp. 215–217). For a critique, see Connover (1974). If the sampling of n units is done at random from a population with two categories (men and women), then the *UMPU* test is still the same as Fisher's test above, conditioned on fixed marginals n_1,(and, therefore, n_2) and r.

Finally, extensive numerical computations in Bhattacharya and Chan (1996) show that the chisquare approximation to the distribution of *Pearson's frequency chi-square* statistic is reasonably good for degrees of freedom 2 and 3, even in cases of small sample sizes, extreme asymmetry, and values of expected cell frequencies much smaller than 5. One theoretical justification for this may be found in the classic work of Esseen (1945), which shows that the error of chisquare approximation is $O(n^{-d/(d+1)})$ for degrees of freedom d.

Acknowledgments

The author acknowledges support from the NSF grant DMS 0806011.

About the Author

For biography *see* the entry ▶Random Walk.

Cross References

▶Binomial Distribution
▶Chi-Square Test: Analysis of Contingency Tables
▶Wilcoxon–Mann–Whitney Test

References and Further Reading

Bhattacharya RN, Chan NH (1996) Comparisons of chisquare, Edgeworth expansions and bootstrap approximations to the distribution of the frequency chisquare. Sankhya Ser A 58:57–68

Bhattacharya RN, Ranga Rao R (1976) Normal approximation and asymptotic expansions. Wiley, New York

Connover WJ (1974) Some reasons for not using Yates' continuity correction on 2 × 2 contingency tables. J Am Stat Assoc 69:374–376

Esseen CG (1945) Fourier analysis of distribution functions: a mathematical study of the Laplace–Gaussian law. Acta Math 77:1–125

Gnedenko BV, Kolmogorov AN (1954) Limit distributions of sums of independent random variables. English translation by K.L. Chung, Reading

Kendall MG, Stuart A (1973) The advanced theory of statistics, vol 2, 3rd edn. Griffin, London

Lehmann EL (1959) Testing statistical hypotheses. Wiley, New York

Lehmann EL (1975) Nonparametrics: statistical methods based on ranks. (With the special assistance of D'Abrera, H.J.M.). Holden-Day, Oakland

Control Charts

Alberto Luceño
Professor
University of Cantabria, Santander, Spain

Introduction

A control chart is a graphical statistical device used to monitor the performance of a repetitive process. Control charts were introduced by Shewhart in the 1920s while working for Western Electric and Bell Labs and, since then, they have been routinely used in Statistical Process Control (SPC). According to Shewhart, control charts are useful to define the standard to be attained for a process, to help

attaining that standard, and to judge whether that standard has been reached.

Variability and Its Causes

Any manufacturing or business process shows some degree of variability. This is obviously true when little effort has been made to try to keep the process stable around a target, but it continues to be true even when a lot of effort has already been dedicated to stabilize the process. In other words, the amount of variability can be reduced (as measured, for example, by the output standard deviation), but cannot be eliminated completely. Therefore, some knowledge about the types of variability that can be encountered in practice and the causes of this variability is necessary.

Concerning the types of variability, one must recognize at least the difference between stationary and non-stationary behavior, the former being desirable, the latter undesirable. A stationary process has fixed mean, variance and probability distribution, so that it is difficult (if not impossible) to perfectly attain this desirable state in practice. A non-stationary process does not have fixed mean, variance or probability distribution, so that its future behavior is unpredictable. Moreover, any natural process, when left to itself, tends to be non-stationary, sometimes in the long run, but most often in the short run. Consequently, some control effort is almost always necessary to, at least, induce stationarity in the process. Control charts are useful for this purpose.

Concerning the causes of variability, the most obvious facts are that there are a lot of causes, that many of them are unknown, and, consequently, that they are difficult to classify. Nevertheless, Shewhart suggested that it is conceptually useful to classify the causes of variability in two groups: common causes and special causes. Common causes are those that are still present when the process has been brought to a satisfactory stationary state of control; they can be described as chance variation, because the observed variation is the sum of many small effects having different causes. Special causes are those that have larger effects and, hence, have the potential to send the process out of control; hopefully, they can eventually be discovered (assigned) and permanently removed from the system.

Control charts are useful tools to detect the presence of special causes of variation worthy of removal. They do so by modelling the likely performance of a process under the influence of the common causes of variation, so that the unexpected behavior (and possible non-stationarity) of the process caused by the emergence of a special cause at any time can be detected efficiently.

Shewhart Charts

When Shewhart presented his control charts, he did not claim any mathematical or statistical optimality for such charts, but he did demonstrate that the cost of controlling a process could often be reduced by using control charts. Consequently, Shewhart control charts are much more justifiable for their practical benefits than for their theoretical properties.

A Basic Chart

Bearing this in mind, a Shewhart control chart for a measurable quality characteristic is constructed in the following way. (1) Select the frequency of sampling and the sample size; e.g., take $n = 4$ observations every 2 h. (2) Calculate the sample average $\bar{X}_t$ for every time interval t (e.g., every 2 h) and plot $\bar{X}_t$ versus t for all the values of t at hand. By doing so, one obtains a run chart. (3) Add a center line (CL) to the run chart. The ordinate of this horizontal line can be a target value for the quality characteristic, a historical mean of past observations, or simply the mean of the observations at hand. (4) Add an upper control limit (UCL) and a lower control limit (LCL). These horizontal lines are usually situated symmetrically around the CL and at a distance of three times the standard deviation of the statistics plotted in the run chart (e.g., three times the standard deviation of $\bar{X}_t$).

This chart is used at every time interval t to take the decision of whether the process should be considered to be in a state of economic control or not. The usual decision rule is: (1) Decide that the process stays in control at time t if the plotted statistics ($\bar{X}_t$) lies between the UCL and the LCL, and continue plotting. (2) Declare an out of control situation otherwise; in this case, a search for an assignable cause should be started, which hopefully will eventually lead to the identification of this cause and its permanent removal from the system. This type of control procedure is sometimes called process monitoring, or process surveillance, and is a part of SPC. Figure 1 shows a Shewhart chart for a random sample of values of $\bar{X}_t$ having mean $\mu = 50$ and standard deviation $\sigma = 2$. The chart does not show any alarm.

Some Modifications of the Basic Chart

Under certain theoretical assumptions, the basic chart can claim some type of optimality. However, it may not be completely satisfactory in practice. Consequently, the form of the basic chart and how it is used can be modified in many different ways. For example, the control limits could not be symmetrically placed around the CL or could not necessarily lie at three standard deviations from the CL.

Warning limits situated at two standard deviations from the CL could also be plotted. Lines at one standard deviation from the CL could be added. The decision rule could correspondingly be modified using, for example, the so called Western Electric rules, etc.

The usefulness of these modifications of the basic chart should be judged, in each particular application, on the bases of the economical or practical advantages they provide. In doing so, the costs of declaring that the process is in control when in fact is not, and vice versa, usually play a role. The elapsed time since the process starts to be out of control until this state is detected can also play a role (true alarm), as well as the time between consecutive declarations of out of control situations when the process stays in control (false alarm rate). These elapsed random times are usually called run lengths (RLs) and their means are called average run lengths (ARLs). Clearly, the frequency distribution (or probability distribution) of the RL will depend on whether the process is in control (RL for false alarms) or out of control (RL for true alarms), and the ARL for false alarms should be much larger than the ARL for true alarms.

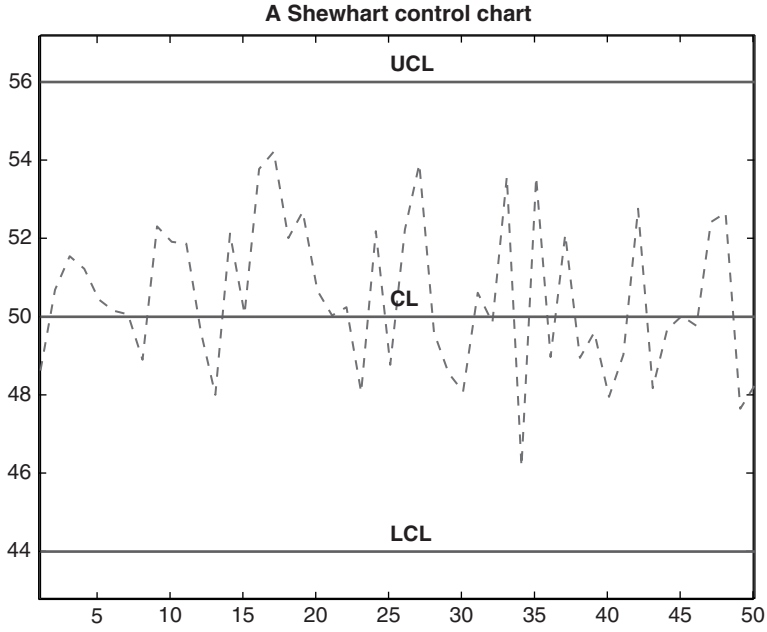
Some More Basic Charts

Control of the mean of a measurable quality characteristic is important, but a process can also be out of control because of excessive variation around its mean. Therefore, in addition to the basic $\bar{X}$ chart, previously described, it is customary to simultaneously run a chart to control the range (R-chart) or the standard deviation (S-chart) of the observations taken every time interval t.

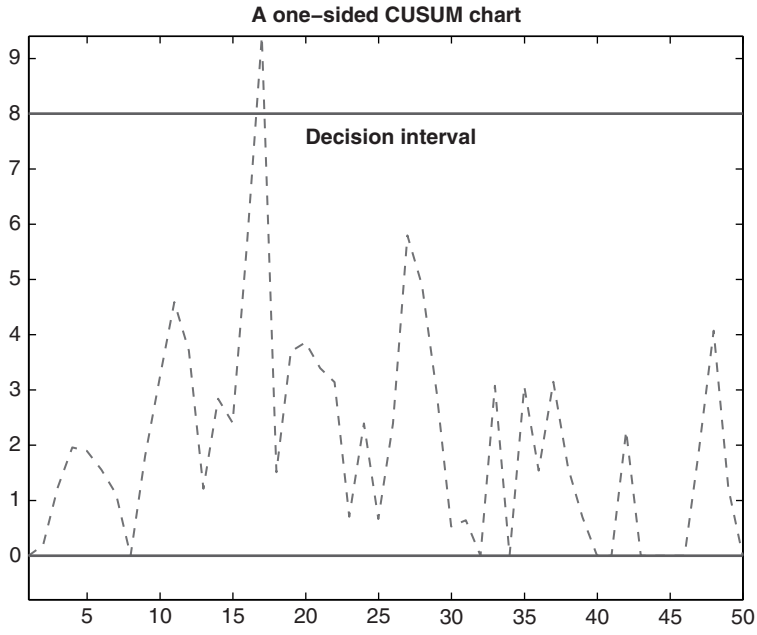
Similarly, when the quality characteristic is not measurable, one can use a p-chart or an np-chart to control the fraction nonconforming for each time interval t, or a c-chart or a u-chart to control the total numbers (counts) of nonconforming items for each period t.

Some Other Types of Control Charts

Basic Shewhart charts are useful to detect relatively large and sporadic deviations from the state of control. However, the control of a process may be jeopardized also by small but persistent deviations from the state of control. The Western Electric rules may be considered as one of many attempts to tackle this problem. However, a more formal approach was suggested by Page (1954, 1957) by introducing the cumulative sum (CUSUM) charts. Moreover, the introduction of the exponentially weighted moving average (EWMA) charts provided an alternative procedure. More recently, cumulative score (CUSCORE) charts, specialized in detecting particular types of deviation from the state of control, have also been suggested (e.g., by Box and Ramírez 1992; Box and Luceño 1997; Box et al. 2009).



Control Charts. Fig. 1 An example of a Shewhart chart



Control Charts. Fig. 2 An example of a one-sided CUSUM chart for the same data as in Fig. 1

CUSUM Charts

To be able to efficiently detect small persistent deviations from target occurring before and at period t, some use of recent observations is necessary. CUSUM charts do so by using the following statistics:

$$S_t^+ = \max\left[S_{t-1}^+ + (\bar{X}_t - k^+); 0\right];$$
$$S_t^- = \max\left[S_{t-1}^- + (-\bar{X}_t - k^-); 0\right]; \quad (1)$$

where k^+ and k^- are called reference values. The process is considered to be in control until the period t at which one of the inequalities $S_t^+ > h^+$ or $S_t^- > h^-$ becomes true, where h^+ and h^- are called decision intervals. At this time, an alarm is declared, and the search for a special cause (or assignable cause, in Deming's words) should begin.

The reference values and decision intervals of the chart are often chosen in the light of the theoretical ARLs that they produce when the process is on target and when the process is out of target by an amount of D times the standard deviation of X_t (or, equivalently, $D\sqrt{n}$ times the standard deviation of $\bar{X}_t$).

If only one of the statistics in (1) is used, the CUSUM chart is called one-sided; if both are used, the CUSUM is called two-sided. The theoretical evaluation of the run length distributions for two-sided CUSUM charts is considerably more difficult than for their one-sided counterparts. Figure 2 shows a one-sided CUSUM chart based on S_t^+, with reference value $\mu + 0.25\sigma$ and decision interval at 4σ, for the sample used in Fig. 1. This chart produces an alarm at $t = 17$.

EWMA Charts

EWMA charts use recent data in a different way than CUSUM charts. The EWMA statistic is

$$\tilde{X}_t = (1-\lambda)\tilde{X}_{t-1} + \lambda\bar{X}_t, \quad (2)$$

where $0 < \lambda < 1$, but most often $0.1 \le \lambda \le 0.4$. The EWMA statistic at time t is an average of all observations taken at time t and before, in which each observation receives a weight that decreases exponentially with its age. In other words, Eq. (2) can be written as

$$\tilde{X}_t = \lambda[\bar{X}_t + (1-\lambda)\bar{X}_{t-1} + (1-\lambda)^2\bar{X}_{t-2} + \cdots]. \quad (3)$$

The smaller the value of λ, the smoother the chart.

The process is usually considered to be in control until the period t at which $|\tilde{X}_t|$ reaches three times the standard deviation of the EWMA statistic $\tilde{X}_t$. It can be shown that the variance of $\tilde{X}_t$ is the product of the variance of $\bar{X}_t$ by a factor $\lambda[1-(1-\lambda)^{2t}]/(2-\lambda)$, where $t = 0$ is the origin of the chart. When an alarm is triggered, the search for a special cause should start.

Information about the above mentioned charts and many possible variants can be found in the bibliography that follows.

About the Author

Professor Luceño was awarded 1998 Brumbaugh Award of the American Society for Quality jointly with Professor George E.P. Box. He is a co-author (with G.E.P. Box) of the well known text *Statistical Control By Monitoring and Feedback Adjustment* (John Wiley & Sons, 1997), and (with G.E.P. Box and M.A. Paniagua-Quiñones) ñ (John Wiley & Sons, 2009). He is currently Associate Editor of *Quality Technology and Quantitative Management*, and *Quality Engineering*.

Cross References

►Acceptance Sampling
►Industrial Statistics
►Multivariate Statistical Process Control
►Six Sigma
►Statistical Quality Control
►Statistical Quality Control: Recent Advances

References and Further Reading

Box GEP, Luceño A (1997) Statistical control by monitoring and feedback adjustment. Wiley, New York
Box GEP, Ramírez JG (1992) Cumulative score charts. Qual reliab Eng Int 8:17–27
Box GEP, Luceño A, Paniagua-Quiñones MA (2009) Statistical control by monitoring and adjustment, 2nd edn. Wiley, New York
Deming WE (1986) Out of the crisis. Massachusetts Institute of Technology, Center for Advanced Engineering Studies, Cambridge
Khattree R, Rao CR (eds) (2003) Handbook of statistics 22: statistics in industry. Elsevier, Amsterdam
Luceño A, Cofiño AS (2006) The random intrinsic fast initial response of two-sided CUSUM charts. Test 15:505–524
Luceño A, Puig-Pey J (2000) Evaluation of the run-length probability distribution for CUSUM charts: assessing chart performance. Technometrics 42:411–416
Montgomery DC (2005) Introduction to statistical quality control, 5th edn. Wiley, New York
NIST/SEMATECH (2009) e-Handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/
Page ES (1954) Continuous inspection schemes. Biometrika 41: 100–114
Page ES (1957) On problems in which a change in a parameter occurs at an unknown point. Biometrika 44:248–252
Ryan TP (1989) Statistical methods for quality improvement. Wiley, New York
Ruggery F, Kenetts RS, Faltin FW (eds) (2007) Encyclopedia of statistics in quality and reliability. Wiley, New York
Shewhart WA (1931) Economic control of quality of manufacturing product. Van Nostrand Reinhold, Princeton, NJ. Republished by Quality Press, Milwaukee, 1980
Western Electronic Company (1956) Statistical quality control handbook. Western Electric Corporation, Indianapolis

Convergence of Random Variables

Pedro J. Rodríguez Esquerdo
Professor, Head
University of Puerto Rico, San Juan, Puerto Rico

C

Introduction

The convergence of a sequence of random variables (RVs) is of central importance in probability theory and in statistics. In probability, it is often desired to understand the long term behavior of, for example, the relative frequency of an event, does it converge to a number? In what sense does it converge? In statistics, a given estimator often has the property that for large samples the values it takes are distributed around and are close to the value of the desired parameter. In many situations the distribution of this estimator can be approximated by a well known distribution, which can simplify the analysis. Thus it is necessary to understand the types of convergence of such sequences, and conditions under which they occur.

Four modes of convergence are presented here.

1. *Weak convergence*, also called *convergence in distribution* or *convergence in law*, refers to the conditions under which a sequence of distribution functions converges to a cumulative distribution function (cdf).
2. A second mode is *convergence in probability*, which studies the limiting behavior of the sequence of probabilities that for each n, a RV deviates by more than a given quantity from a limiting RV.
3. *Convergence with probability one*, or almost sure convergence, studies the conditions under which the probability of a set of points in the sample space for which a sequence of RVs converges to another RV is equal to one.
4. *Convergence in the rth mean* refers to the convergence of a sequence of expected values. As it is to be expected, there are some relations between the different modes of converge.

The results here are explained, for their formal proof, the reader is referred to the included references. In general, the RVs $\{X_n\}$ cannot be assumed to be independent or identically distributed. For each value of the subscript n, the distribution of X_n may change (Casella and Berger 2002). In many cases, however, the sequence of *dfs* converge to another df.

A large amount of literature exists on the convergence of random variables. An excellent reference for understanding the definitions and relations is Rohatgi (1976). For

a discussion of some of these modes of convergence and as they apply to statistics, see Casella and Berger (2002). Chow and Teicher (1997), Loeve (1976) and Dudley (1989) present a more formal and general approach to the concept of convergence of random variables. In this paper, the notation $\{X_n\}$ is used to represent the sequence $X_1, X_2, X_3, \ldots$

Convergence in Distribution

Let $\{X_n\}$ be a sequence of RVs defined on a sample space (Ω, F, P), and let $\{F_n\}$ be the corresponding sequence of cdfs. Let X be a RV with cdf F. The sequence $\{X_n\}$ is said to *converge in distribution* to X if $\lim_{n\to\infty} F_n(x) = F(x)$ at every point where $F(x)$ is continuous. This type of convergence is sometimes also called *convergence in law* and denoted $X_n \xrightarrow{\mathcal{L}} X$. A sequence of distribution functions does not have to converge, and when it does:

1. *The limiting function does not have to be a cdf itself.* Consider the sequence given by $F_n(x) = 0$ if $x < n$ and $F_n(x) = 1$ if $x \geq n$; $n = 1, 2, \ldots$ Then, at each real value x, $F_n(x) \to 0$, which is not a cdf.
2. *Convergence in distribution does not imply that the sequence of moments converges.* For $n = 1, 2, \ldots$, consider a sequence of cdfs $\{F_n\}$ defined by $F_n(x) = 0$, if $x < 0$; $F_n(x) = 1 - 1/n$, for $0 \leq x < n$; and $F_n(x) = 1$ for $x \geq n$. The sequence of cdfs converges to the cdf $F(x) = 1$ for $x \geq 0$, and $F(x) = 0$ otherwise. For each n, the cdf F_n corresponds to a discrete RV X_n that has probability function (pf) given by $P\{X_n = 0\} = 1 - 1/n$ and $P\{X_n = n\} = 1/n$. The limiting cdf F, corresponds to a RV X with pf $P(X = 0) = 1$. For $k \geq 1$, the kth moment of X_n is $E\left(X_n^k\right) = 0(1 - 1/n) + n^k(1/n) = n^{k-1}$. Finally, $E(X^k) = 0$, so that $E\left(X_n^k\right)$ does not converge to $E(X^k)$.
3. *Convergence in distribution does not imply convergence of their pfs or probability density functions* (pdfs). Let a and b be fixed real numbers, and $\{X_n\}$ a sequence of RVs with pfs given by $P\{X_n = x\} = 1$ for $x = b + a/n$ and $P\{X_n = x\} = 0$ otherwise. None of the pfs assigns any probability to the point $x = b$. Then $P\{X_n = x\} \to 0$, which is not a pf, but the sequence of cdfs $\{F_n\}$ of the RVs X_n converges to a cdf, $F(x) = 1$ for $x \geq b$ and $F(x) = 0$ otherwise.
4. *For integer valued RVs, its sequence of pfs converges to another pf if and only if the corresponding sequence of RVs converges in distribution.*
5. *If a sequence of RVs $\{X_n\}$ converges in distribution to X and c is a real constant, then $\{X_n + c\}$, and $\{cX_n\}$ converge in distribution to $\{X + c\}$, and $\{cX\}$, respectively.*

Convergence in Probability

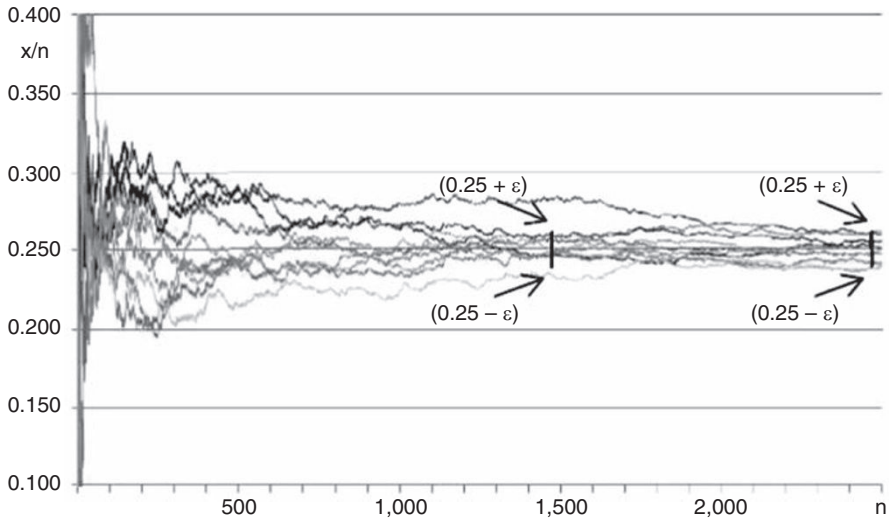
Let $\{X_n\}$ be a sequence of RVs defined on a sample space (Ω, F, P). The sequence $\{X_n\}$ is said to *converge in probability* to a RV X, denoted by $X_n \xrightarrow{p} X$, if for every real number $\varepsilon > 0$,

$$\lim_{n\to\infty} P\{|X_n - X| > \varepsilon\} = 0.$$

Convergence in probability of $\{X_n\}$ to the RV X refers to the convergence of a sequence of probabilities, real numbers to 0. It means that the probability that the distance between X_n and X is larger than $\varepsilon > 0$ tends to 0 as the n increases to infinity. It does not mean that given $\varepsilon > 0$, we can find N such that $|X_n - X| < \varepsilon$ for all $n \geq N$.

Convergence in probability, behaves in many respects as one would expect with respect to common arithmetic operations and under continuous transformations. The following results hold (Rohatgi 1976):

1. $X_n \xrightarrow{p} X$ if and only if $X_n - X \xrightarrow{p} 0$.
2. If $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$, then $P\{X = Y\} = 1$.
3. If $X_n \xrightarrow{p} X$, then $X_n - X_m \xrightarrow{p} 0$, as $n, m \to \infty$.
4. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y$, and $X_n - Y_n \xrightarrow{p} X - Y$.
5. If $X_n \xrightarrow{p} X$ and k is a real constant then $kX_n \xrightarrow{p} kX$.
6. If $X_n \xrightarrow{p} k$ then $X_n^2 \xrightarrow{p} k^2$.
7. If $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$; a, b real constants, then $X_n Y_n \xrightarrow{p} ab$.
8. If $X_n \xrightarrow{p} 1$ then $1/X_n \xrightarrow{p} 1$.
9. If $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$; a, b real constants, $b \neq 0$, then $X_n/Y_n \xrightarrow{p} a/b$.
10. If $X_n \xrightarrow{p} X$ and Y is a RV then $X_n Y \xrightarrow{p} XY$.
11. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n Y_n \xrightarrow{p} XY$.
12. Convergence in probability is stronger than convergence in distribution; that is, if $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{\mathcal{L}} X$.
13. Let k be a real number, then convergence in distribution to k implies convergence in probability to k, that is, if $X_n \xrightarrow{\mathcal{L}} k$ then $X_n \xrightarrow{p} k$.
14. In general, convergence in distribution does not imply convergence in probability. For an example, consider the identically distributed RVs $X, X_1, X_2, \ldots$ with sample space $\{0, 2\}$, such that for every n, $P(X_n = 0, X = 0) = P(X_n = 2, X = 2) = 0$ and $P(X_n = 2, X = 0) = P(X_n = 0, X = 2) = {}^1/_2$. Because X, X_n, are identically distributed RVs, $X_n \xrightarrow{\mathcal{L}} X$, but $P\{|X_n - X| > {}^1/_2\} \geq P\{|X_n - X| = 2\} = 1 \neq 0$. (Rohatgi 1976).



Ten series of 2,500 trials each, of a Binomial (4, 0.25) RV X were simulated. The ratio of the running total of successes x, to the number of trials n is plotted for each series. For X/n to converge in probability to 0.25 implies for this experiment, that as n increases, for fixed ε, the probability of observing a series outside the interval $(0.25 - \varepsilon, 0.25 + \varepsilon)$, will decrease to zero. It does not mean there is a value N such that all the series that we can possibly simulate n will be found inside the interval for all $n > N$.

Convergence of Random Variables. Fig. 1 Illustration of convergence in probability

15. Convergence in probability does not imply that the kth moments converge, that is, $X_n \xrightarrow{p} X$ does not imply that $E\left(X_n^k\right) \to E(X^k)$ for any integer $k > 0$. This is illustrated by the example in (14) above.

Figure 1 illustrates the concept of convergence in probability for series of sample means of RVs from a Binomial(4, .025) distribution. The following results further relate convergence in distribution and convergence in probability. Let $\{X_n, Y_n\}, n = 1, 2, \ldots$ be a sequence of pairs of random variables, and let c be a real number.

16. If $|X_n - Y_n| \xrightarrow{p} 0$ and $Y_n \xrightarrow{\mathcal{L}} Y$, then $X_n \xrightarrow{\mathcal{L}} Y$.
17. If $X_n \xrightarrow{\mathcal{L}} X$ and $Y_n \xrightarrow{p} c$, then $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$. This is also true for the difference $X_n - Y_n$.
18. If $X_n \xrightarrow{\mathcal{L}} X$ and $Y_n \xrightarrow{p} c$ then $X_n Y_n \xrightarrow{\mathcal{L}} cX$ (for $c \neq 0$) and $X_n Y_n \xrightarrow{p} 0$ (for $c = 0$).
19. If $X_n \xrightarrow{\mathcal{L}} X$ and $Y_n \xrightarrow{p} c$ then $X_n/Y_n \xrightarrow{\mathcal{L}} X/c$ (for $c \neq 0$).

Almost Sure Convergence

Let $\{X_n\}$ be a sequence of RVs defined on a sample space (Ω, F, P). The sequence $\{X_n\}$ is said to *converge to X with probability one* or almost surely, denoted $X_n \xrightarrow{as} X$ if

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Almost sure convergence of a sequence of RVs $\{X_n\}$ to an RV X, means that the probability of the event $\left\{\omega;\ \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}$ is one (see also ▶Almost Sure Convergence of Random Variables). That is, the set of all points ω in the sample space Ω, where $X_n(\omega)$ converges to $X(\omega)$, has probability one. It is not required that the sequence of functions $\{X_x(\omega)\}$ converge to the function $X(\omega)$ pointwise, for all ω in the sample space, only that the set of such ω has probability one.

1. Convergence almost surely implies convergence in probability. If the sequence of random variables $\{X_n\}$ converges almost surely to X then it converges in probability to X.

2. Skorokhod's representation theorem shows that if a sequence of RVs $\{X_n\}$ converges in distribution to an RV X, then there exists a sequence of random variables $\{Y_n\}$, identically distributed as $\{X_n\}$ such that $\{Y_n\}$ converges almost surely to a RV Y, which itself is identically distributed as X (Dudley 1989).
3. Continuity preserves convergence in distribution, in probability, and almost sure convergence. If X_n converges in any of these modes to X, and f is a continuous function defined on the real numbers, then $f(X_n)$ converges in the same mode to $f(X)$.
4. If $\{X_n\}$ is a strictly decreasing sequence of positive random variables, such that X_n converges in probability to 0, then X_n converges almost surely to 0.
5. Convergence in probability does not imply convergence almost surely. Consider (Casella and Berger 2002) the sample space given by the interval $[0, 1]$, and the uniform probability distribution. Consider the RV $X(\omega) = \omega$ and let $\{X_n\}$ be defined by

$$X_1(\omega) = \omega + I_{[0,1]}(\omega), \quad X_2(\omega) = \omega + I_{[0,1/2]}(\omega),$$
$$X_3(\omega) = \omega + I_{[1/2,1]}(\omega), \quad X_4(\omega) = \omega + I_{[0,1/3]}(\omega),$$
$$X_5(\omega) = \omega + I_{[1/3,2/3]}(\omega), \quad X_6(\omega) = \omega + I_{[2/3,1]}(\omega),$$

and so on. Here $I_A(\omega)$ is the indicator function of the set A. Then $\{X_n\}$ converges in probability to X, but does not converge almost surely since the value $X_n(\omega)$ alternates between ω and $\omega + 1$ infinitely often.

Convergence in the *r*th Mean

Definition Let $\{X_n\}$ be a sequence of RVs defined on a sample space (Ω, F, P) The sequence $\{X_n\}$ is said to *converge to X in the rth mean*, $r \geq 1$, if $E(|Xn|^r) < \infty$, $E(|X|^r) < \infty$ and $\lim_{n\to\infty} E\left(|X_n - X|^r\right) = 0$.

1. When $r = 1$ we say that $\{X_n\}$ converges in the mean, while for $r = 2$, we say that $\{X_n\}$ converges in the mean square.
2. If a sequence $\{X_n\}$ converges in the rth mean, and $s < r$, then $\{X_n\}$ converges in the sth mean. For example, convergence in the mean square implies convergence in the mean. This means that if the variances of a sequence converge, so do the means.
3. Convergence in the rth mean implies convergence in probability, if $\{X_n\}$ converges in the rth mean to X, then $\{X_n\}$ converges in probability to X. However, the converse is not true. For an example, consider the sequence $\{X_n\}$ with probability function defined by $P(X_n = 0) = 1 - \frac{1}{n^3}$ and $P(X_n = n) = \frac{1}{n^3}$ for $r > 0$. (Rohatgi 1976).

About the Author

Dr. Pedro J. Rodríguez Esquerdo is Professor at the Department of Mathematics, College of Natural Sciences and is currently Professor and Head, Institute of Statistics and Computer Information Science, College of Business Administration, both at the University of Puerto Rico, Rio Piedras. He is past Associate Dean for Academic Affairs at the University of Puerto Rico (1988–1993). Professor Rodríguez Esquerdo has served on several advisory boards, including the Advisory Board of Gauss Research Laboratory (2000–) in San Juan, Puerto Rico. He has been a consultant on education, technology, statistics, mathematics, and intellectual property law. He designed and maintains his statistics education web site www.eduesta.org since 1997. Prof. Rodríguez Esquerdo coauthored a book with professors Ana Helvia Quintero and Gloria Vega, *Estadística Descriptiva* (Publicaciones Puertorriqueñas, 1997), and is currently participating in a distance learning project in applied mathematics.

Cross References

►Almost Sure Convergence of Random Variables
►Binomial Distribution
►Central Limit Theorems
►Ergodic Theorem
►Glivenko-Cantelli Theorems
►Laws of Large Numbers
►Limit Theorems of Probability Theory
►Probability Theory: An Outline
►Random Variable
►Strong Approximations in Probability and Statistics
►Uniform Distribution in Statistics
►Weak Convergence of Probability Measures

References and Further Reading

Casella G, Berger RL (2002) Statistical inference. Duxbury, Pacific Grove

Chow Y, Teicher H (1997) Probability theory: independence, interchangeability, martingales, 3rd edn. Springer, New York

Dudley RM (1989) Real analysis and probability. Chapman & Hall, New York

Loeve M (1977) Probability theory, vol I, 4th edn. Springer, New York

Rohatgi VK (1976) An introduction to probability theory and mathematical statistics. Wiley, New York

Cook's Distance

R. Dennis Cook
Professor
University of Minnesota, Minneapolis, MN, USA

Introduction

Prior to 1975 there was little awareness within statistics or the applied sciences generally that a single observation can influence a statistical analysis to a point where inferences drawn with the observation included can be diametrically opposed to those drawn without the observation. The recognition that such *influential observations* do occur with notable frequency began with the 1977 publication of *Cook's Distance*, which is a means to assess the influence of individual observations on the estimated coefficients in a linear regression analysis (Cook 1977). Today the detection of influential observations is widely acknowledged as an important part of any statistical analysis and Cook's distance is a mainstay in linear regression analysis. Generalizations of Cook's distance and of the underlying ideas have been developed for application in diverse statistical contexts. Extensions of Cook's distance for linear regression along with a discussion of surrounding methodology were presented by Cook and Weisberg (1982).

Cook's distance and its direct extensions are based on the idea of contrasting the results of an analysis with and without an observation. Implementation of this idea beyond linear and ▶generalized linear models can be problematic. For these applications the related concept of *local influence* (Cook 1986) is used to study the touchiness of an analysis to local perturbations in the model or the data. Local influence analysis continues to be an area of active investigation (see, for example, Zhu et al. 2007).

Cook's Distance

Consider the linear regression of a response variable Y on p predictors $X_1, \ldots, X_p$ represented by the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \boldsymbol{\varepsilon}_i,$$

where $i = 1, \ldots, n$ indexes observations, the β's are the regression coefficients and ε is an error that is independent of the predictors and has mean 0 and constant variance σ^2. This classic model can be represented conveniently in matrix terms as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Here, $\mathbf{Y} = (Y_i)$ is the $n \times 1$ vector of responses, $\mathbf{X} = (X_{ij})$ is the $n \times (p+1)$ matrix of predictor values X_{ij}, including a constant column to account for the intercept β_0, and $\boldsymbol{\varepsilon} = (\varepsilon_i)$ is the $n \times 1$ vector of errors. For clarity, the ith response Y_i in combination with its associated values of the predictors $X_{i1}, \ldots, X_{ip}$ is called the ith *case*. Let $\widehat{\boldsymbol{\beta}}$ denote the ordinary least squares (OLS) estimator of the coefficient vector $\boldsymbol{\beta}$ based on the full data and let $\widehat{\boldsymbol{\beta}}_{(i)}$ denote the OLS estimator based on the data after removing the ith case. Let s^2 denote estimator of σ^2 based on the OLS fit of the full dataset – s^2 the residual sum of squares divided by $(n-p-1)$.

Cook (1977) proposed to assess the influence of the ith case on $\widehat{\boldsymbol{\beta}}$ by using a statistic D_i, which subsequently became known as Cook's distance, that can be expressed in three equivalent ways:

$$D_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^T \mathbf{X}^T \mathbf{X} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{(p+1)s^2} \tag{1}$$

$$= \frac{(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})^T (\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}_{(i)})}{(p+1)s^2} \tag{2}$$

$$= \frac{r_i^2}{p+1} \times \frac{h_i}{1-h_i}. \tag{3}$$

The first form (1) shows that Cook's distance measures the difference between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(i)}$ using the inverse of the contours of the estimated covariance matrix $s^2(\mathbf{X}^T\mathbf{X})^{-1}$ of $\widehat{\boldsymbol{\beta}}$ and scaling by the number of terms $(p+1)$ in the model. The second form shows that Cook's distance can be viewed also as the squared length of the difference between the $n \times 1$ vector of fitted values $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ based on the full data and the $n \times 1$ vector of fitted values $\widehat{\mathbf{Y}}_{(i)} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{(i)}$ when $\boldsymbol{\beta}$ is estimated without the ith case.

The final form (3) shows the general characteristics of cases with relatively large values of D_i. The ith *leverage* h_i, $0 \le h_i \le 1$, is the ith diagonal of the projection matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$ that puts the "hat" on $\mathbf{Y}$, $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. It measures how far the predictor values $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ for the ith case are from the average predictor value $\overline{\mathbf{X}}$. If $\mathbf{X}_i$ is far from $\overline{\mathbf{X}}$ then the ith case will have substantial pull on the fit, h_i will be near its upper bound of 1, and the second factor of (3) will be very large. Consequently, D_i will be large unless the first factor in (3) is small enough to compensate. The second factor tells us about the leverage or pull that $\mathbf{X}_i$ has on the fitted model, but it does not depend on the response and thus says nothing about the actual fit of the ith case. That goodness of fit information is provided by r_i^2 in first factor of (3): r_i is the *Studentized residual* for the ith case – the ordinary residual for the ith case divided by $s\sqrt{1-h_i}$. The squared Studentized residual r_i^2 will be large when Y_i does not fit the model and thus can be regarded as an *outlier*, but it says nothing about leverage. In short, the

first factor gives information on the goodness of the fit of Y_i, but it says nothing about leverage, while the second factor gives the leverage information but says nothing about goodness of fit. When multiplied, these factors combine to give a measure of the influence of the ith case.

The Studentized residual r_i is a common statistic for testing the hypothesis that Y_i is not an outlier. That test is most powerful when h_i is small, so $\mathbf{X}_i$ is near $\overline{\mathbf{X}}$, and least powerful when h_i is relatively large. However, leverage or pull is weakest when h_i is small and strongest when h_i is large. In other words, the ability to detect ►outliers is strongest where the outliers tend to be the least influential and weakest where the outliers tend to be the most influential. This gives another reason why influence assessment can be crucial in an analysis.

Cook's distance is not a test statistic and should not by itself be used to accept cases or reject cases. It may indicate an anomalous case that is extramural to the experimental protocol or it may indicate the most important case in the analysis, one that points to a relevant phenomenon not reflected by the other data. Cook's distance does not distinguish these possibilities.

Illustration

The data that provided the original motivation for the development of Cook's distance came from an experiment on the absorption of a drug by rat livers. Nineteen rats were given various doses of the drug and, after a fixed waiting time, the rats were sacrificed and the percentage Y of the dose absorbed by the liver was measured. The predictors were dose, body weight and liver weight. The largest absolute Studentized residual is $\max|r_i| = 2.1$, which is unremarkable when adjusting for multiple testing. The case with the largest leverage 0.85 has a modest Studentized residual of 0.80, but a relatively large Cook's distance of 0.93 – the second largest Cook's distance is 0.27. Body weight and dose have significant effects in the analysis of the full data, but there are no significant effects after the influential case is removed. It is always prudent to study the impact of cases with relatively large values of D_i and all case for which $D_i > 0.5$. The most influential case in this analysis fits both of these criteria. The rat data are discussed in Cook and Weisberg (1999) and available from the accompanying software.

Acknowledgments

Research for this article was supported in part by National Science Foundation Grant DMS-0704098.

About the Author

Dennis Cook is Full Professor, School of Statistics, University of Minnesota. He served a ten-year term as Chair of the Department of Applied Statistics, and a three-year term as Director of the Statistical Center, both at the University of Minnesota. He has served as Associate Editor of the *Journal of the American Statistical Association* (1976–1982; 1988–1991; 2002–2005), *The Journal of Quality Technology*, *Biometrika* (1991–1993), *Journal of the Royal Statistical Society, Seies B* (1992–1997) and *Statistica Sinica* (1999–2005). He is a three-time recipient of the Jack Youden Prize for Best Expository Paper in *Technometrics* as well as the Frank Wilcoxon Award for Best Technical Paper. He received the 2005 COPSS Fisher Lecture and Award. He is a Fellow of ASA and IMS, and an elected member of the ISI.

Cross References

►Influential Observations
►Regression Diagnostics
►Robust Regression Estimation in Generalized Linear Models
►Simple Linear Regression

References and Further Reading

Cook RD (1977) Detection of influential observations in linear regression. Technometrics 19:15–18. Reprinted in 2000 under the same title for the Technometrics Special 40th Anniversary Issue 42, 65–68

Cook RD (1986) Assessment of local influence (with discussion). J R Stat Soc Ser B 48:133–169

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall, London/New York. This book is available online without charge from the University of Minnesota Digital Conservancy: http://purl.umn.edu/37076

Cook RD, Weisberg S (1999) Applied regression including computing and graphics. Wiley, New York

Zhu H, Ibrahim JG, Lee S, Zhang H (2007) Perturbation selection and influence measures in local influence analysis. Ann Stat 35:2565–2588

Copulas

Carlo Sempi
Professor, Dean of the Faculty of Mathematical, Physical and Natural Sciences
Università del Salento, Lecce, Italy

Copulas were introduced by Sklar in 1959 (Sklar 1959). In a statistical model they capture the dependence structure of the random variables involved, whatever the distribution

functions of the single random variables. They also allow the construction of families of bivariate or multivariate distributions.

The definition of the notion of copula relies on those of d-box (Definition 1) and of H-volume (Definition 2). Here, and in the following, we put $\mathbb{I} := [0,1]$.

Definition 1 *Let* $\boldsymbol{a} = (a_1, a_2, \ldots, a_d)$ *and* $\mathrm{b} = (b_1, b_2, \ldots, b_d)$ *be two points in* $\overline{\mathbb{R}}^d$, *with* $0 \le a_j \le b_j \le 1$ $(j \in \{1, 2, \ldots, d\})$; *the* d-box $[\mathbf{a}, \mathbf{b}]$ *is the cartesian product*

$$[\mathbf{a}, \mathbf{b}] = \prod_{j=1}^{d} [a_j, b_j],$$

Definition 2 *For a function* $H : \overline{\mathbb{R}}^d \to \overline{\mathbb{R}}$, *the* H-volume V_H *of the* d-box $[\mathbf{a}, \mathbf{b}]$ *is defined by*

$$V_H([\mathbf{a}, \mathbf{b}]) := \sum_{\mathbf{v}} sign(\mathbf{v})\, H(\mathbf{v}),$$

where the sum is taken over the 2^d *vertices* $\mathbf{v}$ *of the box* $[\mathbf{a}, \mathbf{b}]$; *here*

$$sign(\mathbf{v}) = \begin{cases} 1, & \text{if } v_j = a_j \text{ for an even number of indices,} \\ -1, & \text{if } v_j = a_j \text{ for an odd number of indices.} \end{cases}$$

Definition 3 *A function* $C_d : \mathbb{I}^d \to \mathbb{I}$ *is a* d-copula *if*

(a) $C_d(x_1, x_2, \ldots, x_d) = 0$, *if* $x_j = 0$ *for at least one index* $j \in \{1, 2, \ldots, d\}$;

(b) *when all the arguments of* C_d *are equal to* 1, *but for the* j*-th one, then*

$$C_d(1, \ldots, 1, x_j, 1, \ldots, 1) = x_j;$$

(c) *the* V_{C_d}*-volume of every* d-box $[\mathbf{a}, \mathbf{b}]$ *is positive,* $V_{C_d}([\mathbf{a}, \mathbf{b}]) \ge 0$.

The set of d*-copulas* $(d \ge 2)$ *is denoted by* C_d; *in particular, the set of (bivariate) copulas is denoted by* C_2.

Property (c) is usually referred to as the "d-increasing property of a d-copula". Thus every copula is the restriction to the unit cube $\mathbb{I}^d$ of a distribution function that concentrates all the probability mass on $\mathbb{I}^d$ and which has uniform margins (and this may also serve as an equivalent definition).

It is possible to show that C_d is a compact set in the set of all continuous functions from $\mathbb{I}^d$ into $\mathbb{I}$ equipped with the product topology, which corresponds to the topology of pointwise convergence. Moreover, in C_d pointwise and uniform convergence are equivalent.

Every d-copula satisfies the Fréchet–Hoeffding bounds: for all $x_1, \ldots, x_d$ in $\mathbb{I}$, one has

$$W_d(x_1, \ldots, x_d) \le C(x_1, \ldots, x_d) \le M_d(x_1, \ldots, x_d), \quad (1)$$

where

$$W_d(x_1, \ldots, x_d) := \max\{0, x_1 + \cdots + x_d - d + 1\}$$
$$M_d(x_1, \ldots, x_d) := \min\{x_1, \ldots, x_d\}.$$

Also relevant is the "independence copula"

$$\Pi_d(x_1, \ldots, x_d) := \prod_{j=1}^{d} x_j.$$

While Π_d and M_d are copulas for every $d \ge 2$, W_d is a copula only for $d = 2$, although the lower bound provided by (1) is the best possible.

- Π_d is the distribution function of the random vector $\mathbf{U} = (U_1, U_2, \ldots, U_d)$ whose components are independent and uniformly distributed on $\mathbb{I}$.
- M_d is the distribution function of the vector $\mathbf{U} = (U_1, U_2, \ldots, U_d)$ whose components are uniformly distributed on $\mathbb{I}$ and such that $U_1 = U_2 = \cdots = U_d$ almost surely.
- W_2 is the distribution function of the vector $\mathbf{U} = (U_1, U_2)$ whose components are uniformly distributed on $\mathbb{I}$ and such that $U_1 = 1 - U_2$ almost surely.

The importance of copulas for the applications in statistics stems from Sklar's theorem.

Theorem 1 (Sklar 1959) *Let* H *be a* d*-dimensional distribution function with margins* $F_1, F_2, \ldots, F_d$, *and let* A_j *denote the range of* F_j, $A_j := F_j(\overline{\mathbb{R}})$ $(j = 1, 2, \ldots, d)$. *Then there exists a* d*-copula* C, *uniquely defined on* $A_1 \times A_2 \times \cdots \times A_d$, *such that, for all* $(x_1, x_2, \ldots, x_d) \in \overline{\mathbb{R}}^d$,

$$H(x_1, x_2, \ldots, x_d) = C(F_1(t_1), F_2(t_2), \ldots, F_d(t_d)). \quad (2)$$

Conversely, if $F_1, F_2, \ldots, F_d$ *are distribution functions, and if* C *is any* d*-copula, then the function* $H : \overline{\mathbb{R}}^d \to \mathbb{I}$ *defined by* (2) *is a* d*-dimensional distribution function with margins* $F_1, F_2, \ldots, F_d$.

For a compact and elegant proof of this result see (Rüschendorf 2009).

The second ("converse") part of Sklar's theorem is especially important in the construction of statistical models, since it allows to proceed in two separate steps:

- Choose the one-dimensional distribution functions F_1, $F_2, \ldots, F_d$ that describe the behavior of the individual statistical quantities (random variables) $X_1, X_2, \ldots, X_d$ that appear in the model.
- Fit these in (2) by means of a copula C that captures the dependence relations among $X_1, X_2, \ldots, X_d$.

These two steps are independent in the sense that, once a copula C has been chosen, any choice of the distribution functions $F_1, F_2, \ldots, F_d$ is possible.

It should be stressed that the copula whose existence is established in Sklar's theorem is uniquely defined only when the distribution functions have no discrete component; otherwise, there are, in general, several copulas that coincide on $A_1 \times A_2 \times \cdots \times A_d$ and which satisfy (2). This lack of uniqueness may have important consequences when dealing with the copula of random variables (see, e.g., (Marshall 1996)).

The introduction of copulas in the statistical literature has allowed an easier way to construct models by proceeding in two separate steps: (i) the specification of the marginal laws of the random variables involved and (ii) the introduction of a copula that describes the dependence structure among these variables. In many applications (mainly in Engineering) this has allowed to avoid the mathematically elegant and easy-to-deal, but usually unjustified, assumption of independence.

In view of possible applications, it is important to have at one's disposal a stock of copulas. Many families of bivariate copulas can be found in the books by Nelsen (2006), by Balakrishnan and Lai (2009) and Jaworski et al. (2010). Here we quote only the gaussian, the meta-elliptical (Fang et al. 2002) and the extreme-value copulas (Ghoudi et al. 1998). A popular family of copulas is provided by the *Archimedean* copulas, which, in the two-dimensional case, are represented in the form

$$C_\varphi(s,t) = \varphi^{[-1]}\left(\varphi(s) + \varphi(t)\right),$$

where the *generator* $\varphi : [0,1] \to [0,+\infty]$ is continuous, strictly decreasing, convex and $\varphi(1) = 0$, and $\varphi^{[-1]}$ is the *pseudo-inverse* of φ, defined by $\varphi^{[-1]}(t) := \varphi^{-1}(t)$, for $t \in [0, \varphi(0)]$, and by 0, for $t \in [\varphi(0), +\infty]$. These copulas depend on a function of a single variable, the generator φ; as a consequence, the statistical properties of a pair of random variables having C_φ as their copula are easily computed in terms of φ (Genest and MacKay 1986; Nelsen 2006). For the multivariate case the reader is referred to the paper by McNeil and Nešlehová (2009), where the generators of a such a copula are completely characterized.

Notice, however, that the choice of a symmetric copula, in particular of an Archimedean one, means that the random variables involved are exchangeable, if they have the same distribution. The effort to avoid this limitation motivates the recent great interest in the construction of nonsymmetric copulas (see, e.g., Liebscher (2008)).

It must also be mentioned that many methods of construction for copulas have been introduced; here we mention

- Ordinal sums (Mesiar and Sempi 2010);
- Shuffles of Min (Mikusiński et al. 1992) and its generalization to an arbitrary copula (Durante et al. 2009);
- The $*$-product (Darsow et al. 1992) and its generalization (Durante et al. 2007a);
- Transformations of copulas, $C_h(u,v) := h^{[-1]}\left(C(h(u),h(v))\right)$, where the function $h : \mathbb{I} \to \mathbb{I}$ is concave (Durante and Sempi 2005);
- Splicing of symmetric copulas (Durante et al. 2007b; Nelsen et al. 2008);
- Patchwork copulas (De Boets and De Meyer 2007; Durante et al. 2009);
- Gluing of copulas (Siburg and Stoimenov 2008).

A strong motivation for the development of much of copula theory in recent years has come from their applications in Mathematical Finance (see, e.g., (Embrechts et al. 2003), in Actuarial Science (Free and Valdez 1998), and in Hydrology (see, e.g., (Genest and Favre 2007; Salvadori et al. 2007)).

About the Author

Carlo Sempi received his Ph.D. in Applied Mathematics in 1974, University of Waterloo, Canada (his advisor was Professor Bruno Forte). He was Chairman of the Department of Mathematics, Università del Salento (2002–2008). Currently, he is Faculty of Mathematical, Physical and Natural Sciences, University of Salento. Professor Sempi has (co-)authored about 75 refereed papers; many of these papers are on Copulas (some written with the leading authorities on this subject (including Abe Sklar and Roger Nelsen). He was the organizer of the conference "Meeting Copulæ: the 50th anniversary," Lecce, June 2009.

Cross References

►Bivariate Distributions
►Copulas in Finance
►Copulas: Distribution Functions and Simulation
►Measures of Dependence
►Multivariate Statistical Distributions
►Multivariate Statistical Simulation
►Non-Uniform Random Variate Generations
►Quantitative Risk Management
►Statistical Modeling of Financial Markets

References and Further Reading

Balakrishnan N, Lai C-D (2009) Continuous bivariate distributions, 2nd edn. Springer, New York

Darsow W, Nguyen B, Olsen ET (1992) Copulas and Markov processes. Illinois J Math 36:600–642

De Baets B, De Meyer H (2007) Orthogonal grid constructions of copulas. IEEE Trans Fuzzy Syst 15:1053–1062

Durante F, Sempi C (2005) Copula and semicopula transforms. Int J Math Math Sci 4:645–655
Durante F, Klement EP, Quesada-Molina JJ (2007a) Remarks on two product-like constructions for copulas. Kybernetika (Prague) 43:235–244
Durante F, Kolesárová A, Mesiar R, Sempi C (2007b) Copulas with given diagonal sections: novel constructions and applications. Int J Uncertain Fuzziness Knowledge-Based Syst 15: 397–410
Durante F, Rodríguez Lallena JA, Úbeda Flores M (2009) New constructions of diagonal patchwork copulas. Inf Sci 179:3383–3391
Durante F, Sarkoci P, Sempi C (2009) Shuffles of copulas. J Math Anal Appl 352:914–921
Embrechts P, Lindskog F, McNeil A (2003) Modelling dependence with copulas and applications to risk management. In: Rachev S (ed) Handbook of heavy tailed distributions in finance, Chapter 8. Elsevier, Amsterdam, pp 329–384
Fang HB, Fang KT, Kotz S (2002) The meta-elliptical distributions with given marginals. J Multivariate Anal 82:1–16
Free EW, Valdez EA (1998) Understanding relationships using copulas. N Am J Actuar 2:1–25
Genest C, Favre AC (2007) Everything you always wanted to know about copula modeling but were afraid to ask. J Hydrol Eng 12(4):347–368
Genest C, MacKay J (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283
Ghoudi K, Khoudraji A, Rivest L-P (1998) Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles. Canad J Stat 26:187–197
Jaworski P, Durante F, Härdle W, Rychlik T (eds) (2010) Copula theory and its applications. Springer, Berlin
Liebscher E (2008) Construction of asymmetric multivariate copulas. J Multivariate Anal 99:2234–2250
Marshall AW (1996) Copulas, marginals, and joint distributions, In: Rüschendorf L, Schweizer B, Taylor MD (eds) Distributions with fixed marginals and related topics, Institute of Mathematical Statistics, Lecture Notes – Monograph Series vol 28, Hayward, pp 213–222
McNeil AJ Nešlehová J (2009) Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions. Ann Stat 37:3059–3097
Mesiar R, Sempi C (2010) Ordinal sums and idempotents of copulas. Mediterr J Math 79:39–52
Mikusiński P, Sherwood H, Taylor MD (1992) Shuffles of min. Stochastica 13:61–74
Nelsen RB (2006) An introduction to copulas, Lecture Notes in Statistics 139, 2nd edn. Springer, New York
Nelsen RB, Quesada Molina JJ, Rodríguez Lallena JA, Úbeda Flores M (2008) On the construction of copula and quasi-copulas with given diagonal sections. Insurance Math Econ 42:473–483
Rüschendorf L (2009) On the distributional transform, Sklar's theorem, and the empirical copula process. J Statist Plan Inference 139:3921–3927
Salvadori G, De Michele C, Kottegoda NT, Rosso R (2007) Extremes in nature. An approach using copulas, Water Science and Technology Library, vol 56. Springer, Dordrecht
Siburg KF, Stoimenov PA (2008) Gluing copulas. Commun Stat Theory and Methods 37:3124–3134
Sklar A (1959) Fonctions de répartition à n dimensions et leurs marges. Publ Inst Stat Univ Paris 8:229–231

Copulas in Finance

Cherubini Umberto
Associate Professor of Mathematical Finance, MatematES
University of Bologna, Bologna, Italy

C

Introduction

Correlation trading denotes the trading activity aimed at exploiting changes in correlation or more generally in the dependence structure of assets or risk factors. Likewise, correlation risk is defined as the exposure to losses triggered by changes in correlation. The copula function technique, which enables analyzing the dependence structure of a joint distribution independently from the marginal distributions, is the ideal tool to assess the impact of changes in market comovements on the prices of assets and the amount of risk in a financial position. As far as the prices of assets are concerned, copula functions enable pricing multivariate products consistently with the prices of univariate products. As for risk management, copula functions enable assessing the degree of diversification in a financial portfolio as well as the sensitivity of risk measures to changes in the dependence structure of risk factors. The concept of consistency between univariate and multivariate prices and risk factors is very similar, and actually parallel, to the problem of compatibility between multivariate probability distributions and distribution of lower dimensions. In finance, this concept is endowed with a very practical content, since it enables designing strategies involving univariate and multivariate products with the aim of exploiting changes in correlation.

Copulas and Spatial Dependence in Finance

Most of the applications of copula functions in finance are limited to multivariate problems in a cross-sectional sense (as econometricians are used to saying), or in a spatial sense (as statisticians prefer). In other words, almost all the applications have to do with the dependence structure of different variables (prices or losses in the case of finance) at the same date. The literature on applications like these is too large to be quoted here in detail, and we refer the reader to the bibliography below and to those in Bouyé et al. (2000) and Cherubini et al. (2004) for more details.

Pricing Applications

Standard asset pricing theory is based on the requirement that the prices of financial products must be such that no arbitrage opportunities can be exploited, meaning that no financial strategy can be built yielding positive return

with no risk. The price consistent with absence of arbitrage opportunities is also known as the "*fair value*" of the product. The fundamental theorem of finance states that this amounts to assert that there must exist a probability measure, called *risk-neutral* measure, under which the expected future returns on each and every asset must be zero, or, which is the same, that the prices of financial assets must be endowed with the martingale property, when measured with that probability measure. Then, the price of each asset promising some payoff in the future must be the expected value with respect to the same probability measure. This implies that if the payoff is a function of one risk factor only, the price is the expected value with respect to a univariate probability measure. If the payoff is instead a function of more than one variable, then it must be computed by taking expectations with respect to the joint distribution of the risk factors. Notice that this implies that there must be a relationship of price consistency between the prices of univariate and multivariate products, and more generally there must be *compatibility* relationships (in the proper meaning of the term in statistics) among prices. This is particularly true for derivative products promising some payments contingent on a multivariate function of several assets. These are the so-called basket derivative products, which are mainly designed on common equity stock (*equity derivatives*), or insurance policies against default of a set of counterparties (*credit derivatives*). The same structure may be used for products linked to commodities or actuarial risks. There are also products called "*hybrids*" that include different risk factors (such as market risk, i.e., the risk of market movements and default of some obligors) in the same product. For the sake of illustration, we provide here two standard examples of basket equity and credit derivatives:

Example 1 (*Altiplano Note*) These are so-called *digital* products, that is, paying a fixed sum if some event takes place at a given future date T. Typically, the event is defined as a set of stocks or market indexes, and the product pays the fixed sum if all of them are above some given level, typically specified as a percentage of the initial level. The price of this product is of course the joint *risk-neutral* probability that all the assets be above a specified level at time T : $Q(S_1(T) > K_1, S_2(T) > K_2, \ldots, S_m(T) > K_m)$, where K_i are the levels (so-called *strike* prices). Consider now that we can actually estimate the marginal distributions from the option markets, so that we can price each $Q_i(S_i(T) > K_i)$. As a result, the only reason why one wants to invest in the multivariate digital product above instead of on a set of univariate ones is to exploit changes in correlation among the assets. To put it in other terms, the value of a multivariate product can increase even if the prices of all the univariate products remain unchanged, and this may occur if the correlation increases. Copula functions are ideal tools to single out this effect.

Example 2 (Collateralized Debt Obligation (*CDO*)) Today it is possible to invest in portfolios of credit derivatives. In nontechnical terms, we can buy and sell insurance ("*protection*" is the term in market jargon) on the first x% losses on defaults of a set of obligors (called "names"). This product is called $0-x$% *equity tranche* of a portfolio of credit losses. For the sake of simplicity assume 100 names and a 0–1% equity tranche, and assume that in case of default, each loss is equal to 1. So, this tranche pays insurance the first time a default occurs (it is also called a *first-to-default* protection). Again, we can easily recover the univariate probabilities of default from other products, namely the so-called credit default swap (*CDS*) market. So, we can price the protection for every single name in the basket. The price of the *first-to-default* must then be compatible with such prices. In fact, with respect to such prices, the multivariate product is different only because it allows to invest in correlation. Again, the equity tranche can increase in value even though the values of single-insurance *CDS* for all the names remain constant, provided that the correlation of defaults increase. Even in this case, copula functions provide the ideal tool to evaluate and trade the degree of dependence of the events of default.

Risk Management

The risk manager faces the problem of measuring the exposure of the position to different risk factors. In the standard practice, he transforms the financial positions in the different assets and markets into a set of *exposures* (*buckets*, in jargon) to a set of risk factors (*mapping* process). The problem is then to estimate the joint distribution of losses $L_1, L_2, L_3, \ldots, L_k$, on these exposures and define a risk measure on this distribution. Typical measures are *Value-at-Risk* (*VaR*) and *Expected Shortfall* (*ES*) defined as

$$VaR(L_i) \equiv \inf(x : H_i(L_i) > 1-\alpha) \qquad ES \equiv E(L_i | L_i \geq VaR)$$

where $H_i(.)$ is the marginal probability distribution of loss L_i. The risk measure of the overall portfolio will analogously be

$$VaR\left(\sum_{i=1}^{k} L_i\right) \equiv \inf\left(x : H\left(\sum_{i=1}^{k} L_i\right) > 1-\alpha\right)$$
$$ES \equiv E\left(\sum_{i=1}^{k} L_i | L_i \geq VaR\right)$$

where $H(.)$ is now the probability distribution of the sum of losses. It is clear that the relationship between

univariate and multivariate risk measures is determined by the dependence structure linking the losses themselves. Again, copula functions are used to merge these risk measures together. Actually, if the max(.,.) instead of the sum were used as the aggregation operator, the risk measure would use the copula function itself as the aggregated distribution of losses.

Copula Pricing and Arbitrage Relationships

Using copula functions is very easy to recover arbitrage relationships (i.e., consistency, or compatibility relationships) among prices of multivariate assets. These relationships directly stem from links between copula functions representing the joint distribution of a set of events and those representing the joint distribution of the complement sets. A *survival copula* is defined as

$$Q(S_1 > K_1, S_2 > K_2, \ldots, S_m > K_m) = \overline{C}(1-u_1, 1-u_2, \ldots, 1-u_m)$$

The relevance of this relationship in finance is clear because it enforces the so-called put-call parity relationships. These relationships establish a consistency link between the price of products paying out if all the events in a set take place and products paying out if none of them take place. Going back to the Altiplano note above, we may provide a straightforward check of this principle.

Example 3 (Put-Call Parity of Altiplano Notes) Assume an Altiplano Note like that in Example 1, with the only difference that the fixed sum is paid if all the assets S_i are below (instead of above) the same predefined thresholds K_i. Clearly, the value of the product will be $Q(S_1(T) \leq K_1, S_2(T) \leq K_2, \ldots, S_m(T) \leq K_m)$. Given the marginal distributions, the dependence structure of this product, which could be called *put*, or *bearish*, Altiplano should be represented by a copula, while the price of the *call* or *bullish* Altiplano in Example 1 should be computed using the survival copula. It can be proved that if this is not the case, one could exploit arbitrage profits (see Cherubini and Luciano 2002; Cherubini and Romagnoli 2009).

Copulas and Temporal Dependence in Finance

So far, we have described correlation in a spatial setting. The flaw of this approach, and of copula applications to finance in general, is that no consistency link is specified, among prices with the same underlying risk factors, but payoffs at different times. We provide three examples here, two of which extend the equity and credit products cases presented above, while the third one refers to a problem arising in risk management applications. Research on this topic, as far as applications to finance are concerned, is at an early stage, and is somewhat covered in the reference bibliography below.

Example 4 (Barrier Altiplano Note) Assume an Altiplano Note with a single asset, but paying a fixed sum at the final date if the price of that asset S remains above a given threshold K on a set of different dates $\{t_1, t_2, t_3, \ldots, t_n\}$. This product can be considered multivariate just like that in Example 1, by simply substituting the concept of *temporal dependence* for that of *spatial dependence*. Again, copula functions can be used to single out the impact of changes in temporal dependence on the price of the product. For some of these products, it is not uncommon to find the so-called memory feature, according to which the payoff is paid for all the dates in the set at the first time that the asset is above the threshold.

Example 5 (Standard Collateralized Debt Obligations) (CDX, iTraxx) In the market there exist *CDO* contracts, like those described in Example 2 above, whose terms are standardized, so that they may be of interest for a large set of investors. These products include 125 "names" representative of a whole market (*CDX* for the US and *iTraxx* for Europe), and on these markets people may trade tranches buying and selling protection on 0–3%, 3–6%, and so on, according to a schedule, which is also standardized. So, for example, you may buy protection against default of the first 3% of the same 125 names, but for a time horizon of 5 or 10 years (the standard maturities are 5, 7, and 10 years). For sure you will pay more for the 10 years insurance than for the 5 years insurance on the same risk. How much more will depend on the relationship between the losses which you may incur in the first 5 years and those that you may face in the remaining 5 years. Clearly, temporal dependence cannot be avoided in this case and it is crucial in order to determine a consistency relationship between the price of insurance against losses on a term of 5 years and those on a term of 10 years. This consistency relationship is known as the *term structure* of *CDX* (or *iTraxx*) premia.

Example 6 (Temporal aggregation of risk measures) We may also think of a very straightforward problem of temporal dependence in risk management, which arises whenever we want to compute the distribution of losses over different time horizons. An instance in which this problem emerges is when one wants to apply risk measures to compare the performance of managed funds over different investment horizons. The same problem arises whenever we have to establish a dependence structure between risk factors that are measured with different time frequencies.

To take the typical example, assume you want to study the dependence structure between market risk and credit risk in a portfolio. The risk measures of market risk are typically computed on losses referred to a period of 1 or 10 days, while credit risk losses are measured in periods of months. Before linking the measures, one has to modify the time horizon of one of the two in order to match that of the other one. The typical "*square root rule*" used in the market obviously rests on the assumption of independent losses with Gaussian distribution, but this is clearly a coarse approximation of reality.

Financial Prices Dynamics and Copulas

The need to extend copulas to provide a representation of both spatial and temporal dynamics of financial prices and risk factors has led to the rediscovery of the relationship between ►copulas and the Markov process (see ►Markov Processes) that was first investigated by Darsow et al. (1992). Actually, even though the Markov assumption may seem restrictive for general applications, it turns out to be consistent with the standard *Efficient Market Hypothesis* paradigm. This hypothesis postulates that all available information must be embedded in the prices of assets, so that price innovations must be unpredictable. This leads to models of asset prices driven by independent increments, which are Markovian by construction. For these reasons, this approach was rediscovered both for pricing and financial econometrics applications (Cherubini et al. 2008, 2009; Cherubini and Romagnoli 2010; Ibragimov 2009; Chen 2009).

We illustrate here the basic result going back to Darsow et al. (1992) with application to asset prices. We assume a set of $\{S_1, S_2, \ldots, S_m\}$ assets and a set of $\{t_0, t_1, t_2, \ldots, t_n\}$ dates, and a filtered probability space generated by the prices and satisfying the usual conditions. Denote S_i^j the price of asset i at time j. First, define the product of two copulas as

$$A * B(u,v) \equiv \int_0^1 \frac{\partial A(u,t)}{\partial t}\frac{\partial B(t,v)}{\partial t}dt$$

and the extended concept of "star-product" as

$$A * B(u_1, u_2, \ldots, u_{m+n-1}) \equiv \int_0^{u_m} \frac{\partial A(u_1, u_2, \ldots, u_{m-1}, t)}{\partial t} \times \frac{\partial B(t, u_{m+1}, u_{m+2}, \ldots, u_{m+n-1})}{\partial t} dt$$

Now, Darsow et al. proved that a stochastic process S_i is a first order ►Markov chain if and only if there exists a set of bivariate copula functions $T_i^{j,j+1}, j = 1, 2, \ldots, n$, such that the dependence among $\{S_i^1, S_i^2, \ldots, S_i^n\}$ can be written as

$$G_i^j\left(u_i^1, u_i^2, \ldots, u_i^j\right) = T_i^{1,2}\left(u_i^1, u_i^2\right) * T_i^{2,3}\left(u_i^1, u_i^2\right) \ldots * T_i^{j-1,j}\left(u_i^1, u_i^2\right)$$

The result was extended to general Markov processes of order k by Ibragimov (2009). Within this framework, Cherubini et al. (2009) provided a characterization of processes with independent increments. The idea is to represent the price S^j (or its logarithm) as $S^{j-1} + Y^j$. Assume that the dependence structure between S^{j-1} and Y^j is represented by copula $C(u,v)$. Then, the dependence between S^{j-1} and S^j may be written as

$$T^{j-1,j}(u,v) = \int_0^u D_1 C\left(w, F_Y\left(F_{S,j}^{-1}(v) - F_{S,j-1}^{-1}(w)\right)\right)\mathrm{d}w$$

where D_1 represents partial derivative with respect to the first variable, $F_Y(.)$ denotes the probability distribution of the increment, and the distribution $F_{S,k}(.)$ the probability distribution of S^k. The probability distribution of S^k is obtained by taking the marginal

$$F(S^j \le s) = T^{j-1,j}(1,v) = \int_0^1 D_1 C\left(w, F_Y\left(s - F_{S,j-1}^{-1}(w)\right)\right)\mathrm{d}w$$

This is a sort of extension of the concept of convolution to the case in which the variables in the sum are not independent. Of course, the case of independent increments is readily obtained by setting $C(u,v) = uv$. The copula linking S^{j-1} and S^j becomes in this case

$$T^{j-1,j}(u,v) = \int_0^u F_Y\left(F_{S,j}^{-1}(v) - F_{S,j-1}^{-1}(w)\right)\mathrm{d}w$$

A well-known special case is

$$T^{j-1,j}(u,v) = \int_0^u \Phi\left(\Phi^{-1}(v) - \Phi^{-1}(w)\right)\mathrm{d}w$$

with $\Phi(x)$ the standard normal distribution, which yields the dependence structure of a Brownian motion (see ►Brownian Motion and Diffusions) upon appropriate standardization. As for pricing applications, Cherubini et al. (2008) applied the framework to temporal dependence of losses and the term structure of *CDX* premia, and Cherubini and Romagnoli (2010) exploited the model to price barrier Altiplanos. This stream of literature, which applies copulas to modeling stochastic processes in discrete time, casts a bridge to a parallel approach, that directly applies copulas to model dependence among

stochastic processes in continuous time: this is the so-called Lévy copula approach (Kallsen and Tankov 2006). Both these approaches aim at overcoming the major flaw of copula functions as a static tool and unification of them represents the paramount frontier issue in this important and promising field of research.

About the Author

Umberto Cherubini is Associate Professor of Mathematical Finance at the University of Bologna. He is the author or coauthor of about 50 papers and is the coauthor of 5 books, of which two are in Italian and *Copula Methods in Finance* (with E. Luciano and W. Vecchiato, 2004), *Structured Finance: The Object Oriented Approach* (with G. Della Lunga, 2007), and *Fourier Methods in Finance* (with G. Della Lunga, S. Mulinacci and P. Rossi, 2010), all with John Wiley, Finance Series. Chichester, UK.

Cross References

►Copulas
►Copulas: Distribution Functions and Simulation
►Quantitative Risk Management
►Statistical Modeling of Financial Markets

References and Further Reading

Bouyé E, Durrleman V, Nikeghbali A, Riboulet G, Roncalli T (2000) Copulas for finance: a reading guide and some applications. Groupe de Recherche, Opérationnelle, Crédit Lyonnais. Working paper

Chen X, Wu SB, Yi Y (2009) Efficient estimation of copula-based semiparametric Markov models. Ann Stat 37(6B):4214–4253

Cherubini U, Luciano E (2002) Bivariate option pricing with copulas. Appl Math Finance 9:69–86

Cherubini U, Romagnoli S (2009) Computing the volume of n-dimensional copulas. Appl Math Finance 16(4):307–314

Cherubini U, Romagnoli S (2010) The dependence structure of running maxima and minima: results and option pricing applications. Mathematical Finance 20(1):35–58

Cherubini U, Luciano E, Vecchiato W (2004) Copula methods in finance. Wiley Finance Series, Chichester

Cherubini U, Mulinacci S, Romagnoli S (2008) A copula-based model of the term structure of CDO tranches. In: Hardle WK, Hautsch N, Overbeck L (eds) Applied quantitative finance. Springer, Berlin, pp 69–81

Cherubini U, Mulinacci S, Romagnoli S (2009) A copula-based model of speculative price dynamics. University of Bologna, Berlin

Darsow WF, Nguyen B, Olsen ET (1992) Copulas and Markov processes. Illinois J Math 36:600–642

Embrechts P, Lindskog F, McNeil AJ (2003) Modelling dependence with copulas with applications to risk management. In: Rachev S (ed) A handbook of heavy tailed distributions in finance. Elsevier, Amsterdam, pp 329–384

Gregory J, Laurent JP (2005) Basket defaults, CDOs and factor copulas. J Risk 7(4):103–122

Ibragimov R (2005) Copula based characterization and modeling for time series. Econom Theory 25:819–846

Kallsen J, Tankov P (2006) Characterization of dependence of multidimensional Lévy processes using Lévy copulas. J Multivariate Anal 97:1151–1172

Li D (2000) On default correlation: a copula function approach. J Fixed Income 9:43–54

Patton A (2002) Modelling asymmetric exchange rate dependence. Int Econ Rev 47(2):527–556

Rosemberg J (2003) Non parametric pricing of multivariate contingent claims federal reserve bank of New York staff reports, n. 162

Van Der Goorberg R, Genest C, Werker B (2005) Bivariate option pricing using dynamic copula models. Insur Math Econ 37:101–114

Copulas: Distribution Functions and Simulation

Pranesh Kumar
Professor
University of Northern British Columbia, Prince George, BC, Canada

Introduction

In multivariate data modelling for an understanding of stochastic dependence the notion of correlation has been central. Although correlation is one of the omnipresent concepts in statistical theory, it is also one of the most misunderstood concepts. The confusion may arise from the literary meaning of the word to cover any notion of dependence. From mathematics point of view, correlation is only one particular measure of stochastic dependence. It is the canonical measure in the world of ►multivariate normal distributions and in general for spherical and elliptical distributions. However empirical research in many applications indicates that the distributions of the real world seldom belong to this class. We collect and present ideas of copula functions with applications in statistical probability distributions and simulation.

Dependence

We denote by (X, Y) a pair of real-valued nondegenerate random variables with finite variances σ_x^2 and σ_y^2 respectively. The correlation coefficient between X and Y is the standardized covariance σ_{xy}, i.e., $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, $\rho \in [-1, 1]$. The correlation coefficient is a measure of linear depen-hh dence only. In case of independent random variables, correlation is zero. Embrechts, McNeil and Straumann (1999) have discussed that in case of imperfect linear dependence, i.e., $-1 < \rho < 1$, misinterpretations of correlation are possible. Correlation is not ideal for a dependence measure and

causes problems when there are heavy-tailed distributions. Independence of two random variables implies they are uncorrelated but zero correlation does not in general imply independence. Correlation is not invariant under strictly increasing linear transformations. Invariance property is desirable for the statistical estimation and significance testing purposes. Further correlation is sensitive to ▶outliers in the data set. The popularity of linear correlation and correlation based models is primarily because it is often straightforward to calculate and manipulate them under algebraic operations. For many ▶bivariate distributions it is simple to calculate variances and covariances and hence the correlation coefficient. Another reason for the popularity of correlation is that it is a natural measure of dependence in multivariate normal distributions and more generally in multivariate spherical and elliptical distributions. Some examples of densities in the spherical class are those of the multivariate t-distribution and the ▶logistic distribution.

Another class of dependence measures is rank correlations. They are defined to study relationships between different rankings on the same set of items. Rank correlation measures the correspondence between two rankings and assess their significance. Two commonly used measures of concordance are Spearman's rank correlation (ρ_s) and Kendall's rank correlation (τ). Assuming random variables X and Y have distribution functions F_1 and F_2 and joint distribution function F, Spearman's rank correlation $\rho_s = \rho(F_1(X), F_2(Y))$ where ρ is the linear correlation coefficient. If (X_1, Y_1) and (X_2, Y_2) are two independent pairs of random variables from the distribution function F, then the Kendall's rank correlation is $\tau = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0]$. The main advantage of rank correlations over ordinary linear correlation is that they are invariant under monotonic transformations. However rank correlations do not lend themselves to the same elegant variance–covariance manipulations as linear correlation does since they are not moment-based.

A measure of dependence like linear correlation summarizes the dependence structure of two random variables in a single number. Scarsini (1984) has detailed properties of copula based concordance measures. Another excellent discussion of dependence measures is by Embrecht et al. (1999). Let $D(X, Y)$ be a measure of dependence which assigns a real number to any real-valued pair of random variables (X, Y). Then dependence measure $D(X, Y)$ is desired to have properties: (i) Symmetry: $D(X, Y) = D(Y, X)$; (ii) Normalization: $-1 \leq D(X, Y) \leq +1$; (iii) Comonotonic or Countermonotonic: The notion of comonotonicity in probability theory is that a random vector is comonotonic if and only if all marginals are non-decreasing functions (or non-increasing functions) of the same random variable. A measure $D(X, Y)$ is comonotonic if $D(X, Y) = 1 \iff X, Y$ or countermonotonic if $D(X, Y) = -1 \iff X, Y$; (iv) For a transformation T strictly monotonic on the range of X, $D(T(X), Y) = D(X, Y)$, $T(X)$ increasing or $D(T(X), Y) = -D(X, Y)$, $T(X)$ decreasing.

Linear correlation ρ satisfies properties (i) and (ii) only. Rank correlations fulfill properties (i)–(iv) for continuous random variables X and Y. Another desirable property is: (v) $D(X, Y) = 0 \iff X, Y$ (Independent). However it contradicts property (iv). There is no dependence measure satisfying properties (iv) and (v). If we desire property (v), we should consider dependence measure $0 \leq D^*(X, Y) \leq +1$. The disadvantage of all such dependence measures $D^*(X, Y)$ is that they can not differentiate between positive and negative dependence (Kimeldorf and Sampson 1978; Tjøstheim 1996).

Copulas

▶Copulas have recently emerged as a means of describing joint distributions with uniform margins and a tool for simulating data. They express joint structure among random variables with any marginal distributions. With a copula we can separate the joint distribution into marginal distributions of each variable. Another advantage is that the conditional distributions can be readily expressed using the copula. An excellent introduction of copulas is presented in Joe (1997) and Nelsen (2006). Sklar's theorem (1959) states that any multivariate distribution can be expressed as the k-copula function $C(u_1, \ldots, u_i, \ldots, u_k)$ evaluated at each of the marginal distributions. Copula is not unique unless the marginal distributions are continuous. Using probability integral transform, each continuous marginal $U_i = F_i(x_i)$ has a uniform distribution (see ▶Uniform Distribution in Statistics) on $I \in [0, 1]$ where $F_i(x_i)$ is the cumulative integral of $f_i(x_i)$ for the random variable $X_i \in (-\infty, \infty)$. The k-dimensional probability distribution function F has a unique copula representation $F(x_1, x_2, \ldots, x_k) = C(F_1(x_1), F_2(x_2), \ldots, F_k(x_k)) = C(u_1, u_2, \ldots, u_k)$. The joint probability density function is written as $f(x_1, x_2, \ldots, x_k) = \Pi_{i=1}^{k} f_i(x_i) \times c(F_1(x_1), F_2(x_2), \ldots, F_k(x_k))$ where $f_i(x_i)$ is each marginal density and coupling is provided by copula density $c(u_1, u_2, \ldots, u_k) = \partial^k C(u_1, u_2, \ldots, u_k)/\partial u_1 \partial u_2 \ldots \partial u_k$ if it exists. In case of independent random variables, copula density $c(u_1, u_2, \ldots, u_k)$ is identically equal to one. The importance of the above equation $f(x_1, x_2, \ldots, x_k)$

is that the independent portion expressed as the product of the marginals can be separated from the function $c(u_1, u_2, \ldots, u_k)$ describing the dependence structure or shape. The dependence structure summarized by a copula is invariant under increasing and continuous transformations of the marginals. This means that suppose we have a probability model for dependent insurance losses of various kinds. If our interest now lies in modelling the logarithm of these losses, the copula will not change, only the marginal distributions will change.

The simplest copula is independent copula $\Pi := C(u_1, u_2, \ldots, u_k) = u_1 u_2 \ldots u_k$ with uniform density functions for independent random variables. Another copula example is the Farlie–Gumbel–Morgenstern (FGM) bivariate copula. The general system of FGM bivariate distributions is given by $F(x_1, x_2) = F_1(x_1) \times F_2(x_2) [1 + \rho (1 - F_1(x_1)) (1 - F_2(x_2))]$ and copula associated with this distribution is a FGM bivariate copula $C(u, v) = uv[1 + \rho (1 - u) (1 - v)]$. A widely used class of copulas is Archimedean copulas which has a simple form and models a variety of dependence structures. Most of the Archimedean copulas have closed-form solutions. To define an Archimedean copula, let ϕ be a continuous strictly decreasing convex function from $[0,1]$ to $[0, \infty]$ such that $\phi(1) = 0$ and $\phi(0) = \infty$. Let ϕ^{-1} be the pseudo inverse of ϕ. Then a k-dimensional Archimedean copula is $C(u_1, u_2, \ldots, u_k) = \phi^{-1} [\phi(u_1) + \ldots + \phi(u_k)]$. The function ϕ is known as a generator function. Thus any generator function satisfying $\phi(1) = 0$; $\lim_{x \to 0} \phi(x) = \infty$; $\phi'(x) < 0$; $\phi''(x) > 0$ will result in an Archimedean copula. For an example, generator function $\phi(t) = \left(t^{-\theta} - 1\right)/\theta$, $\theta \in [-1, \infty) \backslash \{0\}$ results in the bivariate Clayton copula $C(u_1, u_2) = \max\left[\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}, 0\right]$. The copula parameter θ controls the amount of dependence between X_1 and X_2.

The Frécht–Hoeffding bounds for copulas: The lower bound for k-variate copula is $W(u_1, u_2, \ldots, u_k) := \max\{1 - n + \sum_{i=1}^{k} u_i, 0\} \leq C(u_1, u_2, \ldots, u_k)$. The upper bound for k-variate copula is $C(u_1, u_2, \ldots, u_k) \leq \min_{i \in \{1,2,\ldots,k\}} u_i =: M(u_1, u_2, \ldots, u_k)$. For all copulas, the inequality $W(u_1, \ldots, u_k) \leq C(u_1, \ldots, u_k) \leq M(u_1, \ldots, u_k)$ is satisfied. This inequality is well known as the Frécht-Hoeffding bounds for copulas. Further, W and M are copulas themselves. It may be noted that the Frécht-Hoeffding lower bound is not a copula in dimension $k > 2$. Copulas M, W and Π have important statistical interpretations (Nelson, 2006). Given a pair of continuous random variables (X_1, X_2), (i) copula of (X_1, X_2) is $M(u_1, u_2)$ if and only if each of X_1 and X_2 is almost surely increasing function of the other; (ii) copula of (X_1, X_2) is $W(u_1, u_2)$ if and only if each of X_1 and X_2 is almost surely decreasing function of the other and (iii) copula of (X_1, X_2) is $\Pi(u_1, u_2) = u_1 u_2$ if and only if X_1 and X_2 are independent.

Three famous measures of concordance Kendall's τ, Spearman's ρ_s and Gini's index γ could be expressed in terms of copulas (Schweizer and Wolff 1981) $\tau = 4 \int \int_{I^2} C(u_1, u_2)\, dC(u_1, u_2) - 1$, $\rho_s = 12 \int \int_{I^2} u_1 u_2\, dC(u_1, u_2) - 3$ and $\gamma = 2 \int \int_{I^2} (|u_1 + u_2 - 1| - |u_1 - u_2|)\, dC(u_1, u_2)$. It may however be noted that the linear correlation coefficient ρ cannot be expressed in terms of copula.

The tail dependence indexes of a multivariate distribution describe the amount of dependence in the upper right tail or lower left tail of the distribution and can be used to analyze the dependence among extremal random events. Tail dependence describes the limiting proportion that one margin exceeds a certain threshold given that the other margin has already exceeded that threshold. Joe (1997) defines tail dependence: If a bivariate copula $C(u_1, u_2)$ is such that $\lambda_U := \lim_{u \to 1} \frac{[1 - 2u + C(u,u)]}{(1-u)}$ exists, then $C(u_1, u_2)$ has upper tail dependence for $\lambda_U \in (0,1]$ and no upper tail dependence for $\lambda_U = 0$. Similarly lower tail dependence in terms of copula is defined $\lambda_L := \lim_{u \to 0} \frac{C(u,u)}{u}$. Copula has lower tail dependence for $\lambda_L \in (0,1]$ and no lower tail dependence for $\lambda_L = 0$. This measure is extensively used in extreme value theory. It is the probability that one variable is extreme given that other is extreme. Tail measures are copula-based and copula is related to the full distribution via quantile transformations, i.e., $C(u_1, u_2) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right)$ for all $u_1, u_2 \in (0,1)$ in the bivariate case.

Simulation

Simulation in statistics has a pivotal role in replicating and analysing data. Copulas can be applied in simulation and Monte Carlo studies. Johnson (1987) discusses methods to generate a sample from a given joint distribution. One such method is a recursive simulation using the univariate conditional distributions. The conditional distribution of U_i given first $i - 1$ components is $C_i(u_i | u_1, \ldots, u_{i-1}) = \frac{\partial^{i-1} C_i(u_1, \ldots, u_i)}{\partial u_1 \ldots \partial u_{i-1}} \Big/ \frac{\partial^{i-1} C_{i-1}(u_1, \ldots, u_{i-1})}{\partial u_1 \ldots \partial u_{i-1}}$. For $k \geq 2$, procedure is as follows: (i) Simulate a random number $u_1$from Uniform (0,1); (ii) Simulate value u_2 from the conditional distribution $C_2(u_2|u_1)$;(iii) Continue in this way; (iv) Simulate a value u_k from $C_k(u_k|u_1, \ldots, u_{k-1})$.

We list some important contributions in the area of copulas under the reference section.

Acknowledgments

Author wish to thank the referee for the critical review and useful suggestions on the earlier draft. This work was

supported by the author's discovery grant from the *Natural Sciences and Engineering Research Council of Canada (NSERC)* which is duly acknowledged.

About the Author

Dr. Pranesh Kumar is Professor of Statistics in the University of Northern British Columbia, Prince George, BC, Canada. He has held several international positions in the past: Professor and Head, Department of Statistics, University of Transkei, South Africa; Associate Professor, Bilkent University, Ankara, Turkey; Associate Professor, University of Dar-es-Salaam, Tanzania; Visiting Professor, University of Rome, Italy; Visiting Senior Researcher at the Memorial University of Newfoundland, Canada; Associate Professor, Indian Agricultural Statistics Research Institute, New Delhi. Dr. Kumar has published his research in many prestigious professional journals. He holds membership in several professional international societies including the International Statistical Institute. Dr. Kumar have membership of the editorial boards of the *Journal of Applied Mathematics and Analysis*, *Journal of Mathematics Research*, *Journal of the Indian Society of Agricultural Statistics and JNANABHA* which has reciprocity agreement with the American Mathematical Society. He has (co-)authored about 75 refereed papers.

Cross References

►Copulas
►Copulas in Finance
►Multivariate Statistical Distributions
►Multivariate Statistical Simulation

References and Further Reading

Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65: 141–151

Cuadras CM, Fortiana J, Rodríguez Lallena JA (2002) Distributions with given marginals and statistical modelling. Kluwer, Dodrecht

Embrechts P, McNeil A, Straumann D (1997) Correlation and dependence in risk management: properties and pitfalls. Risk 12(5):69–71

Fang K-T, Kotz S, Ng K-W (1987) Symmetric multivariate and related distributions. Chapman & Hall, London

Frank MJ (1979) On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$. Aequationes Mathematicae 19:194–226

Frécht M (1951) Sue les tableaux de corrélation dont les marges son données. Ann Univ Lyon Sect A 9:53–77

Genest C (1987) Frank's family of bivariate distributions. Biometrika 74:549–555

Genest C, Mackay J (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283

Genest C, Rivest L (1993) Statistical inference procedures for bivariate Archimedean copulas. J Am Stat Assoc 88:1034–1043

Genest C, Ghoudi K, Rivest L (1995) A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82:543–552

Gumbel EJ (1960) Bivariate exponential distributions. J Am Stat Assoc 55:698–707

Hougaard P (1986) A class of multivariate failure time distributions. Biometrika 73:671–678

Hutchinson TP, Lai CD (1990) Continuous bivariate distributions emphasizing applications. Rumsby Scientific, Adelaide, South Australia

Joe H (1997) Multivariate models and dependent concepts. Chapman & Hall, New York

Johnson ME (1987) Multivariate statistical simulation. Wiley, New York

Kimeldorf G, Sampson AR (1978) Monotone dependence. Ann Stat 6:895–903

Marshall AW, Olkin I (1988) Families of multivariate distributions. J Am Stat Assoc 83:834–841

Nelsen R (2006) An introduction to copulas. Springer, New York

Nelsen RB, Quesada Molina JJ, Rodríguez Lallena JA, Úbeda Flores M (2001) Bounds on bivariate distribution functions with given margins and measures of association. Commun Stat Theory Meth 30:1155–1162

Scarsini M (1984) On measures of concordance. Stochastica 8:201–219

Schweizer B, Sklar A (1983) Probabilistic metric spaces. North Holland, New York

Schweizer B, Wolff E (1981) On nonparametric measures of dependence for random variables. Ann Stat 9:879–885

Sklar A (1959) Fonctions de répartition án dimensional et leurs marges. Publ Inst Stat Univ Paris 8:229–231

Tjøstheim D (1996) Measures of dependence and tests of independence. Statistics 28:249–284

Cornish–Fisher Expansions

Vladimir V. Ulyanov
Professor
Lomonosov Moscow State University, Moscow, Russia

Introduction

In statistical inference it is of fundamental importance to obtain the sampling distribution of statistics. However, we often encounter situations where the exact distribution cannot be obtained in closed form, or even if it is obtained, it might be of little use because of its complexity. One practical way of getting around the problem is to provide reasonable approximations of the distribution function and its quantiles, along with extra information on their possible errors. This can be accomplished with the help of Edgeworth and Cornish–Fisher expansions. Recently, interest in Cornish–Fisher expansions has increased

because of intensive study of VaR (Value at Risk) models in financial mathematics and financial risk management (see Jaschke (2002)).

Expansion Formulas

Let X be a univariate random variable with a continuous distribution function F. For $\alpha : 0 < \alpha < 1$, there exists x such that $F(x) = \alpha$, which is called the (lower) $100\alpha\%$ *point* of F. If F is strictly increasing, the inverse function $F^{-1}(\cdot)$ is well defined and the $100\alpha\%$ point is uniquely determined. We also speak of "quantiles" without reference to particular values of α meaning the values given by $F^{-1}(\cdot)$.

Even in the general case, when $F(x)$ is not necessarily continuous nor is it strictly increasing, we can define its inverse function by the formula

$$F^{-1}(u) = \inf\{x; F(x) > u\}.$$

This is a right-continuous nondecreasing function defined on the interval (0,1) and $F(x_0) \geq u_0$ if $x_0 = F^{-1}(u_0)$.

Let $F_n(x)$ be a sequence of distribution functions and let each F_n admit the ▶Edgeworth expansion (EE) in the powers of $\epsilon = n^{-1/2}$ or n^{-1}:

$$\begin{aligned} F_n(x) &= G_{k,n}(x) + O(\epsilon^k) \quad \text{with} \\ G_{k,n}(x) &= G(x) + \{\epsilon a_1(x) + \cdot + \epsilon^{k-1} a_{k-1}(x)\} g(x), \end{aligned} \tag{1}$$

where $g(x)$ is the density function of the limiting distribution function $G(x)$. An important approach to the problem of approximating the quantiles of F_n is to use their asymptotic relation to those of G's. Let x and u be the corresponding quantiles of F_n and G, respectively. Then we have

$$F_n(x) = G(u). \tag{2}$$

Write $x(u)$ and $u(x)$ to denote the solutions of (2) for x in terms of u and u in terms of x, respectively [i.e. $u(x) = G^{-1}(F_n(x))$ and $x(u) = F_n^{-1}(G(u))$]. Then we can use the EE (1) to obtain formal solutions $x(u)$ and $u(x)$ in the form

$$x(u) = u + \epsilon b_1(u) + \epsilon^2 b_2(u) + \cdots \tag{3}$$

and

$$u(x) = x + \epsilon c_1(x) + \epsilon^2 c_2(x) + \cdots. \tag{4}$$

Cornish and Fisher (1937) and Fisher and Cornish (1946) obtained the first few terms of these expansions when G is the standard normal distribution function (i.e., $G = \Phi$). We call both (3) and (4) the *Cornish–Fisher expansions*, (CFE). Concerning CFE for random variables obeying limit laws from the family of Pearson distributions see Bol'shev (1963). Hill and Davis (1968) gave a general algorithm for obtaining each term of CFE when G is an analytic function:

Theorem 1 *Assume that the distribution function G is analytic. Then the following relation for x and u satisfying $F_n(x) = G(u)$ holds:*

$$x = u - \sum_{r=1}^{\infty} \frac{1}{r!} \{-[g(u)]^{-1} d_u\}^{r-1} \left[\{z_n(u)\}^r / g(u)\right], \tag{5}$$

where $d_u = d/du$ and $z_n(u) = F_n(u) - G(u)$.

A similar relation can be written for u as a function of x.

In many statistical applications, $F_n(x)$ is known to have an asymptotic expansion of the form

$$F_n(x) = G(x) + g(x)\left[n^{-a} p_1(x) + n^{-2a} p_2(x) + \cdots\right],$$

where $p_r(x)$ may be polynomials in x and $a = 1/2$ or 1. Then the formulas (5) can be written as

$$x = u - \sum_{r=1}^{\infty} \frac{1}{r!} d_{(r)} \{g_n(u)\}^r, \tag{6}$$

where $q_n(u) = n^{-a} p_1(u) + n^{-2a} p_2(u) + \cdots$,
$m(x) = -g'(x)/g(x)$,

$d_{(1)}$ = the identity operator,
$d_{(r)} = \{m(u) - d_u\}\{2m(u) - d_u\}\cdots\{(r-1)m(u) - d_u\}$,
$r = 2, 3, \ldots$

The rth term in (6) is of order $O(n^{-ra})$.

It is a tedious process to rewrite (6) in the form of (3) and to express the adjustment terms $b_k(u)$ directly in terms of the cumulants (see Hill and Davis (1968)). Lee and Lin developed a recurrence formula for $b_k(u)$, which is implemented in the algorithm AS269 (see Lee and Lin (1992, 1993)).

Usually the CFE are applied in the following form with $k = 1, 2$, or 3:

$$x_k(u) = u + \sum_{j=1}^{k-1} \epsilon^j b_j(u) + O(\epsilon^k), \tag{7}$$

In order to find the explicit expressions for $b_1(u)$ and $b_2(u)$ we substitute (7) with $k = 2$ to (1) and using (2) we have

$$\begin{aligned} F_n(x) &= F_n(u + \epsilon b_1 + \epsilon^2 b_2 + \cdots) \\ &= G(u + \epsilon b_1 + \epsilon^2 b_2) + g(u + \epsilon b_1 + \epsilon^2 b_2) \\ &\quad \times \{\epsilon a_1(u + \epsilon b_1) + \epsilon^2 a_2(u)\} + O(\epsilon^2). \end{aligned}$$

By Taylor's expansions for G, g, and a_1, we obtain

$$\begin{aligned} F_n(x) =& G(u) + \epsilon g(u)\{b_1 + a_1(u)\} \\ &+ \epsilon^2 \left[g(u) b_2 + \frac{1}{2} g'(u) b_1^2 + g(u) a_1'(u) b_1 \right. \\ &\left. + g(u) a_2(u) + g'(u) b_1 a_1(u)\right] + O(\epsilon^3), \end{aligned}$$

which should be $G(u)$. Therefore,

$$b_1 = -a_1(u),$$
$$b_2 = \frac{1}{2}\{g'(u)/g(u)\}a_1^2(u) - a_2(u) + a_1'(u)a_1(u).$$

An application of general formulas (6) in the case of normal limit distribution see the entry ►Edgeworth Expansion.

Suppose that

$$F_n(x) = G_f(x) + \frac{f\gamma}{4n}[G_f(x) + 2G_{f+2}(x) + G_{f+4}(x)] + \frac{f}{960n^2}\sum_{j=0}^{4}(-1)^j c_j G_{f+2j}(x) + o(n^{-2}),$$

where $G_f(x) = \Pr\{\chi_f^2 \le x\}$; that is, the distribution function of the ►chi-square distribution with f degrees of freedom, γ is a constant, and c_j are constants such that $\sum_{j=0}^{4}(-1)^j c_j = 0$. Then $G(u) = G_f(u)$,

$$g(u) = g_f(u) = \left[\Gamma(f/2)2^{f/2}\right]^{-1} u^{f/2-1}\exp(-u/2),$$
$$m(u) = -g_f'(u)/g_f(u) = \frac{1}{2} - \frac{1}{u}\left(\frac{f}{2} - 1\right).$$

Thus, we can write

$$q_n(u) = -\frac{u(u-f-2)}{2n(f+2)} - \frac{u}{48n^2(f+2)(f+4)(f+6)}[c_4u^3 + (c_4 - c_3)(f+6)u^2 + (c_1 - c_0)(f+4)(f+6)u - c_0(f+2)(f+4)(f+6)] + o(n^{-2}).$$

Therefore, we obtain

$$x = u - q_n(u) - [\gamma^2 u(u-f-2)/\{16n^2(f+2)^2\}] \times \{u^2 - 2(f+4)u + (f+2)^2\} + o(n^{-2}).$$

The upper and lower bounds for the quantiles $x = x(u)$ and $u = u(x)$, satisfying the equation (2), i.e.

$$\underline{x}_n(u) \ge x(u) \ge \bar{x}_n(u), \quad \underline{u}_n(x) \ge u(x) \ge \bar{u}_n(x)$$

were obtained for some special distributions by Wallace (1959).

Validity of Cornish–Fisher Expansions

In applications, the CFE are usually used in the form (7). It is necessary to remember that the approximations for α-quantiles provided by the CFE

(i) become less and less reliable for $\alpha \to 0$ and $\alpha \to 1$;
(ii) do not necessarily improve (converge) for a fixed F_n and increasing order of approximation k.

Let x_α and x_α^* be the upper $100\alpha\%$ points of F_n and $G_{k,n}$ from (1), respectively; that is, they satisfy

$$F_n(x_\alpha) = G_{k,n}\left(x_\alpha^*\right) = 1 - \alpha.$$

The approximate quantile x_α^* based on the Edgeworth expansion is available in numerical form but cannot be expressed in explicit form. Suppose that the remainder term, $R_{k,n}(x) = F_n(x) - G_{k,n}(x)$, is such that

$$|R_{k,n}| \le \epsilon^n C_k.$$

Then

$$\left|F_n\left(x_\alpha^*\right) - (1-\alpha)\right| = \left|F_n\left(x_\alpha^*\right) - G_{k,n}\left(x_\alpha^*\right)\right| \le \epsilon^n C_k.$$

This gives an error bound for the absolute differences between the probabilities based on the true quantiles and their approximations.

The other validity of the CFE was obtained by considering the distribution function $\tilde{F}_{k,n}$ of

$$\tilde{X} = U + \sum_{j=1}^{k-1}\epsilon^j b_j(U),$$

where U is the standard normal variable. Takeuchi and Takemura (1988) showed that if $|F_n(x) - G_{k,n}(x)| = o(\epsilon^{k-1})$, then $|F_n(x) - \tilde{F}_{k,n}| = o(\epsilon^{k-1})$.

Function of Sample Mean

Usually the conditions that are sufficient for validity of EE are sufficient as well for validity of CFE. Under the conditions of section "►Function of Sample Means" in the entry ►Edgeworth Expansion and in its notation we have (see Hall (1992)):

$$\sup_{\epsilon<\alpha<1-\epsilon}\left|x_\alpha - u_\alpha - \sum_{j=1}^{k-2}\frac{b_j(u_\alpha)}{n^{j/2}}\right| = o\left(\frac{1}{n^{(k-2)/2}}\right),$$

where $x_\alpha = \inf\{x; \Pr(\sqrt{n}H(\bar{\mathbf{Y}})/\sigma \le x) > \alpha\}$, $u_\alpha = \Phi^{-1}(\alpha)$, ϵ is any constant in (0,1/2) and b_j's are polynomials depending on Q_j's.

Error Bounds

It is possible to get error bounds for approximation given by the CFE provided we have error bounds for EE. For simplicity, we give error bounds for the first-order CFE (see Chap. 5 in Fujikoshi et al. (2010)):

Theorem 2 *Suppose that for the distribution function of U we have*

$$F(x) \equiv Pr\{U \le x\} = G(x) + R_1(x),$$

where for remainder term $R_1(x)$ there exists a constant c_1 such that

$$|R_1(x)| \le d_1 \equiv c_1\epsilon.$$

Let x_α and u_α be the upper $100\alpha\%$ points of F and G, respectively; that is,

$$Pf\{U \le x_\alpha\} = G(u_\alpha) = 1 - \alpha.$$

Then, for any α such that $1 > \alpha > d_1$ and $1 > \alpha + d_1$:

1. $u_{\alpha+d_1} \le x_\alpha \le u_{\alpha-d_1}$.
2. $|x_\alpha - u_\alpha| \le d_1 / g(u_{(1)})$, *where*

$$g(u_1) = \min_{u \in [u_{\alpha+d_1}, u_{\alpha-d_1}]} g(u).$$

About the Author

DSc Vladimir V. Ulyanov is Professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics at Lomonosov Moscow State University. He has received the State Prize of the USSR for Young Scientists (1987), Alexander von Humboldt Research Fellowship, Germany (1991–1993), JSPS Research Fellowship, Japan (1999, 2004). He was visiting Professor/Researcher at Bielefeld University, Germany, University of Leiden, Université de Paris V, University of Hong Kong, Institute of Statistical Mathematics in Tokyo, National University of Singapore, the University of Melbourne etc. He is a member of the Bernoulli Society. Professor Ulyanov is the author of more than 50 journal articles and a book *Multivariate Statistics: High-Dimensional and Large-Sample Approximations* (with Y. Fujikoshi and R. Shimizu, John Wiley and Sons, 2010).

Cross References

►Edgeworth Expansion
►Multivariate Statistical Distributions

References and Further Reading

Bol'shev LN (1963) Asymptotically Pearson transformations. Theor Probab Appl 8:121–146
Cornish EA Fisher RA (1937) Moments and cumulants in the specification of distributions. Rev Inst Int Stat 4:307–320
Fisher RA, Cornish EA (1946) The percentile points of distributions having known cumulants. J Am Stat Assoc 80:915–922
Fujikoshi Y, Ulyanov VV, Shimizu R (2010) Multivariate statistics : high-dimensional and large-sample approximations. Wiley Series in Probability and Statistics. Wiley, Hoboken
Hall P (1992) The bootstrap and Edgeworth expansion. Springer, New York
Hill GW, Davis AW (1968) Generalized asymptotic expansions of Cornish–Fisher type. Ann Math Stat 39:1264–1273
Jaschke S (2002) The Cornish–Fisher-expansion in the context of delta-gamma-normal approximations. J Risk 4(4):33–52
Lee YS, Lin TK (1992) Higher-order Cornish–Fisher expansion. Appl Stat 41:233–240
Lee YS, Lin TK (1993) Correction to algorithm AS269: higher-order Cornish–Fisher expansion. Appl Stat 42:268–269
Takeuchi K, Takemura A (1988) Some results on univariate and multivariate Cornish–Fisher expansion: algebraic properties and validity, Sankhyā A 50:111–136
Wallace DL (1959) Bounds on normal approximations to Student's and the chisquare distributions. Ann Math Stat 30:1121–1130

Correlation Coefficient

Nitis Mukhopadhyay
Professor
University of Connecticut-Storrs, Storrs, CT, USA

Introduction

A covariance term loosely aims at capturing some essence of *joint dependence* between two random variables. A correlation coefficient is nothing more than an appropriately scaled version of the covariance.

Section "Population Correlation Coefficient" introduces the concepts of a covariance and the population correlation coefficient. Section "Correlation Coefficient and Independence" highlights some connections between the correlation coefficient, independence, and dependence.

Section "A Sample Correlation Coefficient" summarizes the notion of a sample correlation coefficient and its distribution, both exact and large-sample approximation, due to Fisher (1915; 1925). Section "Partial Correlations" gives a brief summary of the concept of partial correlation coefficients.

Population Correlation Coefficient

A covariance term tries to capture a sense of *joint dependence* between two real valued random variables. A correlation coefficient, however, is an appropriately scaled version of a covariance.

Definition 1 *The covariance between two random variables X_1 and X_2, denoted by $Cov(X_1, X_2)$, is defined as*

$$Cov(X_1, X_2) = E\left[(X_1 - \mu_1)(X_2 - \mu_2)\right]$$

$$\text{or equivalently } E\left[X_1 X_2\right] - \mu_1 \mu_2,$$

where $\mu_i = E(X_i)$, $i = 1, 2$ and $E\left[X_1 X_2\right]$, μ_1, μ_2 are assumed finite.

Definition 2 *The correlation coefficient between two random variables X_1 and X_2, denoted by ρ_{X_1,X_2}, is defined as*

$$\rho_{X_1,X_2} = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2},$$

whenever one has $0 < \sigma_1^2 = V(X_1) < \infty$ *and* $0 < \sigma_2^2 = V(X_2) < \infty$.

One may note that we do not explicitly assume $-\infty < Cov(X_1, X_2) < \infty$. In view of the assumption $0 < \sigma_1^2, \sigma_2^2 < \infty$, one can indeed claim the finiteness of $Cov(X_1, X_2)$ by appealing to Cauchy–Schwartz inequality. It should also be clear that

$$Cov(X_1, X_2) = Cov(X_2, X_1) \text{ and } Cov(X_1, X_1) = V(X_1),$$

as long as those terms are finite.

Two random variables X_1, X_2 are respectively called negatively correlated, uncorrelated, or positively correlated if and only if ρ_{X_1,X_2} *is neg*ative, zero or positive.

Theorem 1 *Consider random variables* X_1 *and* X_2 *and assume that* $0 < V(X_1), V(X_2) < \infty$. *Then, we have the following results:*

1. *Let* $Y_i = c_i + d_i X_i$ *w.p.*1 *where* $-\infty < c_i < \infty$ *and* $0 < d_i < \infty$ *are fixed numbers,* $i = 1, 2$. *Then,* $\rho_{Y_1,Y_2} = \rho_{X_1,X_2}$.
2. $|\rho_{X_1,X_2}| \le 1$, *the equality holds if and only if* $X_1 = a + bX_2$ *w.p.*1 *for some real numbers* a *and* b.

More details can be found from Mukhopadhyay (2000, Sect. 3.4).

Correlation Coefficient and Independence

If ρ_{X_1,X_2} is finite and X_1, X_2 are independent, then $\rho_{X_1,X_2} = 0$. Its converse is not necessarily true. In general, $\rho_{X_1,X_2} = 0$ may not imply independence between X_1, X_2. An example follows.

Example 1 Let X_1 be $N(0,1)$ and $X_2 = X_1^2$. Then, $Cov(X_1, X_2) = 0$, and surely $0 < V(X_1), V(X_2) < \infty$, so that $\rho_{X_1,X_2} = 0$. But, X_1 and X_2 are dependent variables. More details can be found from Mukhopadhyay (2000, Sect. 3.7). The recent article of Mukhopadhyay (2010) is relevant here.

Now, we state an important result which clarifies the role of zero correlation in a bivariate normal distribution.

Theorem 2 *Suppose that* (X_1, X_2) *has the* $N_2\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\right)$ *distribution where* $-\infty < \mu_1, \mu_2 < \infty$, $0 < \sigma_1, \sigma_2 < \infty$ *and* $-1 < \rho (= \rho_{X_1,X_2}) < 1$. *Then,* X_1 *and* X_2 *are independent if and only if* $\rho = 0$.

Example 2 A zero correlation coefficient implies *independence* not merely in the case of a bivariate normal distribution. Consider random variables X_1 and X_2 whose joint probability distribution puts mass only at four points $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. Now, if $Cov(X_1, X_2) = 0$, then X_1 and X_2 must be independent.

A Sample Correlation Coefficient

We focus on a bivariate normal distribution. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid $N_2\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho\right)$ where $-\infty < \mu_1, \mu_2 < \infty, 0 < \sigma_1^2, \sigma_2^2 < \infty$ and $-1 < \rho < 1, n \ge 2$. Let us denote

$$\overline{X} = n^{-1}\Sigma_{i=1}^n X_i \quad \overline{Y} = n^{-1}\Sigma_{i=1}^n Y_i$$

$$S_1^2 = (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X})^2 \quad S_2^2 = (n-1)^{-1}\Sigma_{i=1}^n (Y_i - \overline{Y})^2$$

$$S_{12} = (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) \quad r = S_{12}/(S_1 S_2).$$

Here, r is called the *Pearson* (or *sample*) *correlation coefficient*. This r is customarily used to estimate ρ.

The probability distribution of r is complicated, particularly when $\rho \ne 0$. But, even without explicitly writing the pdf of r, it is simple enough to see that the distribution of r can not involve μ_1, μ_2, σ_1^2 and σ_2^2.

Francis Galton introduced a numerical measure, r, which he termed "reversion" in a lecture at the Royal Statistical Society on February 9, 1877 and later called "regression." The term "cor-relation" or "correlation" probably appeared first in Galton's paper to the Royal Statistical Society on December 5, 1888. At that time, "correlation" was defined in terms of deviations from the median instead of the mean. Karl Pearson gave the definition and calculation of correlation r in 1897. In 1898, Pearson and his collaborators discovered that the standard deviation of r happened to be $(1-\rho^2)/\sqrt{n}$ when n was large. "Student" derived the "probable error of a correlation coefficient" in 1908. Soper (1913) gave large-sample approximations for the mean and variance of r which performed better than those proposed earlier by Pearson. Refer to DasGupta (1980) for more historical details.

The unsolved problem of finding the exact pdf of r for normal variates came to R. A. Fisher's attention via Soper's 1913 paper. The pdf of r was published in the year 1915 by Fisher for all values of $\rho \in (-1,1)$. Fisher, at the age of 25, brilliantly exploited the n-dimensional geometry to come up with the solution, reputedly within one week. Fisher's genius immediately came into limelight. Following the publication of Fisher's results, however, Karl Pearson set up a major cooperative study of the correlation. One will notice that in the team formed for this cooperative project (Soper et al. 1917) studying the distribution of the sample correlation coefficient, the young Fisher was not included. This happened in spite of the fact that Fisher was right there and he already earned quite some fame. Fisher felt hurt as he was left out of this project. One thing led to another. RA. Fisher and Karl Pearson continued criticizing each other even more as each held on to his own philosophical stand.

We will merely state the pdf of r when $\rho = 0$. This pdf is given by

$$f(r) = c\,(1-r^2)^{\frac{1}{2}(n-4)} \text{ for } -1 < r < 1,$$

where $c = \Gamma\left(\frac{1}{2}(n-1)\right)\left\{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}(n-2)\right)\right\}^{-1}$ for $n \geq 3$. Using a simple transformation technique, one can easily derive the following result:

$r(n-2)^{1/2}(1-r^2)^{-1/2}$ has the Student's t distribution with $(n-2)$ degrees of freedom when $\rho = 0$.

Fisher's geometric approach (1915) also included the exact pdf of r in the form of an infinite power series for all values of $\rho \neq 0$. One may also look at Rao (1973, pp. 206–209) for a non-geometric approach.

Large-Sample Distribution

But, now suppose that one wishes to construct an *approximate* $100(1-\alpha)\%$ confidence interval for ρ, $0 < \alpha < 1$. In this case, one needs to work with the *non-null* distribution of r. We mentioned earlier that the *exact* distribution of r, when $\rho \neq 0$, was found with an ingenious geometric technique by Fisher (1915). That exact distribution being very complicated, Fisher (1915) proceeded to derive the following asymptotic distribution when $\rho \neq 0$:

$$\sqrt{n}(r-\rho) \xrightarrow{\mathcal{L}} N\left(0, (1-\rho^2)^2\right) \text{ as } n \to \infty.$$

For a proof, one may look at Sen and Singer (1993, pp. 134–136) among other sources.

One should realize that a variance stabilizing transformation may be useful here. We may invoke Mann-Wald Theorem (see Mukhopadhyay 2000, pp. 261–262) by requiring a suitable function $g(.)$ such that the asymptotic variance of $\sqrt{n}\,[g(r) - g(\rho)]$ becomes free from ρ. That is, we want to have:

$$g'(\rho)(1-\rho^2) = k, \text{ a constant.}$$

So, $g(\rho) = k\int \frac{1}{(1-\rho^2)}d\rho$. Hence, we rewrite

$$g(\rho) = \tfrac{1}{2}k\int\left\{\tfrac{1}{1-\rho} + \tfrac{1}{1+\rho}\right\}d\rho = \tfrac{1}{2}k\log\left\{\tfrac{1+\rho}{1-\rho}\right\} + \text{constant.}$$

It is clear that we should look at the transformations:

$$U = \frac{1}{2}\log\left\{\frac{1+r}{1-r}\right\} \text{ and } \xi = \frac{1}{2}\log\left\{\frac{1+\rho}{1-\rho}\right\},$$

and consider the asymptotic distribution of $\sqrt{n}[U-\xi]$. Now, we can claim that

$$\sqrt{n}[U-\xi] \xrightarrow{\mathcal{L}} N(0,1) \text{ as } n \to \infty,$$

since with $g(\rho) = \frac{1}{2}\log\left\{\frac{1+\rho}{1-\rho}\right\}$, one has $g'(\rho) = \frac{1}{1-\rho^2}$. That is, for large n, we should consider the following pivot:

$\sqrt{n}[U-\xi]$, which is approximately $N(0,1)$ for large n.

These transformations can be *equivalently* stated as

$$U = \tanh^{-1}(r) \text{ and } \xi = \tanh^{-1}(\rho),$$

which are referred to as Fisher's Z transformations introduced in 1925.

Fisher obtained the first four moments of $\tanh^{-1}(r)$ which were later updated by Gayen (1951). It turns out that the variance of $\tanh^{-1}(r)$ is approximated better by $\frac{1}{n-3}$ rather than $\frac{1}{n}$ when n is moderately large. Hence, in many applications, one uses an alternate pivot (for $n > 3$):

$$\sqrt{n-3}\left[\tanh^{-1}(r) - \tanh^{-1}(\rho)\right], \text{ which is approximately } N(0,1),$$

for large n whatever be ρ, $-1 < \rho < 1$.

For large n, one customarily uses Fisher's Z transformations to come up with an *approximate* $100(1-\alpha)\%$ confidence interval for ρ. Also, to test a null hypothesis $H_0 : \rho = \rho_0$, for large n, one uses the test statistic

$$Z_{calc} = \sqrt{n-3}\left[\tanh^{-1}(r) - \tanh^{-1}(\rho_0)\right]$$

and comes up with an *approximate* level α test against an appropriate alternative hypothesis. These are customarily used in all areas of statistical science whether the parent population is bivariate normal or not.

Partial Correlations

Suppose that in general $\mathbf{X} = (X_1, \ldots, X_p)$ has a p-dimensional probability distribution with all pairwise correlations finite. Now, ρ_{X_i,X_j} will simply denote the correlation coefficient between X_i, X_j based on their joint bivariate distribution derived from the distribution of $\mathbf{X}$, for any $i \neq j = 1, \ldots, p$.

Next, $\rho_{X_i,X_j.X_k}$ is simply the correlation coefficient between X_i, X_j based on their joint bivariate conditional distribution given X_k that is derived from the distribution of $\mathbf{X}$, for any $i \neq j \neq k = 1, \ldots, p$.

Similarly, $\rho_{X_i,X_j.X_k,X_l}$ is simply the correlation coefficient between the pair of random variables X_i, X_j based on their joint bivariate conditional distribution given X_k, X_l derived from the distribution of $\mathbf{X}$, for any $i \neq j \neq k \neq l = 1, \ldots, p$. Clearly, one may continue further like this.

Such correlation coefficients $\rho_{X_i,X_j.X_k}$, $\rho_{X_i,X_j.X_k,X_l}$ are referred to as partial correlation coefficients. Partial correlation coefficients have important implications in multiple linear regression analysis. One may refer to Ravishanker and Dey (2002, pp. 160–164) among other sources.

About the Author

For biography *see* the entry ►Sequential Sampling.

Cross References

►Autocorrelation in Regression
►Intraclass Correlation Coefficient
►Kendall's Tau
►Measures of Dependence
►Rank Transformations
►Spurious Correlation
►Tests of Independence
►Weighted Correlation

References and Further Reading

DasGupta S (1980) Distributions of the correlation coefficient. In: Fienberg SE, Hinkley DV (eds) R. A. Fisher: an appreciation. Springer, New York, pp 9–16
Fisher RA (1915) Frequency distribution of the values of the correlation coefficients in samples from an indefinitely large population. Biometrika 10:507–521
Fisher RA (1925) Theory of statistical estimation. Proc Camb Phil Soc 22:700–725
Gayen AK (1951) The frequency distribution of the product moment correlation coefficient in random samples of any size drawn from non-normal universes. Biometrika 38:219–247
Mukhopadhyay N (2000) Probability and statistical inference. Marcel Dekker, New York
Mukhopadhyay N (2010) When finiteness matters: Counterexamples to notions of covariance, correlation, and independence. The Amer. Statistician, in press
Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, New York
Ravishanker N, Dey DK (2002) A first course in linear model theory. Chapman & Hall/CRC Press, Boca Raton
Sen PK, Singer JO (1993) Large sample methods in statistics. Chapman & Hall, New York
Soper HE (1913) On the probable error of the correlation coefficient to a second approximation. Biometrika 9:91–115
Soper HE, Young AW, Cave BM, Lee A, Pearson K (1917) On the distribution of the correlation coefficient in small samples. A cooperative study. Biometrika 11:328–413
"Student" (Gosset WS) (1908) The probable error of a mean. Biometrika 6:1–25

Correspondence Analysis

Jörg Blasius
Professor
University of Bonn, Bonn, Germany

Correspondence analysis (CA) has been developed in the 1960s in France by Jean-Paul Benzécri and his collaborators; it is the central part of the French "Analyse des Données," or in English, geometric data analysis (cf. Benzécri et al. 1973; Greenacre 1984, 2007; Lebart et al. 1984; Le Roux and Rouanet 2004). The method can be applied to any data table with nonnegative entries. The main objective of CA is to display rows and columns of data tables in two-dimensional spaces, called "maps." This kind of data description via visualization reflects a way of thinking that is typical for the social sciences in France, especially associated with the name of Pierre Bourdieu, and of many statisticians in the 1970s and 1980s in France, who at that time published almost only in French. The philosophy behind their work can be expressed by the famous quotation of Jean-Paul Benzécri who pointed out that "The model must follow the data, and not the other way around." Instead of limiting the data to restrictive and subjectively formulated statistical models, they show the importance of the data and of the features in the data themselves. The discussion outside of France started with the textbooks by Greenacre (1984) and Lebart et al. (1984).

CA translates deviations from the independence model in a contingency table into distances as the following brief introduction shows. In the simple case, there is a two-way table $\mathbf{N}$ with I rows and J columns. In cases where the data are from survey research, the cells n_{ij} of $\mathbf{N}$ contain the frequencies of a bivariate cross-tabulation of two variables, with $\sum_{ij} n_{ij} = n$. Dividing n_{ij} by the sample size n provides the percentages of the total p_{ij}, or, for the entire table, with the $(I \times J)$ correspondence matrix $\mathbf{P}$. Thereby, $\mathbf{r} = \mathbf{P1}$ is the vector of the "row masses," or the "average column profile" with elements $r_i = n_{i+}/n$, and $\mathbf{c} = \mathbf{P}^T\mathbf{1}$ is the vector of "column masses" or the "average row profile" with elements $c_j = n_{+j}/n$; $\mathbf{D}_r$ and $\mathbf{D}_c$ are the diagonal matrices of the row and column masses, respectively.

The matrix of row profiles can be defined as the rows of the correspondence matrix $\mathbf{P}$ divided by their respective row masses, $D_r^{-1}P$; for the matrix of columns profiles yields PD_c^{-1}. As a measure of similarity between two row profiles (or between two column profiles, respectively), a weighted Euclidian or chi-square distance in the metric D_r^{-1} (or, D_c^{-1}, respectively) is used. For chi-square calculations, the weighted deviations from independence over all cells of the contingency table are used. For each cell, the unweighted deviation of the observed from the expected value can be calculated by $(n_{ij} - \hat{n}_{ij})$, with $\hat{n}_{ij} = (n_{i+} \times n_{+j})/n$. Dividing $(n_{ij} - \hat{n}_{ij})$ by n provides with $(p_{ij} - r_i c_j)$, or, in matrix notation, $(\mathbf{P} - \mathbf{rc}^T)$, with the unweighted deviations from the independence model for the entire table.

To fulfill the chi-square statistic, this matrix is weighted by the product of the square root of the row and column masses to give the standardized residuals $s_{ij} = (p_{ij} - r_i c_j)/\sqrt{r_i c_j}$, or in matrix notation, the $(I \times J)$ matrix of

standardized residuals $S = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$. The similarity to chi-square analysis and total inertia as a measure for the variation in the data table, which is defined as $\sum_{ij} s_{ij}^2 = \frac{\chi^2}{n} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$, becomes apparent. Applying singular value decomposition to $\mathbf{S}$ results in $SVD(\mathbf{S}) = U\mathbf{\Gamma}V^T$, where $\mathbf{\Gamma}$ is a diagonal matrix with singular values in descending order $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_S > 0$, with S = rank of $\mathbf{S}$. The columns from $\mathbf{U}$ are the left singular vectors, the columns from $\mathbf{V}$ are the right singular vectors, with $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.

The connection between SVD as used in CA and the well-known canonical decomposition is shown by $S^TS = V\mathbf{\Gamma}\mathbf{U}^TU\mathbf{\Gamma}\mathbf{V}^T = V\mathbf{\Gamma}^2\mathbf{V}^T = V\mathbf{\Lambda}V^T$, with $SS^T = U\mathbf{\Gamma}\mathbf{V}^TV\mathbf{\Gamma}\mathbf{U}^T = U\mathbf{\Gamma}^2\mathbf{U}^T = U\mathbf{\Lambda}U^T$, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_S > 0$; $\chi^2/n = \sum_s \lambda_s$ = total inertia, since trace $(\mathbf{SS}^T)$ = trace $(\mathbf{S}^T\mathbf{S})$ = trace $(\mathbf{\Gamma}^2)$ = trace $(\mathbf{\Lambda})$.

As in ▶principal component analysis (PCA), the first axis is chosen to explain the maximum variation in the data; the second axis captures the maximum of the remaining variation, and so on. Again, analogous to PCA, it is possible to interpret the variable categories in relation to the axes, which can be considered the latent variables. And furthermore, as in PCA and other data reduction methods, only the s major components are used for interpretation. The number of interpretable dimensions depends on criteria such as the eigenvalue criteria, theory (how many latent variables can be substantively interpreted), or the scree test (for more details, see Blasius 1994).

For the graphical representation, we use $F = D_r^{-1/2}U\mathbf{\Gamma}$ providing the principal coordinates of the rows, and $G = D_c^{-1/2}V\mathbf{\Gamma}$ providing the principal coordinates of the columns (for further details see Greenacre 1984, 2007). The maps drawn on the basis of principal coordinates are called "symmetric maps." In the full space, the distances between the rows and the distances between the columns can be interpreted as Euclidian distances, whereas the distances between the rows and the columns are not defined.

As in PCA, the input data can be factorized. Understanding correspondence analysis as a model (see, e.g., van der Heijden et al. 1989, 1994), the row and column coordinates can be used for recomputing the input data. Adding the latent variables successively models the deviations from independency. This is similar to the loglinear model and other modeling approaches such as the latent class model or the log-multiplicative model (see, e.g., van der Heijden et al. 1989, 1994; Goodman 1991). In loglinear analysis, for example, these deviations are modeled by using higher-order interaction effects; in correspondence analysis latent variables are used. For any cell yields $n_{ij} = nr_ic_j\left(1 + \sum_{s=1}^{S} f_{is}g_{js}/\gamma_s\right)$, or $p_{ij} = r_ic_j\left(1 + \sum_{s=1}^{S} f_{is}g_{js}/\gamma_s\right)$, and in matrix notation $P = rc^T + D_rF\mathbf{\Gamma}^{-1}G^TD_c$. The left part of the equation reflects the independence model and the right part, the modeling from independency by including the S factors in successive order. Including all factors in the model fully reconstructs the original data table $\mathbf{N}$.

The interpretation of CA is similar to the one of PCA, both methods provide eigenvalues and their explained variances, factor loadings, and factor values. While PCA is restricted to metric data, CA can be applied to any kind of data table with nonnegative entries, among others, to indicator and Burt matrices – in these two cases the method is called multiple correspondence analysis (MCA).

Whereas simple correspondence analysis is applied to a single contingency table or to a stacked table, MCA uses the same algorithm to an indicator or a Burt matrix. In the case of survey research, input data to simple CA is usually a matrix of raw frequencies of one or more contingency tables. In this context, there is usually one variable to be described, for example, preference for a political party, and one or more describing variables, for example, educational level and other sociodemographic indicators such as age groups, gender, and income groups. The number of variables can be quite high, apart from theoretical considerations there is no real limitation by the method. In the given case, each of the describing variables is cross-tabulated with the variable to be described in order to investigate the importance of this association. Concatenating, or stacking the tables before applying CA allows to visualize and interpret several relationships of "preferred political party" with the sociodemographic indicators in the same map.

Applying CA to the indicator matrix $\mathbf{Z}$ (=MCA), the table of input data has as many rows as there are respondents, and as many columns as there are response alternatives in all variables included in the analysis. A "1" in a given row indicates the respondent who chose that specific response category; otherwise there is a "0" for "specific response category not chosen." Considering all categories of all variables provides row sums that are constant and equal to the number of variables, the column sums reflect the marginals. An alternative to the indicator matrix as input to MCA is the Burt matrix $\mathbf{B}$. This matrix can either be generated by cross-tabulating all variables by all variables, including the cross-tabulations of the variables by themselves, and stacking them row- and column-wise. Further, $\mathbf{B}$ can be computed by multiplying the transposed indicator matrix by itself, that is $\mathbf{B} = \mathbf{Z}^T\mathbf{Z}$. The solutions from $\mathbf{Z}$ can be directly converted to those of $\mathbf{B}$ by rescaling the solution; for example, the squared eigenvalues of $\mathbf{Z}$

are equal to those of **B**. As it is true for PCA, MCA contains all first-order interaction effects, the method can be understood as a generalization of PCA to categorical data.

Taking a two-way contingency table with $I = 7$ rows, $J = 5$ columns, and $n = 500$ cases as an example, input data of the simple CA would be the frequencies of the (7×5) cross-table. Turning to MCA, input data is an indicator matrix with 500 rows (the number of cases) and 12 columns (the number of variable categories). MCA is also known under the names "homogeneity analysis" (see Gifi 1990; Heiser and Meulman 1994), "dual scaling" (Nishisato 1980, 2007), and "quantification of qualitative data III" (Hayashi 1954); CA procedures are available in all major statistic packages as well as in *R* (Greenacre and Nenadić 2006). For details regarding the history of CA and related methods, we refer to Nishisato (2007, Chap. 3).

CA employs the concept of inertia: the farther the categories are from the centroid along a given axis (squared distances) and the higher their masses (their marginals), the more the categories determine the geometric orientation of that axis. In the graphical solution, the locations of all variable categories can be compared to each other (except in simple CA and using symmetric maps, in this case the distances between rows and columns are not defined), short distances imply high similarities and long distances imply high dissimilarities. For all dimensions, CA supplies principal inertias that can be interpreted as canonical correlation coefficients (they are the singular values of the solution, i.e., the square roots of the eigenvalues), correlation coefficients between the item categories and the latent variables as well as scores for all item categories and all respondents.

There are several extensions of simple CA and MCA. With respect to the Burt matrix **B**, it is apparent that most of the variation in this super matrix is caused by the main diagonal blocks. These sub-matrices contain the cross-tabulations of the variables by themselves; the main diagonal elements of them contain the marginals of the variables while their off-diagonal elements are equal to zero. Excluding this variation in an iterative procedure and visualizing the variation of the off-diagonal blocks of **B** only is the objective of joint correspondence analysis. The aim of subset correspondence analysis is to concentrate on some response categories only, while excluding others from the solution. For example, applying subset MCA to a set of variables, the structure of non-substantive responses ("don't know," "no answer") can be analyzed separately, or these responses can be excluded from the solution while concentrating on the substantive responses. Variables can also be included in the model as supplementary or passive ones; in this case they do not have any impact on the geometric orientation of the axes but they can be interpreted together with the active variables. CA can not only be applied to single and stacked contingency tables or to indicator matrix, it can also be used to analyze rank and metric data, multiple responses, or squared tables. The statistical background and examples of these kinds of data can be found in the textbook of Greenacre (2007) as well as in the readers of Greenacre and Blasius (1994, 2006), and Blasius and Greenacre (1998).

About the Author

Jörg Blasius is the President of RC33 (Research Committee of Logic and Methodology in Sociology) at ISA (International Sociological Association) (2006–2010). Together with Michael Greenacre (Barcelona) he founded CARME (Correspondence Analysis and Related Methods Network) and edited three books on Correspondence Analysis, in June 2009 they had a special issue on this topic in *Computational Statistics and Data Analysis* (further coeditors: Patrick Groenen and Michel van de Velden).

Cross References

►Data Analysis
►Distance Measures
►Multivariate Data Analysis: An Overview

References and Further Reading

Benzécri JP et al (1973) L'analyse des Données. L'analyse des Correspondances. Dunod, Paris

Blasius J (1994) Correspondence analysis in social science research. In: Greenacre M, Blasius J (eds) Correspondence analysis in the social sciences. Recent developments and applications. Academic, London, pp 23–52

Blasius J, Greenacre M (eds) (1998) Visualization of categorical data. Academic, London

Gifi A (1990) Nonlinear multivariate analysis. Wiley, Chichester

Goodman LA (1991) Measures, models, and graphical display in the analysis of cross-classified data (with discussion). J Am Stat Assoc 86:1085–1138

Greenacre MJ (1984) Theory and applications of correspondence analysis. Academic, London

Greenacre MJ (2007) Correspondence analysis in practice. Chapman & Hall, Boca Raton

Greenacre MJ, Blasius J (eds) (1994) Correspondence analysis in the social sciences. Recent developments and applications. Academic, London

Greenacre MJ, Blasius J (eds) (2006) Multiple correspondence analysis and related methods. Chapman & Hall, Boca Raton

Greenacre MJ, Oleg Nenadić (2006) Computation of multiple correspondence analysis, with code in R. In: Greenacre M, Blasius J (eds) Multiple correspondence analysis and related methods. Chapman & Hall, Boca Raton, pp 523–551

Hayashi C (1954) Multidimensional quantification – with the applications to the analysis of social phenomena. Ann Inst Stat Math 5:231–245

Heiser WJ, Meulman JJ (1994) Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships. In: Greenacre M, Blasius J (eds) Correspondence analysis in the social sciences. Recent developments and applications. Academic, London, pp 179–209
Lebart L, Morineau A, Warwick KM (1984) Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Wiley, New York
Le Roux B, Rouanet H (2004) Geometric data analysis. North Holland, Amsterdam
Nishisato S (1980) Analysis of categorical data: dual scaling and its applications. University of Toronto Press, Toronto
Nishisato S (2007) Multidimensional nonlinear descriptive analysis. Chapman & Hall, Boca Raton
Van der Heijden PGM, de Falguerolles A, de Leeuw J (1989) A combined approach to contingency table analysis using correspondence analysis and loglinear analysis. Appl Stat 38:249–292
Van der Heijden PGM, Mooijaart A, Takane Y (1994) Correspondence analysis and contingency table models. In: Greenacre M, Blasius J (eds) Correspondence analysis in the social sciences. Recent developments and applications. Academic, London, pp 79–111

C_p Statistic

Colin Mallows
Basking Ridge, Avayalabs, NJ, USA

The C_p statistic was invented by C. Mallows in 1963. It facilitates the comparison of many subset-regression models, by giving for each model an unbiased estimate of the (scaled) total mean-square-error for that model. There is an associated graphical technique called the "C_p plot" in which values of C_p (one for each subset of regressors) are plotted against p.

The problem in choosing a subset-regression model for predicting a response is that including too many unnecessary terms will add to the variance of the predictions, while including too few will result in biased predictions.

In more detail, if we have n observations, and k regressors are available (possibly including a constant term), let P denote some subset of these. (Usually if a constant term is to be considered, this will appear in each subset). Let p be the number of regressors in the subset P. Then C_p (for the P-subset model) is defined to be

$$C_p = \frac{RSS_P}{s^2} - n + 2p$$

where RSS_P is the residual sum of squares for this P-model, and s^2 is an estimate of the residual variance when all relevant terms are included in the model. Usually this is taken to be RSS_K where K is the set of all available regressors.

Under the usual assumptions, that the vector of observations y equals $v + z$ where v is the vector of true means, and the z's are independent with mean zero and constant variance σ^2, $s^2 C_p$ is an unbiased estimate of $\sigma^2 E(J_P)$ where J_P is $|\hat{v}_P - v|^2$, and where $\hat{v}_P$ is the estimate of v that is obtained by fitting the P model. Thus J_P is a measure of the adequacy for prediction of the P model. This result holds even when the true model v is not expressible in terms of the available regressors. However in this case we cannot use the residual sum of squares from the full (K) model as an estimate of σ^2.

The C_p statistic is often used to guide selection of a subset-model, but this cannot be recommended; while for each P separately, C_p gives an unbiased estimate of the scaled mean-square error for that subset, this is not true if the subset is chosen to minimise C_p. In fact this approach can lead to worse results than are obtained by simply fitting all available regressors. In a 1995 paper, Mallows has attempted to quantify this effect.

The C_p statistic is similar to ►Akaike's Information criterion.

About the Author

Colin L. Mallows spent 40 years at AT&T Bell Labs and one of its descendants, AT&T Labs. Since retiring he has been a consultant at another descendant, Avaya Labs. He is a Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the Royal Statistical Society. He has served on several committees of the IMS and ASA. He was an Associate editor of JASA (1966–1972). He has been COPSS Fisher Lecturer and ASA Deming Lecturer, and has received the ASA Wilks Medal. He has written over 150 papers, and has edited two books.

Cross References

►Akaike's Information Criterion
►General Linear Models

References and Further Reading

Daniel C, Wood FS (1980) Fitting equations to data, rev edn. Wiley, New York
Gorman JW, Toman RJ (1966) Selection of variables for fitting equations to data. Technometrics 8:27–51
Mallows CL (1973) Some comments on Cp. Technometrics 15:661–675
Mallows CL (1995) More comments on Cp. Technometrics 37:362–372

Cramér–Rao Inequality

Maarten Jansen[1], Gerda Claeskens[2]
[1]Université libre de Bruxelles, Brussels, Belgium
[2]K. U. Leuven, Leuven, Belgium

The Cramér–Rao Lower Bound

The Cramér–Rao inequality gives a lower bound for the variance of an unbiased estimator of a parameter. It is named after work by Cramér (1946) and Rao (1945). The inequality and the corresponding lower bound in the inequality are stated for various situations. We will start with the case of a scalar parameter and independent and identically distributed random variables $X_1, \ldots, X_n$, with the same distribution as X.

Denote $\boldsymbol{X} = (X_1, \ldots, X_n)$ and denote the common probability mass function or probability density function of X at a value x by $f(x;\theta)$ where $\theta \in \Theta$, which is a subset of the real line $\mathbb{R}$ and $x \in \mathbb{R}$. Denote the support of X by R, that is, $R = \{x : f(x;\theta) > 0\}$.

Assumptions

1. The partial derivative $\frac{\partial}{\partial\theta} \log f(x;\theta)$ exists for all $\theta \in \Theta$ and all $x \in R$ and it is finite. This is equivalent to stating that the Fisher information value $I_X(\theta) = E\left[\left(\frac{\partial}{\partial\theta} \log f(X;\theta)\right)^2\right]$ is well defined, for all $\theta \in \Theta$.
2. The order of integration and differentiation is interchangeable in $\int \frac{\partial}{\partial\theta} \log f(x;\theta) dx$. If the support of X, that is, the set R, is finite, then the interchangeability is equivalent with the condition that the support does not depend on θ. A counter-example on uniformly distributed random variables is elaborated below.

The Cramér–Rao inequality

Under assumptions (i) and (ii), if $\hat{\theta} = g(\boldsymbol{X})$ is an unbiased estimator of θ, this means that $E[\hat{\theta}] = \theta$, then

$$\mathrm{var}(\hat{\theta}) \geq 1/\left[n \cdot I_X(\theta)\right].$$

The lower bound in this inequality is called the Cramér–Rao lower bound.

The proof starts by realizing that the correlation of the score $V = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \log f_X(X_i;\theta)$ and the unbiased estimator $\hat{\theta}$ is bounded above by 1. This implies that $\left(\mathrm{var}(V) \cdot \mathrm{var}(\hat{\theta})\right)^{1/2} \geq \mathrm{cov}(V, \hat{\theta})$. The assumptions are needed to prove that the expected score $E(V)$ is zero. This implies that the covariance $\mathrm{cov}(V, \hat{\theta}) = 1$, from which the stated inequality readily follows.

A second version of the Cramér–Rao inequality holds if we estimate a functional $\kappa = H(\theta)$. Under assumptions (i) and (ii), if $\boldsymbol{X}$ is a sample vector of independent observations from random variable X with density function $f(x;\theta)$ and $\hat{\kappa} = h(\boldsymbol{X})$ is an unbiased estimator of $H(\theta)$, such that the first derivative $\frac{dH(\theta)}{d\theta}$ exists and is finite for all θ, then

$$\mathrm{var}(\hat{\kappa}) \geq \left[\frac{dH(\theta)}{d\theta}\right]^2 / \left[n \cdot I_X(\theta)\right].$$

Similar versions of the inequality can be phrased for observations that are independent but not identically distributed.

In the case of a vector parameter $\boldsymbol{\theta}$, the variance of the single parameter estimator $\mathrm{var}(\hat{\theta})$ is replaced by the covariance matrix of the estimator vector $\Sigma_{\hat{\boldsymbol{\theta}}}$. This matrix is bounded by a matrix expression containing the inverse of the Fisher information matrix, where bounded means that the difference between the covariance matrix and its "upper bound" is a negative semidefinite matrix.

The Cramér–Rao inequality is important because it states what the best attainable variance is for unbiased estimators. Estimators that actually attain this lower bound are called efficient. It can be shown that maximum likelihood estimators asymptotically reach this lower bound, hence are asymptotically efficient.

Cramér–Rao and UMVUE

If $\boldsymbol{X}$ is a sample vector of independent observations from the random variable X with density function $f_X(x;\theta)$ and $\hat{\theta} = g(\boldsymbol{X})$ is an unbiased estimator of θ, then $\mathrm{var}(\hat{\theta}) = 1/\left[n \cdot I_X(\theta)\right] \Leftrightarrow \hat{\theta} = aV + b$ with probability one, where V is the score and a and b are some constants. This follows from the proof of the Cramér–Rao inequality: the lower bounded is reached if the correlation between the score and the estimator is one. This implies that $\mathrm{var}\left(\frac{V}{\sigma_V} + \frac{\hat{\theta}}{\sigma_{\hat{\theta}}}\right) = 0 \Rightarrow \frac{V}{\sigma_V} + \frac{\hat{\theta}}{\sigma_{\hat{\theta}}} = c$ almost surely for some constant c. We here used the notation σ_X to denote the standard deviation of a random variable X.

The coefficients a and b may depend on θ, but $\hat{\theta}$ should be observable without knowing θ.

If a and b exist such that $\hat{\theta}$ is unbiased and observable, then $\hat{\theta}$ has the smallest possible variance among all unbiased estimators: it is then certainly the uniformly minimum variance unbiased estimator (UMVUE).

It may, however, be well possible that no a and b can be found. In that case, the UMVUE, if it exists, does not reach the Cramér-Rao lower bound. In that case, the notion of *sufficiency* can be used to find such UMVUE.

Counter example: estimators for the upperbound of uniform data

Let $X \sim \text{unif}[0,a]$, so $f_X(x) = \frac{1}{a} I(0 \le x \le a)$, where $I(c \le x \le d)$ is the *indicator* function of the interval $[c,d]$. We want to estimate a. The maximum likelihood estimator (MLE) is $\hat{a}_{\text{MLE}} = \max_{i=1,\ldots,n} X_i$, which is biased. Define $\hat{a}_u = \frac{n}{n-1}\hat{a}_{\text{MLE}}$, which is unbiased. The method of moments leads to an estimator $\hat{a}_{\text{MME}} = 2\overline{X}$, which is also unbiased. The score is $V_i = \frac{\partial}{\partial a} \log f_X(X_i;a) = -\frac{1}{a}$. This is a constant (so, not a random variable), whose expected value is of course *not zero*. This is because the partial derivative and expectation cannot be interchanged, as the boundary of the support of X depends on a. As a consequence, the Cramér–Rao lower bound is *not* valid here. We can verify that $\text{var}(\hat{a}_{\text{MLE}}) = \frac{n}{(n+2)(n+1)^2} a^2$ and $\text{var}(\hat{a}_u) = \frac{1}{n(n+2)} a^2$. This is (for $n \to \infty$) one order of magnitude smaller than $\text{var}(\hat{a}_{\text{MME}}) = \frac{1}{3n} a^2$ and also one order of magnitude smaller than what you would expect for an unbiased estimator if the Cramér–Rao inequality would hold.

A Bayesian Cramér–Rao Bound

It should be noted that biased estimators can have variances below the Cramér–Rao lower bound. Even the MSE (mean squared error), which equals the sum of the variance and the squared bias can be lower than the Cramér–Rao lower bound (and hence lower than any unbiased estimator could attain). A notable example in this respect is Stein's phenomenon on shrinkage rules (Efron and Morris 1977).

In practice, large classes of estimators, for example most nonparametric estimators, are biased. An inequality that is valid for biased or unbiased estimators is due to van Trees (1968, p. 72), see also Gill and Levit (1995) who developed multivariate versions of the inequality.

We assume that the parameter space Θ is a closed interval on the real line and denote by g some probability distribution on Θ with density $\lambda(\theta)$ with respect to the Lebesgue measure. This is where the Bayesian flavor enters. The θ is now treated as a random variable with density λ. We assume that λ and $f(x;\cdot)$ are absolutely continuous and that λ converges to zero at the endpoints of the interval Θ. Moreover we assume that $E[\frac{\partial}{\partial\theta} \log f(X;\theta)] = 0$. We denote $I(\lambda) = E[\{\log \lambda(\theta)\}^2]$ and have that $E[I_X(\theta)] = \int I_X(\theta) g(\theta) d\theta$. Then, for an estimator $\hat{\theta} = \hat{\theta}(X)$, it holds that

$$E[\{\hat{\theta} - \theta\}^2] \ge \frac{1}{E[I_X(\theta)] + I(\lambda)}.$$

A second form of this inequality is obtained for functionals $\kappa = H(\theta)$. Under the above assumptions, for an estimator $\hat{\kappa} = h(\boldsymbol{X})$ of $H(\theta)$, such that the first derivative $\frac{dH(\theta)}{d\theta}$ exists and is finite for all θ,

$$E[\{\hat{\kappa} - H(\theta)\}^2] \ge \frac{\{E[\frac{d}{d\theta} H(\theta)]\}^2}{E[I_X(\theta)] + I(\lambda)}.$$

About the Authors

For the biography of Maarten Jansen *see* the entry ▶Nonparametric Estimation.

For the biography of Gerda Claeskens *see* the entry ▶Model Selection.

Cross References

▶Estimation
▶Minimum Variance Unbiased
▶Sufficient Statistical Information
▶Unbiased Estimators and Their Applications
▶Uniform Distribution in Statistics

References and Further Reading

Cramér H (1946) Mathematical methods of statistics. Princeton University Press, Princeton
Efron B, Morris C (1977) Stein's paradox in statistics. Scient Am 236:119–127
Gill RD, Levit BY (1995) Applications of the van Trees inequality: a Bayesian Cramér–Rao bound. Bernoulli 1(1–2): 59–79
Rao C (1945) Information and the accuracy attainable in the estimation of statistical parameters. Bull Calcutta Math Soc 37:81–89
van Trees HL (1968) Detection, estimation and modulation theory: part I. Wiley, New York

Cramér–Von Mises Statistics for Discrete Distributions

Michael A. Stephens
Professor Emeritus
Simon Fraser University, Burnaby, B.C., Canada

Introduction

Cramér–von Mises statistics are well established for testing fit to continuous distributions; see Anderson (2010) and Stephens (2010), both articles in this encyclopedia. In this paper, the corresponding statistics for testing discrete distributions will be described.

Consider a discrete distribution with k cells labeled $1, 2, \ldots, k$, and with probability p_i of falling into cell i. Suppose n independent observations are given; let o_i be the observed number of observations and $e_i = np_i$ be the expected number in cell i. Let $S_j = \sum_{i=1}^{j} o_i$ and $T_j = \sum_{i=1}^{j} e_i$.

Then S_j/n and $H_j = T_j/n$ are the cumulated histograms of observed and expected values and correspond to the empirical distribution function $F_n(z)$ and the cumulative distribution function $F(.)$ for continuous distributions. Suppose $Z_j = S_j - T_j, j = 1, 2, \ldots, k$; the weighted mean of the Z_i is $\bar{Z} = \sum_{j=1}^{k} Z_j t_j$, where $t_j = (p_j + p_{j+1})/2$, with $p_{k+1} = p_1$. The modified Cramér–von Mises statistics are then defined as follows:

$$W_d^2 = n^{-1} \sum_{j=1}^{k} Z_j^2 t_j; \tag{1}$$

$$U_d^2 = n^{-1} \sum_{j=1}^{k} (Z_j - \bar{Z})^2 t_j; \tag{2}$$

$$A_d^2 = n^{-1} \sum_{j=1}^{k} Z_j^2 t_j / \{H_j(1 - H_j)\}. \tag{3}$$

note that $Z_k = 0$ in these summations, so that the last term in W_d^2 is zero. The last term in A_d^2 is of the form 0/0, and is set equal to zero.

The well-known Pearson χ^2 statistic is

$$\chi^2 = \sum_{i=1}^{k} (o_i - e_i)^2 / e_i.$$

Statistics corresponding to the Kolmogorov–Smirnov statistics (see ▶Kolmogorov-Smirnov Test) for continuous observations are

$$D_d^+ = \max_j (Z_j)/\sqrt{n}, D_d^- = \max_j (-Z_j)/\sqrt{n},$$

$$D_d = \max_j |Z_j|/\sqrt{n}.$$

Comments on the Definitions

1. Several authors have examined distributions of the Kolmogorov–Smirnov family, see Pettitt and Stephens (1977) and Stephens (1986) for tables and references. In general, for continuous data, the Kolmogorov–Smirnov statistic is less powerful as an omnibus test statistic than the Cramér–von Mises family; limited Monte Carlo studies suggest that this holds also for D_d.
2. The Cramér–von Mises and Kolmogorov–Smirnov statistics take into account the order of the cells, in contrast to the Pearson χ^2 statistic.
3. Use of t_j in these definitions ensures that the value of the statistic does not change if the cells are labelled in reverse order.

   For instance, in testing the ▶binomial distribution, one statistician might record the histogram of successes, and another the histogram of failures; or in a test involving categorical data such as the tones of a photograph, the histogram of cells with light to dark observations might be recorded, or *vice versa*.
4. The statistic U_d^2 is intended for use with a discrete distribution around a circle, since its value does not change with different choices of origin; this is why p_{k+1} is set equal to p_1.

Matrix Formulation

To obtain asymptotic distributions it is convenient to put the above definitions into matrix notation. Let a prime, e.g., Z', denote the transpose of a vector or matrix. Let $\mathbf{I}$ be the $k \times k$ identity matrix, and let p' be the $1 \times k$ vector $(p_1, p_2, \ldots, p_k)$. Suppose $\mathbf{D}$ is the $k \times k$ diagonal matrix whose j-th diagonal entry is $p_j, j = 1, \ldots, k$ and let $\mathbf{E}$ be the diagonal matrix with diagonal entries t_j, and $\mathbf{K}$ be the diagonal matrix whose (j,j)-th element is $K_{jj} = 1/\{H_j(1 - H_j)\}, j = 1, \ldots, k-1$ and $K_{kk} = 0$. Let o_i and e_i be arranged into column vectors $\mathbf{o}$, $\mathbf{e}$ (so that, for example, the j-th component of $\mathbf{o}$ is $o_j, j = 1, \ldots, k$). Then $Z = Ad$, where $d = o - e$ and A is the $k \times k$ partial- sum matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix}.$$

The definitions become

$$W_d^2 = Z'EZ/n, \tag{4}$$

$$U_d^2 = Z'(I - E11')E(I - 11'E)Z/n, \tag{5}$$

$$A_d^2 = Z'EKZ/n, \tag{6}$$

$$X^2 = (d'D^{-1}d)/n = Z'A^{-1'}D^{-1}A^{-1}Z/n. \tag{7}$$

Asymptotic Theory

All Parameters Known

All four statistics above are of the general form $S = Y'MY$, where $Y = Z/\sqrt{n}$ and $\mathbf{M}$ is symmetric. For W_d^2, $M = E$, for U_d^2, $M = (I - E11')E(I - 11'E)$, and for A_d^2, $M = EK$. Also Y has mean 0. Suppose its covariance matrix is $\mathbf{\Sigma}_y$, to be found below; then S may be written

$$S = Y'MY = \sum_{i=1}^{k-1} \lambda_i (w_i' Y)^2, \tag{8}$$

where λ_i are the $k-1$ non-zero eigenvalues of $M\boldsymbol{\Sigma}_y$ and w_i are the corresponding eigenvectors, normalized so that $w_i'\boldsymbol{\Sigma}_y w_j = \delta_{ij}$ where δ_{ij} is 1 if $i=j$ and 0 otherwise.

As $n \to \infty$, the s_i tend to standard normal, and they are independent; the limiting distribution of S is that of S_∞ where

$$sinf S_\infty = \sum_{i=1}^{k-1} \lambda_i s_i^2 \tag{9}$$

which is a sum of independent weighted χ_1^2 variables.

Recall that $Y = Z/\sqrt{n} = Ad/\sqrt{n}$; its covariance $\boldsymbol{\Sigma}_y$ is found as follows. Calculate the $k \times k$ matrix

$$\boldsymbol{\Sigma}_0 = D - pp'; \tag{10}$$

this is the covariance matrix of $(o-e)/\sqrt{n}$. Then $\boldsymbol{\Sigma}_y = A\boldsymbol{\Sigma}_0 A'$, with entries $\boldsymbol{\Sigma}_{y,ij} = \min(H_i, H_j) - H_i H_j$.

For the appropriate M for the statistic required, the eigenvalues λ_i, $i = 1, \cdots, k$ of $M\boldsymbol{\Sigma}_y$ are used in (9) to obtain the limiting distribution of the statistic. The limiting distributions have been examined in detail in Choulakian et al. (1994).

Parameters Unknown

Cramér–von Mises statistics when the tested distribution contains unknown parameters θ_i have been investigated by Lockhart et al. (2007). The θ_i must be estimated efficiently, for example by maximum likelihood (ML). Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)'$ is the vector of m parameters.

The log-likelihood is (omitting irrelevant constants)

$$L^* = \sum_{i=1}^{k} o_i \log p_i,$$

and p_i contains the unknown parameters. The ML estimation consists of solving the m equations

$$\frac{\partial L^*}{\partial \theta_j} = \sum_{i=1}^{k} \frac{o_i}{p_i} \frac{\partial p_i}{\partial \theta_j} = 0,$$

for $j = 1, \ldots, m$.

Let $\hat{\boldsymbol{\theta}}$ be the ML estimate of $\boldsymbol{\theta}$, let $\hat{p}$ be the estimate of p, evaluated using $\hat{\boldsymbol{\theta}}$, and let $\hat{e}$ be the estimated vector of expected values in the cells, with components $\hat{e}_j = n\hat{p}_j$. Then let $\hat{d} = (o - \hat{e})$ and $\hat{Z} = A\hat{d}$.

Define a k by m matrix B with entries

$$B_{i,j} = \partial p_i / \partial \theta_j$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, m$. The matrix $B'D^{-1}B$ is the Fisher Information matrix for the parameter estimates. Define $V = (B'D^{-1}B)^{-1}$. The asymptotic covariance of $\hat{\boldsymbol{\theta}}$ is then V/n, the covariance of $\hat{d}/\sqrt{n}$ is $\boldsymbol{\Sigma}_d = \boldsymbol{\Sigma}_0 - BVB'$, where $\boldsymbol{\Sigma}_0$ is defined in (10), and the covariance of $\hat{Z}/\sqrt{n} = A\hat{d}/\sqrt{n} = \hat{Y}$ is

$$\boldsymbol{\Sigma}_u = A\boldsymbol{\Sigma}_d A'.$$

Then, as in the previous section, where parameters were known, the weights λ_i in the asymptotic distribution (9) are the k eigenvalues of $M\boldsymbol{\Sigma}_u$ for the appropriate M for the statistic required.

In practice, in order to calculate the statistics, using (4–7), the various vectors and matrices must be replaced by their estimates where necessary. For example, let matrix $\hat{D}$ be D with p replaced by $\hat{p}$ and similarly obtain $\hat{B}, \hat{E}, \hat{V}, \hat{K}$ and $\hat{\boldsymbol{\Sigma}}_0$ using estimates in an obvious way. The eigenvalues will also be found using the estimated matrices $\hat{\boldsymbol{\Sigma}}_u$ and $\hat{M}$. Consistent estimates of the λ_i will be obtained and (9) used to find the estimated asymptotic distribution.

Thus the steps are :

1. Calculate $\hat{V} = (\hat{B}'\hat{D}^{-1}\hat{B})^{-1}$.
2. Calculate $\hat{\boldsymbol{\Sigma}}_d = \hat{\boldsymbol{\Sigma}}_0 - \hat{B}\hat{V}\hat{B}'$ and $\hat{\boldsymbol{\Sigma}}_u = A\hat{\boldsymbol{\Sigma}}_d A'$.
3. For the statistic required, let $\hat{M}$ be the estimate of the appropriate M. Find the $k-1$ eigenvalues of $\hat{M}\hat{\boldsymbol{\Sigma}}_u$, or (equivalently) those of the symmetric matrix $\hat{M}^{1/2}\hat{\boldsymbol{\Sigma}}_u\hat{M}^{1/2}$ and use them in (9) to obtain the asymptotic distribution.

For practical purposes, percentage points of S_∞ using exact or estimated λs, can be used for the distributions of the statistics for finite n; this has been verified by many Monte Carlo studies. One therefore needs good approximate points in the upper tail of S_∞; these can be found from the percentage points of $S1$, where $S1$ has the distribution $a + b\chi_p^2$, and the a, b, p are chosen so that the first three cumulants of $S1$ match those of S_∞ in (9). These cumulants are $\kappa_j = 2^{j-1}(j-1)! \sum_{i=1}^{k-1} \lambda_i^j$. In particular, the mean κ_1 is $\sum_{i=1}^{k-1} \lambda_i$, the variance κ_2 is $\sum_{i=1}^{k-1} 2\lambda_i^2$ and κ_3 is $8\sum_{i=1}^{k-1} \lambda_i^3$. Then for the $S1$ approximation, $b = \kappa_3/(4\kappa_2)$, $p = 8\kappa_2^3/\kappa_3^2$, and $a = \kappa_1 - bp$. This approximation is generally accurate in the upper tail, at levels $\alpha < 0.15$. More accurate points can be obtained by the method of Imhof (1961).

About the Author

Michael A. Stephens is Professor Emeritus of Mathematics and Statistics at Simon Fraser University in Burnaby, British Columbia, Canada. Prior to that he taught at several universities including McGill, Nottingham, McMaster, and Toronto, and was a visiting professor at Stanford, Wisconsin-Madison, and Grenoble. He has (co-)authored over 100 papers on the analysis of directional data, continuous proportions, curve-fitting, and tests of fit. Professor Stephens was President of the Statistical Society of Canada in 1983. He is a Fellow of the Royal Statistical Society, and

his honors include membership in the International Statistical Institute, and fellowships of the American Statistical Association and the Institute of Mathematical Statistics. Dr. Stephens received the B.Sc. degree (1948) from Bristol University and A.M. degree (1949) in physics from Harvard University, where he was the first Frank Knox Fellow, and Ph.D. degree (1962) from the University of Toronto. In 1989 he was awarded the Gold Medal, Statistical Society of Canada for two main areas of research: analysis of directional data, and statistical theory and methods associated with goodness of fit.

Cross References

►Anderson-Darling Tests of Goodness-of-Fit
►Exact Goodness-of-Fit Tests Based on Sufficiency
►Kolmogorov-Smirnov Test
►Tests of Fit Based on The Empirical Distribution Function

References and Further Reading

Anderson TW (2010) Anderson–Darling tests of goodness-of-fit. Article in this encyclopedia
Anderson TW, Darling DA (1952) Asymptotic theory of certain goodness of fit criteria based on stochastic processes. Ann Math Stat 23:193–212
Choulakian V, Lockhart RA, Stephens MA (1994) Cramer–von Mises Tests for discrete distributions. Can J Stat 22:125–137
Darling DA (1955) The Cramér–Smirov test in the parametric case. Ann Math Stat 26:1–20
Imhof JP (1961) Computing the distribution of quadratic forms in normal variables. Biometrika 48:419–426
Lockhart RA, Spinelli JJ, Stephens MA (2007) Cramér–von Mises statistics for discrete distributions with unknown parameters. Can J Stat 35:125–133(9)
Pettitt AN, Stephens MA (1977) The Kolmogorov–Smirov test for discrete and grouped data. Technometrics 19(2):205–210
Stephens MA (1976) Asymptotic results for goodness-of-fit statistics with unknown parameters. Ann Stat 4:357–369
Stephens MA (1986) Tests based on EDF statistics. In: D'Agostino R, Stephens MA (eds) Chap. 4 in Goodness-of-fit techniques. Marcel Dekker, New York
Stephens MA (2010) EDF tests of fit. Article in this encyclopedia

Cross Classified and Multiple Membership Multilevel Models

Harvey Goldstein
Professor of Social Statistics
University of Bristol, Bristol, UK

Hierarchically Structured Data

Interesting real life data rarely conform to classical text book assumptions about data structures. Traditionally these assumptions are about observations that can be modelled with independently, and typically identically, distributed "error" terms. More often than not, however, the populations that generate data samples have complex structures where measurements on data units are not mutually independent, but depend on each other through complex structural relationships. For example, a household survey of voting preferences will typically show variation among households and voting constituencies (constituencies and households differ on average in their political preferences). This implies that the replies from individual respondents within a household or constituency will be more alike than replies from individuals in the population at large. Another example of such "hierarchically structured data" would be measurements on students in different schools, where, for example, schools differ in terms of the average attainments of their students. In epidemiology we would expect to find differences in such things as fertility and disease rates across geographical and administrative areas.

Techniques for modelling such data have come to be known as "multilevel" or "hierarchical data" models and basic descriptions of these are dealt with in other articles (*see* ►Multilevel Analysis). In the present article we shall consider two particular extensions to the basic multilevel model that allow us to fit structures that have considerable complexity and are quite commonly found, especially in the social and medical sciences.

The Basic Multilevel Model

A simple multilevel model for hierarchical data structures with normally distributed responses can be written as:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_j + e_{ij},\ u_j \overset{iid}{\sim} N(0, \sigma_u^2),\ e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2). \quad (1)$$

This might be applied to a sample, say, of school students where i indexes students (level 1), who are grouped within schools (level 2). The response y might be an attainment measure and x a predictor such as a prior test score. Often referred to as a "variance components" model this may be extended in a number of ways to better fit a data set. For example, we may introduce further covariates and we may allow the coefficients of such covariates to vary at level 2, so that, say, β_1, may vary from school to school. Another possibility is to allow the level 1 variance to depend on a set of explanatory variables, so that, for example, we can allow the variance between male students to be different from that for female students. We can have several responses that are correlated leading to a multivariate model, and we can consider non-normal responses, such as binary ones, in order to fit generalised linear multilevel models. We can also have several further levels; for example schools may be

grouped within school boards or authorities, so yielding a three level structure. Goldstein (2010) provides further details and discusses estimation methods.

Cross Classified Structures

The above only describes purely hierarchical models. In practice, however, data structures are often more complicated. Consider an educational example where students move through both their primary and secondary education with the response being attainment at the end of secondary school. For any given primary school, students will generally move to different secondary schools, and any given secondary school will draw students from a number of primary schools. We therefore have a *cross classification* of primary by secondary schools where each cell of the classification will be populated by students (some may be empty). When we model such a structure we have a contribution to the response that is the sum of an effect from the primary and an effect from the secondary school attended by a student. A basic, variance components, cross classified model may be written as

$$\begin{aligned} y_i^{(1)} &= \beta_0 + \beta_1 x_{1i} + u^{(2)}_{\text{primary school}(i)} + u^{(3)}_{\text{secondary school}(i)} \\ &\quad + u^{(1)}_{\text{student}(i)} \\ u^{(2)}_{\text{primary school}(i)} &\overset{iid}{\sim} N(0, \sigma^2_{u(2)}), \\ u^{(3)}_{\text{secondary school}(i)} &\overset{iid}{\sim} N(0, \sigma^2_{u(3)}) \\ u^{(1)}_{\text{student}(i)} &\overset{iid}{\sim} N(0, \sigma^2_{u(1)}),\ i = 1, \ldots, N. \end{aligned} \tag{2}$$

We have changed the notation to make it more general and flexible. The superscript refers to the set of units, or classification; 1 being students, 2 primary school and 3 secondary school. Model (2) thus assumes that there are separate, additive, contributions from the primary and the secondary school attended. As with the simple hierarchical model we can extend (2) in several ways by introducing random coefficients, complex variance structures and further cross classifications and levels. There are many examples where cross classified structures are important. Thus, for example, students will generally be grouped by the neighborhood where they live and this will constitute a further classification. In a repeated measures study where there is a sample of subjects and a set of raters or measurers, if the subjects are rated by different people at each occasion we would have a cross classification of subjects by raters.

Multiple Membership Structures

In many circumstances units can be members of more than one higher level unit at the same time. An example is friendship patterns where at any time individuals can be members of more than one friendship group. In an educational system students may attend more than one school over time. In all such cases we shall assume that for each higher level unit to which a lower level unit belongs there is a known weight (summing to 1.0 for each lower level unit), which represents, for example, the amount of time spent in the higher level unit. The choice of weights may be important but is beyond the scope of this article. For more details about choosing weights see Goldstein et al. (2007).

Using the general notation we used for cross classifications we can write a basic variance components multiple membership model as

$$\begin{aligned} y_i^{(1)} &= \beta_0 + \beta_1 x_i + \sum_{j \in school(i)} w^{(2)}_{i,j} u^{(2)}_{(j)} + u^{(1)}_i \\ u^{(2)}_{(j)} &\sim N(0, \sigma^2_{u(2)}),\ u^{(1)}_{(i)} \sim N(0, \sigma^2_{u(1)}) \\ \sum_{j \in school(i)} w^{(2)}_{i,j} &= 1. \end{aligned} \tag{3}$$

This assumes that the total contribution from the level 2 units (schools) is a weighted sum over all the units of which the level 1 unit has been a member. Thus, for example, if every student spends half their time in one school and half their time in another (randomly selected) then the variance at level 2 will be

$$\text{var}(0.5u^{(2)}_{j_1} + 0.5u^{(2)}_{j_2}) = \sigma^2_{u(2)}/2. \tag{2}$$

Thus, a failure to account for the multiple membership of higher level units in this case will lead us to treat the estimate of the level 2 variance, $\sigma^2_{u(2)}/2$ as if it were a consistent estimate of the true level 2 variance $\sigma^2_{u(2)}$. More generally, ignoring a multiple membership structure will lead to an underestimation of the higher level variance.

Finally, we can combine cross classified and multiple membership structures within a single model and this allows us to handle very complex structures. An example where the response is a binary variable is given in Goldstein (2010, Chap. 13). It is possible to use maximum likelihood estimation for these models, but apart from small scale datasets, MCMC estimation is more efficient and flexible. The MLwiN software package (Rasbash et al. 2009; Browne 2009. http://www.cmm.bristol.ac.uk) is able to fit these models.

About the Author

Professor Goldstein is a chartered statistician, has been editor of the *Royal Statistical Society's Journal, Series A*, a member of the Society's Council and was awarded the Society's Guy medal in silver in 1998. He was elected a member of the International Statistical Institute in 1987, and a Fellow of the British Academy in 1996. He was awarded an honorary doctorate by the Open University in 2002.

His most important research interest is in the methodology of multilevel modelling. He has had research funding for this since 1986 and has been involved in the production (with Jon Rasbash and William Browne) of a widely used software package (MLwiN) and made a number of theoretical developments. These include multiple membership models, multilevel structural equation models and more recently the use of multilevel multivariate latent normal models and especially their application to missing data problems. The major text on multilevel modelling is his book *Multilevel Statistical Models* (New York, Wiley, Fourth Edition, 2010). He has also written extensively on the use of statistical modelling techniques in the construction and analysis of educational tests. The implications of adopting such models have been explored in a series of papers since 1977.

Cross References

►Multilevel Analysis
►Psychology, Statistics in

References and Further Reading

Browne WJ (2009) MCMC estimation in MLwiN. Version 2.10. Bristol, Centre for Multilevel Modelling, University of Bristol

Goldstein H (2010) Multilevel statistical models, 4th edn. New York, Wiley

Goldstein H, Burgess S, McConell B (2007) Modelling the effect of pupil mobility on school differences in educational achievement. J R Stat Soc Ser A 170(4):941–954

Rasbash J, Steele F, Browne W, Goldstein H (2009) A user's guide to MLwiN version 2.10. Bristol, Centre for Multilevel Modelling, University of Bristol

Cross-Covariance Operators

Charles R. Baker
Professor Emeritus
University of North Carolina, Chapel Hill, NC, USA

This article will initially treat joint probability measures and their associated cross-covariance operators. Subsequently, attention will be shifted to three examples of problems on capacity of information channels.

Cross-covariance operators were introduced in Baker (1970) as a tool in solving a basic problem in information theory, and treated more extensively in Baker (1973). Related results are in Gualtierotti (1979) and Fortet (1995). The empahsis in Baker (1973) was in two directions: showing the added power of analysis obtained by introducing the cross-covariance operator of a joint measure, and providing new results for actually computing likelihood ratios for joint measures. Applications to date have included results on absolute continuity of probability measures, mutual information for pairs of ►stochastic processes, and analysis of information capacity for communication channels. More recently, there has been interest in this topic by researchers in machine learning, who have applied theory from Baker (1973) in a number of interesting publications (e.g., Fukumizu et al. 2004, 2009; Gretton et al. 2005).

The joint measures to be discussed are probability measures on the product of two real separable Hilbert spaces, H_1 and H_2, with Borel sigma fields θ_1. and θ_2. Denote the inner products by $<,>_1$ on H_1 and $<,>_2$ on H_2. H_1 x H_2 is then a real separable Hilbert space under the inner product defined by $<(x,u),(v,y)>_{12} = <x,v>_1 + <u,y>_2$. Next, introduce a joint measure π_{12} on the measurable space (H_1 x H_2, θ_1 x θ_2). Only strong second-order probability measures will be considered: those joint measures ∂_{12} such that $E_{\partial_{1,2}}|||(x,y)||||^2_{12} = \int_{H_1xH_2}(||x||^2_1 + ||y||^2_2)d\partial_{12}(x,y)$ is finite. All Gaussian measures on H_1xH_2 are strong second order, as are their projections on H_1 and H_2. From the measure π_{12} one has projections π_i on (H_i, θ_i), $i = 1,2$. Let m_1 and m_2 denote the mean elements and R_1 and R_2 the covariance operators of π_1 and π_2.

The first result of note is the definition and properties of the cross-covariance operator for the joint measure π_{12}. Denoting that operator by C_{12}, it is defined for all (u,v) in H_1xH_2 by

$$<C_{12}v,u>_1 = \int_{H_1xH_2} <x - m_1,u>_1 <y - m_2,v>_2 \, d\pi_{12}(x,y).$$

Theorem 1 *C_{12} has representation $C_{12} = R_1^{1/2}VR_2^{1/2}$, V: $H_2 \to H_1$ a unique linear operator having $||V|| \le 1$ and $P_1VP_2 = V$, P_i the projection of H_i onto the closure of range(R_i).* ⊡

Next, we turn to the definition and properties of the covariance operator $\mathbf{R}_{12}$ of π_{12}. By direct computation (Baker 1973), one can show that this operator is defined on every element (u,v) in H_1xH_2 by

$$\begin{aligned}\mathbf{R}_{12}(u,v) &= (R_1u + C_{12}v, R_2v + C_{12}^*u)\\ &= (R_1 \otimes R_2)(u,v) + (C_{12}^* \otimes C_{12})(u,v).\end{aligned}$$

We now give a result that illustrates both similarity and difference between a joint measure and the usual measure.as defined on one of the spaces H_1 or H_2. We define a self-adjoint operator $\mathbf{V}$ in H_1xH_2 by $\mathbf{V}(u,v) = (Vv, V^*u) = (V^* \otimes V)(u,v)$ and denote by $\mathbf{I}$ the identity operator in H_1xH_2; it is shown in Baker (1973) that $||\mathbf{V}|| \le 1$ and that the non-zero eigenvalues of VV^* are squares of the non-zero eigenvalues of $\mathbf{V}$.

Theorem 2 *The covariance operator* $\boldsymbol{R}_{12}$ *of the measure* π_{12} *on* $H_1 x H_2$ *has representation* $\boldsymbol{R}_{12} = \boldsymbol{R}_{1\otimes 2}^{1/2}(\boldsymbol{I}+\boldsymbol{V})\boldsymbol{R}_{1\otimes 2}^{1/2}$, *where* $\boldsymbol{R}_{1\otimes 2}$ *is the covariance operator of the product measure* $\pi_1 \otimes \pi_2$. *If* π_{12} *is Gaussian, then* π_{12} *and* $\pi_1 \otimes \pi_2$ *are mutually absolutely continuous if and only if* $\boldsymbol{V}$ *is Hilbert-Schmidt with* $||\boldsymbol{V}|| < 1$, *and otherwise orthogonal.*⊡

The preceding results give some of the basic properties of the covariance operator of a joint measure, and it is seen that the cross-covariance operator is an essential component in the definition and properties of the covariance operator. In Baker (1973), considerable attention is given to Gaussian joint measures. However, it should be noted that the definition of the cross-covariance operator and its relation to the covariance operator hold for any strong-second order joint probability measure When the joint measure at hand is not Gaussian, one still has the cross-covariance operator available as well as the mean and the covariance operator. These functions can frequently be estimated from data and used to develop suboptimum but effective operations using (for example) second moment criteria.

We now turn to a brief introduction to three problems on the capacity of a Gaussian channel without feedback (Baker 1978, 1987; Baker and Chao 1996a, b). The cited papers provide examples of the use of results from Baker (1973) in applications to information theory. The definition of the channel capacity is as follows. We have a joint measure $\pi_{S,AS+N}$ where S is the actual signal, AS is the transmitted coded signal (from a measurable space (Ω, Θ)) and AS+N is the received waveform of signal+noise from a measurable space (Ψ, Γ). The (average) mutual information will be finite if $\pi_{S,AS+N}$ is absolutely continuous with respect to its product measure $\pi_{S\otimes AS+N}$, and its value is then given by

$$\int_{\Omega x \Psi} \log[(d\pi_{S,AS+N}/d\pi_{S\otimes AS+N})(x,y)]d\pi_{S,AS+N}(x,y).$$

The transmitted signal AS and the received AS+N can vary with choices by the coder (and the jammer in the third example below), and the channel capacity is the supremum of the mutual information over all admissible S and AS+N.pairs.

In each case, the transmitted signal has a constraint given in terms of the ambient noise process. When the constraint on the transmitted signal is given in terms of the channel noise covariance, one says that the channel is "matched" (coder constraint is matched to the channel noise covariance) (Baker 1978). The second type of channel is "mismatched" (the signal constraint is not given in terms of the channel noise covariance) (Baker 1987). The third class is the jamming channel without feedback, wherein the noise in the channel consists of a known Gaussian ambient noise (nature's contribution) plus an independent noise that is under the control of a hostile jammer (Baker and Chao 1996a, b). In this channel, there is a constraint on the jammer's noise as well as one on the coder's.transmitted signal.

In the jamming channel, the jammer has no constraints on the choice of the probability distributions of the noise at his command. However, it is known (Ihara 1978) that if the channel noise due to nature is Gaussian, then the information capacity is minimized by the jammer choosing (among all processes satisfying the constraints) a Gaussian process. Thus, the original problem becomes a problem involving an ambient Gaussian noise (which is used to calculate the coder's constraint) and an independent Gaussian process (jamming) giving the covariance constraint that the jammer uses.

Cross References

►Canonical Analysis and Measures of Association
►Measure Theory in Probability
►Statistical Signal Processing
►Statistical View of Information Theory

References and Further Reading

Baker CR (1970) Mutual information for Gaussian processes. SIAM J Appl Math 19(2):451–458
Baker CR (1973) Joint measures and cross-covariance operators. Trans Am Math Soc 186:273–289
Baker CR (1978) Capacity of the Gaussian channel without feedback. Inf Cont 37:70–89
Baker CR (1987) Capacity of the mismatched Gaussian channel. IEEE Trans Inf Theory 33:802–812
Baker CR, Chao IF (1996a) Information capacity of channels with partially unknown noise. I. Finite-dimensional channels. SIAM J Appl Math 56:946–963
Baker CR, Chao IF (1996b) Information capacity of channels with partially unknown noise. II. Infinite-dimensional channels. SIAM J Cont Optimization 34:1461–1472
Fortet RM (1995) Vecteurs, fonctions et distributions aleatoires dans les espaces de Hilbert. Hermes, Paris (see pp. 331 ff.)
Fukumizu K, Bach FR, Jordan MJ (2004) Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. J Mach Learning Res 5:73–99
Fukumizu K, Bach FR, Jordan MJ (2009) Kernel dimensionality reduction in regression. Ann Stat 37:1871–1905
Gretton A, Bousquet O, Smola AJ, Schölkopf B (2005) Measuring statistical dependence with Hilbert–Schmidt norms. MPI Technical Report, Max Planck Institute for Biological Cybernetics. Tübingen, Germany' Report 140
Gualtierotti AF (1979) On cross-covariance operators. SIAM J Appl Math 37(2):325–329
Ihara S (1978) On the capacity of channels with additive non-Gaussian noise. Inf Cont 37:34–39

D

Data Analysis

Alfredo Rizzi
Professor
Sapienza Università di Rome, Rome, Italy

The Development of Data Analysis

Data analysis began to be developed for use in statistical methodology in the early 1960s. Today, it includes a number of techniques which allow an acceptable synthesis of information collected from n statistical units or objects which are each characterized by p qualitative or quantitative variables.

The basics of certain data analysis techniques were established in the last century, principal components at the beginning of the 1900s, factor analysis in the 1930s, and automated classification methods in the 1940s.

During the same period, statistical methodology also saw development in statistical inference (classic and Bayesian) and in studies of interpretative models of complex phenomena (linear, loglinear, generalized linear, analysis of time series such as ARMA, ARIMA, autoregressive, etc.). Data analysis has made use and at times stimulated the development of two important "tools": the language of matrices and appropriate computer software requiring hardware with large memory capacity and very fast access to information.

The development of data analysis has been stimulated by the operational needs of its application in various sectors, in particular, business and the social sciences. For example, automatic classification methods are interesting applications used for classifying clients and market segments and for establishing homogeneity between company units and territorial administrative divisions. These methods are also applied in medicine for pattern recognition (see ►Pattern Recognition, Aspects of and ►Statistical Pattern Recognition Principles), for the automatic assessment of words used by speakers in different situations, as well as in many other sectors. Cluster analysis (see ►Cluster Analysis: An Introduction), when used in sample surveys of a population with known structural characteristics, allows one to reduce the variability within an equal sample size, thereby increasing precision. Multiway matrix analysis is especially interesting when seeking trends in a single location at different times or in different locations at the same time.

Methods such as ►multidimensional scaling allow us to represent a set of n objects belonging to a space with $p > 3$ dimensions in spaces of two or three dimensions, so that the distortion of the matrix of the distance between points is minimal. ►Principal component analysis has also been the subject of in-depth studies which have opened the way for its use in a wide variety of fields. ►Correspondence analysis, introduced by Benzécri at the end of the 1960s, is considered to be the acme of data analysis. The French school has, in some way, introduced this new sector of statistics to scholars and has carried out research in social sciences, economics, ►demography, and business using methodologies which have proved to be very useful in many research situations.

Data analysis uses not only the matrix language and advanced software mentioned above but also methods of mathematical optimization, the theory of eigenvalues (see ►Eigenvalue, Eigenvector and Eigenspace) which are of great importance in many analyses, certain results from vector analysis, the theory of graphs and operations research along with some theorems of mathematical analysis, and the theory of linear spaces and fuzzy sets. Important and interesting contributions to classification have been made by scientists studying mathematical optimization methods, who have introduced rigorous algorithms to establish the partitions of a set of n objects characterized by p variables, which constitute the statistical population or group of reference.

The studies carried out by Minkowski and Fréchet and other mathematicians on metrics are applicable to the realm of data analysis. Euclidean metrics is one of many but is often preferred by certain software for reasons of tradition. The *L1* norm, for example, and the geometry based on it is, in many research situations, better suited than other metrics to represent complex phenomena in various fields. But metrics in the traditional mathematical sense have proved completely inadequate for the applications.

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

Therefore, these interesting studies of distance and similarity indices and, most importantly, of ultra-metrics have become an important part of data analysis.

It is also of great importance in data analysis to be very careful about conclusions drawn from surveys which have been carried out using methodologies which are not justified by the hypotheses on which the research is based. The great progress and widespread availability of hardware and software has made it possible for those who are not students or specialists in the field to carry out research. This may lead to data processing done without in-depth knowledge of the methodology, techniques, or parametrization used; so, the default settings are chosen by the software without an understanding of the meaning of many statistical parameters which allow an a posteriori evaluation of the goodness of the data model adopted.

The data coding step is of fundamental importance for the evaluation of any outcome although it is often given less attention than needed by those preferring to concentrate on the mathematical aspects of the analysis.

Inferential problems also arise in data analysis. Data analysis has, in the first stage, been viewed only as a methodology for synthesizing information with any probabilistic or inferential approach excluded. Attempts to apply classical inferential methodologies requiring strong distribution hypotheses, especially in the multidimensional field, have not led to any worthwhile outcome. In some techniques – such as principal components and factor analysis – inferential aspects were considered to be of primary importance and were the subject of wide studies during the first half of the last century.

As is often the case in scientific research, techniques used in statistical induction have proved to be unsuited to solving inferential problems in data analysis. Furthermore, in this field, induction has a wider meaning that is merely passing from "the sample to the population," both in terms of an estimate of the parameters and of a verification of the hypotheses. The themes of fundamental importance here are those concerning the stability of results in terms of knowing to what extent the relations found between individuals and/or between variables can be considered valid. The relations between variables – linear and nonlinear – are often very strong.

The verification of the hypotheses and the estimation of the parameters are not specific interests. One is nearly always concerned with data gathered from populations of a finite number of individuals, which makes the stability of the results extremely important. In sampling, on the other hand, it is important to carefully consider what the expected outcome is and what the meaning of the extension is. We are unaware of any studies of the sampling plans capable of guaranteeing an acceptable level of representation of the samples extracted. This is also a completely new field of research regarding the methodological instruments usable in this important field.

The Techniques of Data Analysis

The expression "data analysis" (*analyse des données in French*) or multidimensional data analysis (MDA) originated from the methodological approach first used in France in the late 1960s (Benzécri 1973; Bertier and Bouroche 1975; Caillez and Pages 1976, etc.). It includes two groups of multivariate statistical methods: classification methods or automatic clustering and linear data analysis and specifically principal component analysis, canonic correlation analysis, both simple and multiple correspondence analysis, and multidimensional scaling.

During the 1980s, MDA spreaded to other countries of Europe and was established as an autonomous branch of statistics. This created the conditions for new developments in both methodology and in applications, which characterize recent developments in multidimensional data analysis. There were important contributions in this area from Italian statisticians.

In addition to those mentioned above, there are other methods available for multivariate statistical analysis: regression, ►analysis of variance, discriminant analysis (see ►Discriminant Analysis: An Overview and ►Discriminant Analysis: Issues and Problems), common factor analysis, etc., in which probabilities and inferential aspects are treated. These may be called validation analyses for assumptions formulated about the multidimensional data set.

Inferential and validation aspects come into play in MDA and can also be introduced, with some conditions, into the methods. In general, the various methods of ►multivariate statistical analysis have different ways, depending on the purpose, to analyze, describe, and synthesize the relations between statistical characteristics and statistical units or cases.

It is useful to this purpose to divide the different methods of multivariate statistical analysis according to the role of the variables that come into play in the analysis. In the original MDA approach, the methods used for multivariate statistical analysis could be divided into:

Symmetric Analyses: with variables of the same role (interdependence). The principals are:

- Cluster analysis
- Principal component analysis
- Canonic correlation analysis
- Factor analysis of correspondences
- Common factor analysis
- Symmetric multidimensional scaling

Asymmetric Analyses: where some variables are explicative, independent attributes of other variables, dependent and endowed with attributes or where pairs of variables or units are ranked as in:

- Regression
- Analysis of variance
- Discriminant analysis
- Nonsymmetrical multidimensional scaling
- Path analysis

In the context of recent developments in MDA, such a split is no longer sharp because asymmetric variants have been introduced in all the main methods of data analysis. For example, in principal component analysis, the supplemental or passive variables can be treated as variables in a reference subspace instead of being handled as supplemental elements in the plots of the principal axes without having been used in the actual computations. Asymmetric variants have also been introduced in simple and multiple correspondence analysis.

Multiway data analysis concerning particular extensions of the methods of data analysis is linked to asymmetric data analysis because of the different role it assigns to variables. Instead of the traditional data classification according the two criteria (statistical units *x* variables), these methods are applied, with different approaches, to data classified according to three or more criteria or statistical ways (statistical units *x* variables *x* occurrences).

The combination of multivariate statistical methods and the automatic data processing necessary for this type of analysis is what brought about MDA. MDA is an important tool of ▶data mining. One can consider data mining to be a procedure that starts from elementary data in a Data Warehouse to arrive at a decision. One can consider data mining to be the heir of artificial intelligence and expert systems. In each case, MDA is the basis for discovering data structure.

Three-Way Data Analysis

Since the last 2 decades of the twentieth century, there has been growing attention paid to the analysis of phenomena characterized by a set **X** of *nkr* variables, observed on a set of *n* units on *r* occasions (different times, places, etc.). These phenomena must be distinguished from the classical multivariate and are referred to as multivariate-multioccasion. The set X is called a three-way data set because its elements can be classified according to modes: units, variables, and occasions. The data structure associated with this data set is a *three-indices* or *three-way array*.

The most widely collected three-way data set occurs when modes are units and variables and occasions are different times. Here, we will especially refer to this data type. The repetition in time of the observation of the units allows us to evaluate the dynamics of the phenomenon differently from the classical case of a multivariate or cross-sectional (two-way) data set. There are several major advantages over a conventional cross-sectional or time series data set in using these, so-called three-way longitudinal data. The researcher has a larger number of data which increases the degree of freedom and reduces collinearity among explanatory variables and the possibility of making inferences about the dynamics of change from cross-sectional evidence.

The three-way longitudinal data set may be given by:

- *Repeated recurring surveys with no overlapping units*. That is, a survey organization repeats a survey on a defined topic, generally at regular time intervals. No overlaps of the sample units are required at different times. Examples of these surveys are given by the repeated analyses made by Central Bureau of Statistics in most countries.
- *Repeated surveys with partially overlapping units*. These surveys are also repeated at regular intervals. The survey design includes rotating units to allow variance reduction, i.e., the units are included in the analysis a number of times, then rotated out of the survey.
- *Longitudinal surveys with no rotation of units*. A set of units is followed over time with a survey designed with this aim. In the economic field, these collected data are called ▶*panel data*. An example of this type is the Survey of Income.
- *Longitudinal surveys with rotation*. A group of units is followed for a period of time, so that new units are introduced and followed over time. These are repeated surveys with a partially overlapping units. An example of this type is the Monthly Retail Trade Survey.

In longitudinal surveys, it is generally the same variable which is observed on the units to allow comparison over time.

Examples of the applications of the analyses of three-way data sets are found in many different fields, including chemometrics*, biology, economics.

Different objectives can be pursued in analyzing three-way data sets:

1. To synthesize a three-way data set
2. To study the multiway variability of a three-way data set
3. To study the interrelationships between sets of elements of **X**:
   (a) Between sets of variates

(b) Between sets of units
(c) Between sets of occasions

4. To define virtual modes or derivational modes for the three sets of modes of **X**:
   (a) Virtual units
   (b) Factors or latent variables
   (c) Virtual occasions
5. To analyze individual differences in the judgments of different sets dissimilarities:
   (a) Three-way scaling of a replicated set of dissimilarities (replicated multidimensional scaling): metric and nonmetric Euclidean models
   (b) Weighted multidimensional scaling: metric and nonmetric weighted Euclidean models
6. To simultaneously classify different sets of the elements of **X**:
   (a) Clustering of units
   (b) Clustering of variables
   (c) Clustering of occasions
7. To give a linear or nonlinear model of **X**

About the Author

Professor Alfredo Rizzi is past President of the Società Italiana di Statistica (1992–1996). He was Head of the Department of Statistica, Probabilità e Statistiche Applicate, Sapienza università di Roma (1981–1985). He was full professor of Teoria dell'inferenza statistica (1975–2008), Facoltà di Scienze statistiche, Sapienza università di Roma. He was an elected member of the Italian Consiglio Nazionale delle Ricerche (CNR) (1987–1994). Professor Rizzi was a member of the Italian council of Istituto Nazionale di Statistica (1991–1999). He was the chairman of the Scientific and Organizing Committee of the 7th and 17th Symposia in Computational Statistics (Compstat) held in Rome in 1986 and 2006, and the 6th conference of the International Federation of Classification Societies (IFCS), held in Rome in 1998. Dr Rizzi has authored and co-authored more than 130 papers and 7 books including *Inferenza Statistica* (Utet, 1991).

Cross References

►Analysis of Areal and Spatial Interaction Data
►Analysis of Multivariate Agricultural Data
►Categorical Data Analysis
►Exploratory Data Analysis
►Functional Data Analysis
►Fuzzy Logic in Statistical Data Analysis
►Interactive and Dynamic Statistical Graphics
►Multi-Party Inference and Uncongeniality
►Multivariate Data Analysis: An Overview
►Statistical Analysis of Longitudinal and Correlated Data
►Statistical Fallacies: Misconceptions, and Myths
►Statistics Education

References and Further Reading

Agresti A (2007) An Introduction to categorical data analysis, 2nd edn. Wiley, Hoboken
Benzécri JP (1973) L'analyse des données, Tome 1: Taxinomie; Tome II Analyse des correspondances. Dunod, Paris
Bertier P, Bouroche JM (1975) Analyse des données, multidimensionel, Ed. P.U.F, France
Billard L, Diday E (2007) Symbolic data analysis. Conceptual statistics and data mining. Wiley, New York
Caillez F, Pages JP (1976) Introduction á l'analyse des données, smash, Paris
Chatfield C, Collins AJ (1980) Introduction to multivariate analysis. Chapman & Hall, London
Greenacre MJ (1984) Theory and applications of correspondence analysis. Academic, London
Lebart L, Morineau A, Piron M (2006) Statistique exploratoire multidimensionelle. Dunod, Paris

Data Depth

Reza Modarres
Chair and Professor of Statistics
The George Washington University, Washington, D.C., USA

Data depth provides a systematic approach to order multivariate observations in R^d. A data depth function is any function $D(\mathbf{t}; F)$ that measures the closeness or centrality of a point $\mathbf{t} \in R^d$ with respect to a distribution function F. Thus, a depth function assigns to each $x \in R^d$ a nonnegative score as its center-outward depth with respect to F. Observations close to the center of F receive high ranks whereas peripheral observations receive low ranks. Hotelling (1929) characterized the univariate median as the point which minimizes the maximum number of observations on one of its sides. That is, $D(\mathbf{t}; F) = \min(F(t), 1 - F(t))$. This notion was generalized by Tukey (1975), giving rise to the definition of half-space or Tukey's multivariate depth (Donoho 1982). The seminal work of Oja (1983) and Liu (1990) has renewed interest in data depth functions. Over the past two decades, various new notions of data depth have emerged as powerful explanatory and inferential tools for nonparametric multivariate analysis. Zuo and Serfling (2000) proposed a formal framework for statistical depth functions based on the following four properties.

Let F be the class of distributions on the Borel set on R^d and $F_{\mathbf{X}}$ be the distribution function of a given random

vector $\mathbf{X}$. Let $D : R^d \times \mathrm{F} \longmapsto R$ be bounded, nonnegative, and satisfy the following four conditions:

- *Affine invariance:* $D(A\mathbf{t} + \mathbf{b}; F_{A\mathbf{X}+\mathbf{b}}) = D(\mathbf{t}; F_{\mathbf{X}})$ holds for any random vector $\mathbf{X}$, any $d \times d$ nonsingular matrix A, and any vectors $\mathbf{t}$ and $\mathbf{b}$ in R^d;
- *Maximality at center:* $D(\theta; F) = sup_{t \in R^d} D(\mathbf{t}; F)$ holds for any $F \in \mathrm{F}$ with center θ_F (a point of symmetry);
- *Monotonicity relative to deepest point:* For any $F \in \mathrm{F}$ with center θ, $D(\mathbf{t}; F) \leq D(\theta + \alpha(\mathbf{t} - \theta); F)$ holds for $\alpha \in [0, 1]$; and
- *Vanishing at infinity:* $D(\mathbf{t}; F) \to 0$ as $\| \mathbf{t} \| \to \infty$, for each $F \in \mathrm{F}$.

A sample version of $D(\cdot; F)$ is denoted by $D(\cdot; F_m)$, where F_m is the sample or empirical distribution function. Several novel depth functions have appeared in the literature, including Mahalanobis (*MD*) (Mahalanobis 1936), half-space (*HSD*) or Tukey's depth (Tukey 1975), convex hull peeling (*CHPD*) (Mosler 2002), simplicial volume (*SVD*) or Oja's depth (Oja 1983), majority (*MJD*) (Singh 1991), simplicial (*SD*) (Liu 1990), spatial (*SPD*) (Vardi and Zhang 2000), projection (*PD*) (Stahel 1981; Donoho, 1982), zonoid (*ZD*) (Koshevoy and Mosler 1997; Mosler, 2002); spherical (*SPHD*) (Elmore et al. 2006); and triangle (Liu and Modarres 2008) depth functions. Some of the above data depth functions are described below.

Let F be a multivariate distribution function in $R^d, d \geq 1$ and $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a random sample from F. The Mahalanobis depth $MD(\mathbf{t}; F)$ is defined as

$$MD(\mathbf{t}; F) = \left[1 + (\mathbf{t} - \mu_F)\Sigma_F^{-1}(\mathbf{t} - \mu_F)\right]^{-1},$$

where μ_F and Σ_F are the mean vector and dispersion matrix of F, respectively. The sample version of $MD(\mathbf{t}; F)$ is obtained by replacing μ_F and Σ_F with their sample estimates. The half-space depth $HSD(\mathbf{t}; F)$ is defined as

$$\begin{aligned} HSD(\mathbf{t}; F) &= \inf_H \{P(H) : H \text{ is a closed half-space in } R^d \\ &\qquad \text{and } \mathbf{t} \in H\} \\ &= \inf_{\mathbf{u} \in R^d} P_F(\{\mathbf{X} : \mathbf{u}^T\mathbf{X} \geq \mathbf{u}^T\mathbf{t}\}) = \inf_{\|\mathbf{u}\|=1} \\ &\qquad P_F(\{\mathbf{X} : \mathbf{u}^T\mathbf{X} \geq \mathbf{u}^T\mathbf{t}\}). \end{aligned}$$

The sample version of $HSD(\mathbf{t}; F)$ is $HSD(\mathbf{t}; F_m) = \frac{1}{m} \min_{\|\mathbf{u}\|=1} \#\{i : \mathbf{u}^T\mathbf{x}_i \geq \mathbf{u}^T\mathbf{t}\}$.

The simplicial depth $SD(\mathbf{t}; F)$ is defined as $SD(\mathbf{t}; F) = P_F\{\mathbf{t} \in S[\mathbf{X}_1, \ldots, \mathbf{X}_{d+1}]\}$ where $S[\mathbf{X}_1, \ldots, \mathbf{X}_{d+1}]$ is a closed simplex formed by $(d + 1)$ random observations from F. The sample version of $SD(\mathbf{t}; F)$ is obtained by replacing F by F_m, or alternatively, by computing the fraction of the sample random simplices containing the point t. The projection depth $PD(\mathbf{t}; F)$ (Zuo 2003) is defined as $PD(\mathbf{t}; F) = (1 + O(\mathbf{t}; F))^{-1}$ where $O(\mathbf{t}; F)$ is a measure of outlyingness of a point $\mathbf{t}$ w.r.t. F. For example, the outlyingness of a point $\mathbf{t}$ can be defined as the maximum outlyingness of $\mathbf{t}$ with respect to the univariate median in any one dimensional projection, that is,

$$O(\mathbf{t}; F) = \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T\mathbf{t} - Med(\mathbf{u}^T\mathbf{X})|}{MAD((\mathbf{u}^T\mathbf{X}))}$$

where $Med(\cdot)$ denotes the univariate median, $MAD(\cdot)$ denotes the univariate median absolute deviation; i.e., $MAD(Y) = Med(|Y - Med(Y)|)$, and $\|\cdot\|$ is the Euclidean norm.

The spatial depth $SPD(\mathbf{t}; F)$ or L_1 depth is defined as

$$SPD(\mathbf{t}; F) = 1 - \left\| E_F\left(\frac{\mathbf{t} - \mathbf{X}}{\|\mathbf{t} - \mathbf{X}\|}\right)\right\|.$$

The spherical depth $SPHD(\mathbf{t}; F)$ is defined as the probability that $\mathbf{t}$ is contained in the unique, closed random hypersphere, $S(\mathbf{X}_1, \mathbf{X}_2)$, formed by two random points $\mathbf{X}_1$ and $\mathbf{X}_2$, which are i.i.d with c.d.f. F . That is, $SPHD(\mathbf{t}; F) = P(\mathbf{t} \in S(\mathbf{X}_1, \mathbf{X}_2))$. The sample version of the spherical depth at a point $\mathbf{t}$ is $SPHD(\mathbf{t}; F_m) = \frac{1}{\binom{m}{2}} \sum_{i<j}^{m} I[\mathbf{t} \in S(\mathbf{x}_i, \mathbf{x}_j)]$.

The empirical convex hull peeling depth $CHPD(\mathbf{t}; F_m)$ at $\mathbf{t}$ w.r.t. the sample $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is defined as $CHPD(\mathbf{t}, F_m) = \min\{k : \mathbf{t} \in E_k\}$, if such a k exists; and zero, otherwise, where E_k denotes the level k convex layer to which $\mathbf{t}$ belongs.

To obtain a convex layer at level k, one initially construct the smallest convex hull which enclose all samples points $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and $\mathbf{t}$. The points on the perimeter are designated the first convex layer and removed. The convex hull for the remaining points is now constructed; the points on the perimeter constitute the second layer. This process is repeated, and a sequence of nested convex layers is formed.

The empirical zonoid depth $ZD(\mathbf{t}; F_m)$ at $\mathbf{t}$ w.r.t. the sample $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is defined as $ZD(\mathbf{t}, F_m) = \sup\{\alpha : \mathbf{t} \in D_\alpha(\mathbf{x}_1, \ldots, \mathbf{x}_m)\}$ where $D_\alpha(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \left\{\sum_{i=1}^m \lambda_i \mathbf{x}_i : \sum_{i=1}^m \lambda_i = 1, \quad 0 \leq \lambda_i, \alpha\lambda_i \leq \frac{1}{n}, \quad \text{for all} \quad i\right\}$.

For a continuous distribution F on R^d, let $L(\mathbf{X}_1, \mathbf{X}_2)$ denote the hyper-lens defined by the intersection of two identical closed hyper-spheres centered at $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, with radius $\|\mathbf{X}_1 - \mathbf{X}_2\|$. The hyper-lens of any two i.i.d. random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ in R^d is defined as $L(\mathbf{X}_1, \mathbf{X}_2) = B(\mathbf{X}_1, \|\mathbf{X}_1 - \mathbf{X}_2\|) \cap B(\mathbf{X}_2, \|\mathbf{X}_1 - \mathbf{X}_2\|)$ where $B(c, r)$ is the closed ball of radius r centered at c. The triangle depth function $TD(\mathbf{t}; F)$ for a vector $\mathbf{t} \in R^d$, with

respect to a distribution F on R^d, is the probability that $\mathbf{t}$ is contained in the random hyper-lens $L(\mathbf{X}_1, \mathbf{X}_2)$. That is, $TD(\mathbf{t}; F) = P(\mathbf{t} \in L(\mathbf{X}_1, \mathbf{X}_2))$ where $\mathbf{X}_1$ and $\mathbf{X}_2$ are i.i.d with c.d.f. F. Let $\mathbf{X}_1, \ldots, \mathbf{X}_m$ be an i.i.d. random sample with c.d.f. F. The sample (empirical) version of $TD(\mathbf{t}; F)$ is the proportion of $L(\mathbf{X}_i, \mathbf{X}_j), 1 \le i < j \le m$, that contain $\mathbf{t}$. That is, $TD(\mathbf{t}; F_m) = \frac{1}{\binom{m}{2}} \sum_{i<j}^{m} I[\mathbf{t} \in L(\mathbf{x}_i, \mathbf{x}_j)]$.

About the Author

For biography *see* the entry ►Measures of Dependence.

Cross References

►Multivariate Outliers
►Multivariate Rank Procedures: Perspectives and Prospectives
►Statistical Quality Control: Recent Advances

References and Further Reading

Donoho DL (1982) Breakdown properties of multivariate location estimators. PhD qualifying paper, Department of Statistics, Harvard University
Elmore RT, Hettmansperger TP, Xuan F (2006) Spherical data depth and a multivariate median. In: Liu R, Serfling R, Souvaine D (eds) Proceedings of data depth: robust multivariate analysis, computational geometry and applications, pp 87–101
Hotelling H (1929) Stability in competition. Econ J 39:41–57
Koshevoy G, Mosler K (1997) Zonoid trimming for multivariate distributions. Ann Stat 25:1998–2017
Liu RY (1990) On a notion of data depth based on random simplices. Ann Stat 18:405–414
Liu Z, Modarres R (2008) Triangle data depth. Technical report. Department of Statistics, George Washington University
Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci India 12:49–55
Mosler K (2002) Multivariate dispersion, central regions and depth. Lecture notes in statistics. Springer, Berlin
Oja H (1983) Descriptive statistics for multivariate distributions. Stat Probab Lett 1:327–333
Singh K (1991) Majority depth. Technical report. Rutgers University
Stahel WA (1981) Robust estimation: infinitesimal optimality and covariance matrix estimators. PhD thesis, ETH, Zurich (in German)

Data Mining

Bruno Scarpa
University of Padua, Padua, Italy

In recent decades technological innovation has made the availability of large and sometimes huge amounts of information on a phenomenon of interest simple and cheap. This is due to two main reasons: on one side the development of automatic methods of data acquisition, and on the other side the progress of storage technology producing the fall of related costs. This new environment involves all areas of human endeavor.

- Every month a supermarket chain releases millions of receipts, one for each shopping trolley checking out. The content of each trolley summarizes the needs, the propensities and the economic behavior of the customer that selected it. The collection of all these shopping lists forms an important information base for the supermarket in order to decide the sales and purchases politics. Such an analysis becomes even more interesting when each shopping list is connected with the customers loyalty cards, allowing to follow the single client behavior by recording the purchases sequences; a similar problem arises in analyzing the usage of credit cards.
- Telecommunication companies generate every day data on millions of phone calls and of other services. Companies are interested in analyzing the customers behavior in order to measure chances of up sell and cross sell, and to identify potential *churners* (customers with high propensity to move to others operators).
- The growth of the World Wide Web in the last years, and its success and usage in research, business and daily life, makes it the largest publicly accessible data source in the world. Analysts often need to extract knowledge from this huge source of available data. Search engines need to identify the small part of the documents that are related to each single query; this operation is complicated by a number of elements: (a) the total size of the set of documents is amazing, (b) data are not collected in a structured form, such as a well organized data base, (c) inside each single document the elements related with the specific query are not in a pre-specified position either with respect to the entire document or among themselves.
- In scientific research huge amount of data are also available, for example in microbiology for the study of DNA structure. The analysis of sequences of portions of DNA produces very large tables, called "DNA microarray", where each column represents a sequence of some thousands of numerical values associated to DNA of an individual and each row represents different individuals. The goal is, for example, to connect the configuration of these sequences with the occurrence of some disease.
- The set of physical, chemical or other measurements finalized to examining the Earth's climate is becoming massive. Even the simple structured organization of

such data is problematic as much as the analysis to summarize all these pieces of information.

- Digital images of the sky will create massive data sets of observational data regarding many different features of hundreds of millions of sky objects. Astronomers are often interested in finding some algorithm that can perform as well as human experts in classifying stars and galaxies.

These are only few examples, but from all of these (and from many other that could be provided), it is clear that nowadays the analysis of huge amount of data is often required in order to solve real world problems. Specific statistical tools are required in order to extract important and useful knowledge from such an abundant amount of information. In fact, from one side the large number of cases are difficult to visualize effectively and, on the other side, as dimensionality increases, it becomes more difficult to describe data. Also, as we have seen in some of the examples above, data may not assume the simple structure of a database and sometimes they are even collected from streaming (e.g., for recording electricity usage).

The exploration and analysis of huge amount of data is often called "data mining" recalling the extraction (of relevant knowledge) from a mine of data:

> Data mining *is the activity of graphical and numerical analysis of large observational data sets or streaming data in order to find useful knowledge for the data owner.*

"Useful knowledge" in this context is quite generic, since very often it is not a priori specified what is the object of interest, which is in general acquired by "mining" the data. From such a definition, it is clear that data mining is an interdisciplinary exercise where statistics, database technology, machine learning, pattern recognition (see ►Pattern Recognition, Aspects of and ►Statistical Pattern Recognition Principles), artificial intelligence and visualization play a role (Hand et al. 2001). Therefore, data mining requires an understanding of both statistical and computational issues.

Looking at it as a statistician, (e.g., Azzalini and Scarpa 2010) it is worthwhile outlining some statistical specificities of data mining:

- The size of the data is huge. When many variables are available, the problem of the curse of dimensionality appears: the number of unit cells in a space as the number of variables increases linearly has an exponential growth rate, so that, in high dimensional spaces, "nearest" points may be very far away. Also, when the number of cases is of the order of millions, computational issues need to be considered. "Every time the amount of data increases by a factor of ten, we should totally rethink how we analyze it" (Friedman 1997).
- The observational context: data are often not collected from an experiment, but purely observational. They just "exists." Often they have been collected by convenience or opportunity, rather than randomly.
- Data used for the analysis have often been collected for some other purpose (e.g., they may have been collected in order to maintain an up-to-date record of all the transactions in a bank).
- Data may be dirty, corrupted, contaminated, missing and need to be preprocessed in order to be used.
- Sampling and population issues: the huge amount of data can be tackled by sampling, if the aim is modeling, but not necessarily so if the aim is pattern detection. When data sets are entire populations (e.g., the entire customer base of a company), the standard statistical notion of inference has no relevance.
- "If you torture data long enough, Nature will always confess" (Coase 1982), with so large data sets it would be easy to find a model that will fit the data well, but often this model will not really describe the phenomenon of interest. In this context the trade off between bias and variance needs to be managed.

In summary, while data mining does overlap considerably with the standard exploratory data analysis techniques of statistics, it also runs into new problems, many of which are consequences of size and the non-traditional nature of the data sets involved.

Using a machine learning classification, data mining tasks may be divided into two groups, the *supervised learning* problems, "supervised" because of the presence of an outcome variable to guide the learning process, and the *unsupervised learning* problems, (also said in a more statistical language *internal analysis* problems) where we observe only the features and have no measurements of the outcome.

Often the aim of supervised learning is to build a model that will permit the prediction of the value of one variable from known values of other variables. If the variables being predicted (called dependent variables or outputs or responses) are categorical, the problem is termed *classification*, while if they are quantitative the problem is called *regression*. A large number of methods have been developed in statistics and machine learning to tackle these problems, from simple parametric models such as ►generalized linear models (McCullagh and Nelder 1989) to nonparametric structures such as generalized additive models (Hastie and Tibshirani 1990), ►neural networks (e.g., Ripley 1996), support vector machines (Vapnik 1996), classification and regression trees (Breiman et al. 1984)

and their resampling versions such as bagging (Breiman 1996), boosting (Freund and Schapire 1996) or random forests (Breiman 2001a) and many others that have been developed in recent years.

Unsupervised learning is a class of problems in which one seeks to determine how the data are organized, directly inferring the properties of the density of a multivariate variable without the help of a "teacher" providing correct answers. Classical statistics and machine learning literature (e.g., Kaufman and Rousseeuw 1990; Ripley 1996) provide many clustering algorithms with the aim of grouping or segmenting a collection of objects into subsets or "clusters," such that those within each cluster are more closely related to one another than objects assigned to different clusters. Another family of algorithmic techniques based on *association rules* (e.g., Hastie, Tibshirani and Friedman 2008) is related to the task of finding combinations of items that occur frequently in transaction databases. Other goals of interest, in this context, are on detecting patterns among data, like spotting fraudulent behavior among customers of a company, or detecting unusual stars or galaxies in astronomical work in order to identify previously unknown phenomena. A particular case of this pattern detection is when a pattern of interest is already known by the user and he is interested in finding similar patterns in the data set. This task is very common in text and image data sets, where the pattern can be a set of keywords, and the user may wish to find relevant documents within a large set of possibly important documents, or the user may have a sample image, and wish to find similar images from a large set of images.

Most of the methods included in data mining are often considered as automatic tools that do not require any human intervention. In particular many algorithms generated in the machine learning context, dealing with really huge amount of data with, sometimes, very complicated computational structure, and seems to present an objective advantage by working without any human intervention. However, from empirical experience, it seems that an analysis driven by a thinking brain is much more powerful. In fact the necessity to "understand the problem" characterizes the style of statistical data mining, in all the phases of the analysis from the data preparation to the interpretation of the results. Even when black box algorithms are used, one can never hope to solve every problem by using software on a powerful computer just by "pushing a button". A deep knowledge of the nature of the tools used and of how the used methods work is crucial for, at least, three broad reasons (Azzalini and Scarpa 2010):

1. Understanding the characteristics of the tools is crucial in choosing the right method.
2. The same mastery is required in order to correctly interpret the results of the algorithms.
3. Some ability in the computational and algorithmic aspects is very useful in order to better evaluate the computer output, considering also its reliability.

The paper by Breiman (2001b), with the discussion that it generated, describes quite well the specificities of the type of statisticians needed to tackle these problems.

About the Author

Dr. Bruno Scarpa joined the Department of Statistical Sciences, University of Padova, in 2007, after four years being at the Department of Applied Statistics at the University of Pavia. Previously, he used to work as a manager for some telecommunication companies where he was in charged of activities of customer profiling, churn analysis, CRM and data mining services for marketing. He is author of several scientific papers and jointly with Adelchi Azzalini is author of the book *Data Mining and Data Analysis* (published in 2004 by Springer in Italian and soon will be published in English).

Cross References

►Business Forecasting Methods
►Data Mining Time Series Data
►Preprocessing in Data Mining
►Statistics: An Overview

References and Further Reading

Azzalini A, Scarpa B (2010) Data analysis and data mining. Oxford University Press, New York

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International, Monterey, CA

Breiman L (1996) Bagging predictors. Mach Learn 26:123–140

Breiman L (2001a) Random forests. Mach Learn 45:5–32

Breiman L (2001b) Statistical modeling: the two cultures (with discussion). Stat Sci 16:199–231

Coase RH (1982) How should economists choose? American Enterprise Institute for Public Policy Research, Washington, DC

Friedman JH (1997) Data mining and statistics: what's the connection? In: Proceedings of the 29th symposium on the interface: computing science and statistics. Houston, TX, May 1997. Available at http://www-stat.stanford.edu/~jhf/ftp/dm-stat.pdf

Freund Y, Schapire R (1996) Experiments with a new boosting algorithm. In: Machine learning: proceedings of the thirteenth international conference, Morgan Kauffman, San Francisco, pp 148–156

Hand D, Mannila H, Smyth P (2001) Principles of data mining. MIT Press, Cambridge

Hastie T, Tibshirani R (1990) Generalized additive models. Chapman and Hall, London

Hastie T, Tibshirani R, Friedman JH (2008) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York

McCullagh P, Nelder JA (1989) Generalized linear models. Chapman and Hall, London
Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
Vapnik V (1996) The nature of statistical learning theory. Springer, New York

Data Mining Time Series Data

Eamonn J. Keogh
Professor, University Scholar
University of California-Riverside, Riverside, CA, USA

Time series data is ubiquitous; large volumes of time series data are routinely created in medical and biological domains, examples include gene expression data (Aach and Church 2001), electrocardiograms, electroencephalograms, gait analysis, growth development charts etc. Although statisticians have worked with time series for more than a century, many of their techniques hold little utility for researchers working with massive time series databases (for reasons discussed below).

Below are the major task considered by the time series data mining community:

- **Indexing** (Query by Content): Given a query time series Q, and some similarity/dissimilarity measure D(Q,C), find the most similar time series in database *DB* (Chakrabarti et al. 2002; Faloutsos et al. 1994; Kahveci and Singh 2001; Popivanov and Miller 2002).
- **Clustering**: Find natural groupings of the time series in database *DB* under some similarity/dissimilarity measure D(Q,C) (Aach and Church 2001; Debregeas and Hebrail 1998; Kalpakis et al. 2001; Keogh and Pazzani 1998).
- **Classification**: Given an unlabeled time series Q, assign it to one of two or more predefined classes (Geurts 2001; Keogh and Pazzani 1998).
- **Prediction** (Forecasting): Given a time series Q containing n datapoints, predict the value at time $n+1$.
- **Association Detection**: Given two or more time series, find relationships between them. Such relationships may or may not be casual and may or may not exist for the entire duration of the time series.
- **Summarization**: Given a time series Q containing n datapoints where n is an extremely large number, create a (possibly graphic) approximation of Q which retains its essential features but fits on a single page, computer screen etc. (Indyk et al. 2000; van Wijk and van Selow 1999).
- **Anomaly Detection** (Interestingness Detection): Given a time series Q, assumed to be normal, and an unannotated time series R. Find all sections of R which contain anomalies or "surprising/interesting/unexpected" occurrences (Guralnik and Srivastava 1999; Keogh et al. 2002; Shahabi et al. 2000).
- **Segmentation**: Given a time series Q containing n datapoints, construct a model $\overline{Q}$, from K piecewise segments ($K << n$) such that $\overline{Q}$ closely approximates Q (Keogh and Pazzani 1998).

Note that indexing and clustering make *explicit* use of a distance measure, and many approaches to classification, prediction, association detection, summarization and anomaly detection make *implicit* use of a distance measure. For this reason, there has been significant research on finding the "best" distance measure for time series (Keogh and Kasetty 2002). Recent work, however, suggests that the simple Euclidean distance is surprisingly competitive (Ding et al. 2008).

It is interesting to note that with the exception of indexing, research into the tasks enumerated above predate not only the decade old interest in ▶data mining, but computing itself. What then, are the essential differences between the classic, and the data mining versions of these problems? The key difference is simply one of size and scalability; time series data miners routinely encounter datasets that are gigabytes in size. As a simple motivating example, consider hierarchal clustering. The technique has a long history, and well-documented utility. If however, we wish to hierarchically cluster a mere million items, we would need to construct a matrix with 10^{12} cells, well beyond the abilities of the average computer for many years to come. A data mining approach to clustering time series, in contrast, must explicitly consider the scalability of the algorithm (Kalpakis et al. 2001).

In addition to the large volume of data, it is often the case that each individual time series has a very high dimensionality (Chakrabarti et al. 2002). Whereas classic algorithms assume a relatively low dimensionality (e.g., a few measurements such as "height, weight, blood sugar etc."), time series data mining algorithms must be able to deal with dimensionalities in the hundreds and thousands. The problems created by high dimensional data are more than mere computation time considerations, the very meanings of normally intuitive terms such as "similar to" and "cluster forming" become unclear in high dimensional space. The reason is that as dimensionality increases, all objects

become essentially equidistant to each other, and thus classification and clustering lose their meaning. This surprising result is known as the "curse of dimensionality" and has been the subject of extensive research (Aggarwal et al. 2001). The key insight that allows meaningful time series data mining is that although the actual dimensionality may be high, the *intrinsic* dimensionality is typically much lower. For this reason, virtually all time series data mining algorithms avoid operating on the original "raw" data, instead they consider some higher-level representation or abstraction of the data.

Time Series Representations

As noted above, time series datasets are typically very large, for example, just 8 h of electroencephalogram data can require in excess of a gigabyte of storage. This is a problem because for almost all data mining tasks, most of the execution time spent by algorithm is used simply to move data from disk into main memory. This is acknowledged as the major bottleneck in data mining, because many naïve algorithms require multiple accesses of the data. As a simple example, imagine we are attempting to do k-means clustering of a dataset that does not fit into main memory. In this case, every iteration of the algorithm will require that data in main memory to be swapped. This will result in an algorithm that is thousands of times slower than the main memory case.

With this in mind, a generic framework for time series data mining has emerged. The basic idea is summarized in (Table 1).

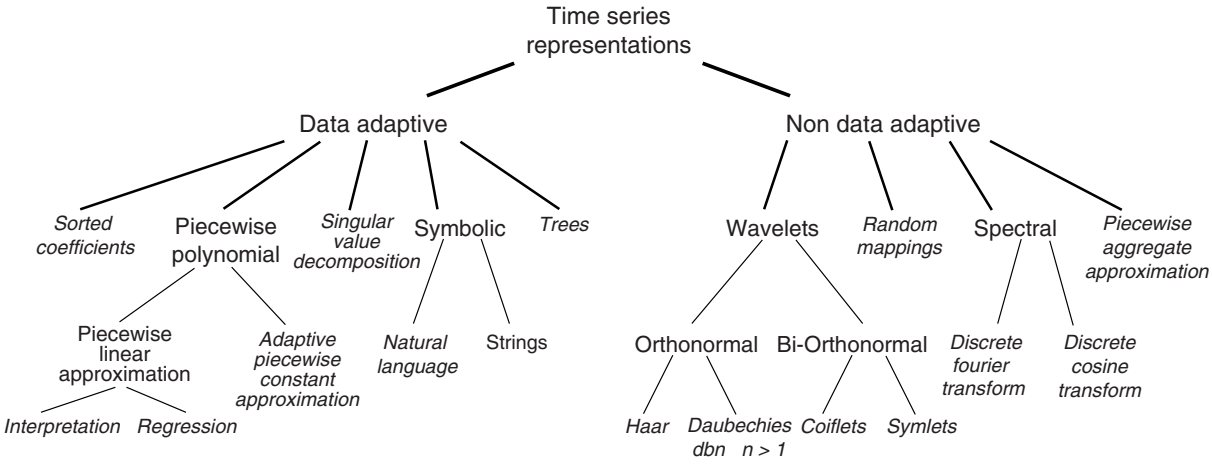
It should be clear that the utility of this framework depends heavily on the quality of the approximation created in Step 1. If the approximation is very faithful to the original data, then the solution obtained in main memory is likely to be the same, or very close to, the solution we would have obtained on the original data. The handful of disk accesses made in Step 2 to confirm or slightly modify the solution will be inconsequential compared to the number of disks accesses required if we had worked on the original data. With this in mind, there has been a huge interest in approximate representation of time series. Figure 1 illustrates a hierarchy of every major representation proposed in the literature.

Data Mining Time Series Data. Table 1 A generic time series data mining approach

1. Create an approximation of the data, which will fit in main memory, yet retains the essential features of interest
2. Approximately solve the problem at hand in main memory
3. Make (hopefully very few) accesses to the original data on disk to confirm the solution obtained in Step 2, or to modify the solution so it agrees with the solution we would have obtained on the original data

Given the plethora of different representations, it is natural to ask which is best. Recall that the more faithful the approximation, the less clarification disks accesses we will need to make in Step 3 of Table 1. There have been many attempts to answer the question of which is the best representation, with proponents advocating their favorite technique (Chakrabarti et al. 2002; Faloutsos et al. 1994; Popivanov and Miller 2002; Rafiei and Mendelzon 1998). The literature abounds with mutually contradictory statements such as "*Several wavelets outperform the . . . DFT*" (Popivanov and Miller 2002), "*DFT-based and DWT-based techniques yield comparable results*" (Wu et al. 2000), and "*Haar wavelets perform. . . better that DFT*" (Kahveci and Singh 2001). However an extensive empirical comparison on 50 diverse datasets suggests that while some datasets favor a particular approach, overall there is little difference between the various approaches in terms of their ability to approximate the data (Ding et al. 2008; Keogh and Kasetty 2002). There are however, other important differences in the usability of each approach (Chakrabarti et al. 2002). We will consider some representative examples of strengths and weaknesses below.

The wavelet transform is often touted as an ideal representation for time series data mining, because the first few wavelet coefficients contain information about the overall shape of the sequence while the higher order coefficients contain information about localized trends (Popivanov and Miller 2002; Shahabi et al. 2000). This multiresolution property can be exploited by some algorithms, and contrasts with the Fourier representation in which every coefficient represents a contribution to the global trend (Faloutsos et al. 1994; Rafiei and Mendelzon 1998). However wavelets do have several drawbacks as a data mining representation. They are only defined for data whose length is an integer power of two. In contrast, the Piecewise Constant Approximation suggested by (Yi and Faloutsos 2000), has exactly the fidelity of resolution of as the Haar wavelet, but is defined for arbitrary length time series. In addition, it has several other useful properties such as the ability to support several different distance measures (Yi and Faloutsos 2000), and the ability to be calculated in an incremental fashion as the data arrives (Chakrabarti et al. 2002). Choosing the right representation for the task



Data Mining Time Series Data. Fig. 1 A hierarchy of time series representations

at hand is the key step in any time series data-mining endeavor. The points above only serve as a sample of the issues that must be addressed.

Readings

The field of time series data mining is relatively new, and ever changing. Because of the length of journal publication delays, the most interesting and useful work tends to appear in top-tier conference proceedings. Interested readers are urged to consult the latest proceedings of the major conferences in the field. These include the ACM Knowledge Discovery in Data and Data Mining, IEEE International Conference on Data Mining and the IEEE International Conference on Data Engineering.

About the Author

Eamonn J. Keogh received his Ph.D. in Computer Science in 2001, UC Irvine. His research areas include data mining, machine learning and information retrieval, specializing in techniques for solving similarity and indexing problems in time-series datasets. Dr. Keogh is a prolific author in data mining/database conferences. He has (co-)authored more than 120 papers. His papers on time series data mining have been referenced well over 5,000 times (his H-index is 34). He received the IEEE ICDM 2007 best paper award, SIGMOD 2001 best paper award, and runner up best paper award in KDD 1997. He has made some important contributions in data mining time series data, including: Symbolic Aggregate approXimation (SEX) (with Jessica Lin, 21002), the first symbolic representation for time series that allows for dimensionality reduction and indexing with a lower-bounding distance measure, and LB_Keogh, a tool for lower bounding various time series distance measure, that makes retrieval of time-warped time series feasible even for large data sets.

Cross References

- ►Business Forecasting Methods
- ►Data Mining
- ►Distance Measures
- ►Hierarchical Clustering
- ►Time Series

References and Further Reading

Aach J, Church G (2001) Aligning gene expression time series with time warping algorithms. Bioinformatics 17: 495–508

Aggarwal C, Hinneburg A, Keim DA (2001) On the surprising behavior of distance metrics in high dimensional space. In: Proceedings of the 8th international conference on database theory, London, UK, 4–6 Jan 2001, pp 420–434

Chakrabarti K, Keogh E, Pazzani M, Mehrotra S (2002) Locally adaptive dimensionality reduction for indexing large time series databases. ACM T Database Syst 27(2):188–228

Ding H, Trajcevski G, Scheuermann P, Wang X, Keogh E (2008) Querying and mining of time series data: experimental comparison of representations and distance measures VLDB

Debregeas A, Hebrail G (1998) Interactive interpretation of kohonen maps applied to curves. In: Proceedings of the 4th Int'l conference of knowledge discovery and data mining, New York, NY, 27–31 Aug 1998, pp 179–183

Faloutsos C, Ranganathan M, Manolopoulos Y (1994) Fast subsequence matching in time-series databases. In: Proceedings of the ACM SIGMOD Int'l conference on management of data, Minneapolis, MN, 25–27 May 1994, pp 419–429

Geurts P (2001) Pattern extraction for time series classification. In: Proceedings of principles of data mining and knowledge discovery, 5th European conference, Freiburg, Germany, 3–5 Sept 2001, pp 115–127

Guralnik V, Srivastava J (1999) Event detection from time series data. In: Proceedings of the 5th ACM SIGKDD Int'l conference on knowledge discovery and data mining, San Diego, CA, 15–18 Aug 1999, pp 33–42

Indyk P, Koudas N, Muthukrishnan S (2000) Identifying representative trends in massive time series data sets using sketches. In: Proceedings of the 26th Int'l conference on very large data bases, Cairo, Egypt, 10–14 Sept 2000, pp 363–372

Kahveci T, Singh A (2001) Variable length queries for time series data. In: Proceedings of the 17th Int'l conference on data engineering, Heidelberg, Germany, 2–6 Apr 2001, pp 273–282

Kalpakis K, Gada D, Puttagunta V (2001) Distance measures for effective clustering of ARIMA time-series. In: Proceedings of the IEEE Int'l conference on data mining, San Jose, CA, Nov 29–Dec 2 2001, pp 273–280

Keogh E, Lonardi S, Chiu W (2002) Finding surprising patterns in a time series database in linear time and space. In: The 8th ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton, Alberta, Canada, 23–26 July 2002, pp 550–556

Keogh E, Pazzani M (1998) An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: Proceedings of the 4th Int'l conference on knowledge discovery and data mining, New York, NY, 27–31 Aug 1998, pp 239–241

Keogh E, Kasetty S (2002) On the need for time series data mining benchmarks: a survey and empirical demonstration. In: The 8th ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton, Alberta, Canada, 23–26 July 2002, pp 102–111

Popivanov I, Miller RJ (2002) Similarity search over time series data using wavelets. In: Proceedings of the 18th Int'l conference on data engineering, San Jose, CA, Feb 26–Mar 1 2002, pp 212–221

Rafiei D, Mendelzon AO (1998) Efficient retrieval of similar time sequences using DFT. In: Proceedings of the 5th Int'l conference on foundations of data organization and algorithms, Kobe, Japan, 12–13 Nov 1998

Shahabi C, Tian X, Zhao W (2000) TSA-tree: a wavelet based approach to improve the efficiency of multi-level surprise and trend queries. In: Proceedings of the 12th Int'l conference on scientific and statistical database management, Berlin, Germany, 26–28 Jul 2000, pp 55–68

van Wijk JJ, van Selow E (1999) Cluster and calendar-based visualization of time series data. In: Proceedings 1999 IEEE symposium on information visualization, IEEE Computer Society, 25–26 October 1999, pp 4–9

Wu Y, Agrawal D, El Abbadi A (2000) A comparison of DFT and DWT based similarity search in time-series databases. In: Proceedings of the 9th ACM CIKM Int'l conference on information and knowledge management, McLean, VA, 6–11 Nov 2000, pp 488–495

Yi B, Faloutsos C (2000) Fast time sequence indexing for arbitrary lp norms. In: Proceedings of the 26th Int'l conference on very large databases, Cairo, Egypt, 10–14 Sept 2000, pp 385–394

Data Privacy and Confidentiality

Stephen E. Fienberg[1], Aleksandra B. Slavković[2]
[1]Maurice Falk University Professor
Carnegie Mellon University, Pittsburgh, PA, USA
[2]Associate Professor
The Pennsylvania State University, University Park, PA, USA

Introduction

Data privacy is an overarching concern in modern society, as government and non-government agencies alike collect, archive, and release increasing amounts of potentially sensitive personal data. Data owners or stewards, in the case of statistical agencies, often critically evaluate both the type of data that they make publicly available and the format of the data product releases. The statistical challenge is to discover how to release important characteristics of existing databases without compromising the privacy of those whose data they contain.

Modern databases, however, pose new privacy problems due to the types of information they hold and their size. In addition to traditional types of information contained in censuses (see ▶Census), surveys, and medical and public health studies, contemporary information repositories store social network data (e.g., cell phone and Facebook data), product preferences (e.g., from commercial vendors), web search data, and other statistical information that was hitherto unavailable in digital format. The information in modern databases is also more commercially exploitable than pure census data (e.g., credit cards, purchase histories, medical history, mobile device locations). As the amount of data in the public realm accumulates and record-linkage methodologies improve, the threat to confidentiality and privacy magnifies. Repeated database breaches demonstrate that removing obviously identifying attributes such as names is insufficient to protect privacy (e.g., Narayanan and Shmatikov 2006; Backstrom et al. 2007). Even supposed anonymization techniques can leak sensitive information when the intruder has modest partial knowledge about the data from external sources (e.g., Coull et al. 2008; Ganta et al. 2008).

Two rich traditions of investigating data confidentiality have developed within the scientific community, one in statistics, e.g., see Federal Committee on Statistical Methodology (2005), and the other more recently in computer science. Both tackle the fundamental trade-off between utility and privacy, but in essentially different ways. In statistics the focus has been on the trade-off

between disclosure risk and data utility while in computer science the focus has been on algorithmic aspects of the problem and more recently on rigorous definitions of privacy that may also allow for utility. A special 2009 issue of the *Journal of Privacy and Confidentiality* features survey articles describing the current approaches and open research problems (http://repository.emu.edu/jpe/).

Statistical Disclosure Limitation

Statistical disclosure limitation (SDL) applies and develops statistical tools for limiting releases of sensitive information from statistical databases while allowing for valid statistical inference. Emanating from official statistics involving censuses and large-scale national surveys, modern SDL emphasizes statistical inference as the main yardstick of utility; e.g., given a statistical model, how well can one "anonymize" the data and still carry out valid model estimation and assessment? The goal of disclosure limitation is to examine and manage a trade-off between data utility and disclosure risk (e.g., Doyle et al. 2001; Trottini and Fienberg 2002; Duncan and Stokes 2009). Data utility is a measure of usefulness of a dataset for an intended analyst. Disclosure risk measures the degree to which a dataset and its released statistics reveal sensitive information. This notion is probabilistic; as released data accumulate in the public domain, the probability of uniquely identifying members of the population increases.

Starting with Dalenius (1977) and continuing to modern work with connections to algebraic geometry, multivariate analysis, optimization, and the generation of synthetic data, SDL has a vast literature with primary applications to social sciences and health data. In general, SDL methods introduce bias and variance to data in order to minimize identity and attribute disclosure while trying to retain sufficient information needed for proper statistical inference. Data masking involves transforming an $n \times p$ (cases by variables) original data matrix Z through pre- and post-multiplication and the possible addition of noise. That is, $Z = AZB + C$, where A is a matrix that operates on the n cases, B is a matrix that operates on the p variables, and C is a matrix that adds noise. Matrix masking can be applied to either microdata or table of counts, and includes a variety of standard approaches to disclosure limitation (*see* ►Statistical Approaches to protecting confidentiality in public use data): (a) recodings such as (e.g., rounding and thresholding), (b) cell suppression, (c) data swapping, and (d) perturbation. More recent methods include *sampling* which involves releasing subsets of data or variables – deleting rows that is columns of Z, and *simulations* such as adding rows to Z (e.g., see Reiter 2005 for generation of synthetic Microdata, Slavkovic and Lee 2010 for creation of synthetic tables, and Dobra et al. 2008 for partial information release, including optimization bounds on contingency tables).

Computer Science Approaches

Closely related to SDL techniques are privacy-preserving data mining (PPDM) methods that aim to construct efficient data mining algorithms (see ►Data Mining) while maintaining privacy with the emphasis on automation and scalability of the anonymization process. Currently proposed methods can be roughly grouped in (a) randomization methods (e.g., multiplicative perturbations, data swapping), (b) group based anonymization, (c) distributed PPDM, and (d) privacy-preservation of application rules (e.g., see Fienberg and Slavković 2005 for association rule hiding). For an overview of the PPDM models and algorithms, and related issues see Aggarwal and Yu (2008).

Research emerging from cryptography has emphasized that privacy and security are properties of the anonymization algorithm, rather than of a particular output. This perspective allows for rigorous definitions and proofs of privacy. In particular, a concept called *differential privacy* (e.g., Dwork et al. 2006) provides rigorous guarantees no matter what external information is available to an intruder. The algorithmic perspective also leads to utility being defined in terms of functional approximation: given a database x and a function f, how well does the released information allow one to approximate $f(x)$.

Some recent developments aim to establish connections between differential privacy and traditional statistical inference; e.g., see Smith (2008) for private parametric estimation and Wasserman and Zhou (2008) for approximation of smooth densities. In some cases, the theory has sufficiently advanced for the development of concrete methodological guidelines; e.g., related to the additive perturbation techniques of Dwork et al. (2006), Vu and Slavković (2009) have begun to develop rules for sample size calculations and statistical ►power analysis. Machanavajjhala et al. (2008) describe the first application of privacy tools to large-scale public-use databases by creating differentially private synthetic data based on Bayesian posterior predictive distributions.

Another more recent promising approach uses cryptographic protocols for distributing privacy-preserving algorithms for valid statistical analysis among a group of servers (or data owners) so as to avoid pooling data in any single location. Using ideas from secure multiparty computation (Lindell and Pinkas 2009), statisticians have been working on variations of secure protocols for generalized linear models analysis (e.g., Ghosh et al. 2007).

Future Directions

Protecting privacy in statistical databases is an increasingly challenging problem. Any useful statistics extracted from a database must reveal some information about the individuals whose data are included. Despite an increasing focus on large and often sparse data sets, the evaluation of disclosure risk and utility of statistical results often involves human-guided tuning of parameters and context-specific notions of privacy. New policies and data privacy research calls for methodology that achieves maximum utility, minimal risk and transparency. Current trends aim at integration of the computationally focused, rigorous definitions of privacy and cryptographic protocols emanating from computer science with notions of utility from statistics.

Acknowledgments

Supported in part by National Science Foundation grant DMS-0631589 and U.S. Army Research Office Contract W911NF-09-1-0360 to Carnegie Mellon University.

Supported in part by National Science Foundation grant SES-0532407 to the Department of Statistics, Pennsylvania State University.

About the Authors

Stephen E. Fienberg is Maurice Falk University Professor of Statistics and Social Science at Carnegie Mellon University, with appointments in the Department of Statistics, the Machine Learning Department, Cylab, and i-Lab. Professor Fienberg is Past President of the Institute of Mathematical Statistics (1998–1999), and Past President of the International Society for Bayesian Analysis (1996–1997) and he has been Vice President of the American Statistical Association. He is an Elected member of the U.S. National Academy of Sciences (1999) and co-chair of its Report Review Committee (2008–2012). He is a fellow of the American Academy of Arts and Sciences (2007), the American Academy of Political and Social Science (2004), and the Royal Society of Canada (2004). Professor Fienberg is an Editor of the *Annals of Applied Statistics* (2006–), and former editor of the *Journal of the American Statistical Association*. He was Co-Founder of the *Journal of Privacy and Confidentiality* and *Chance*. He has received the Wilks Award (2000) and the Founders Award (2009) from the American Statistical Association, and the Lise Manchester Award from the Statistical Society of Canada (2008).

Aleksandra Slavkovic received her Ph.D. from Carnegie Mellon University in 2004. She is currently an Associate Professor of Statistics, with appointments at the Department of Statistics and the Institute for CyberScience at Penn State University, University Park, and at the Department of Public Health Sciences, Pennsylvania State College of Medicine, Hershey. She is currently serving as an Associate Editor of the Annals of Applied Statistics, Journal of Privacy and Confidentiality and Journal of Statistical Computation and Simulation. Dr. Slavkovic has served as a consultant to the National Academy of Sciences/National Research Council Committee to Review the Scientific Evidence of Polygraph in 2001 and part of 2002. In 2003, she received and honorable mention for the best student paper from the Committee on Statisticians in Defense and National Security of the American Statistical Association.

Cross References

►Census
►Federal Statistics in the United States, Some Challenges
►Multi-Party Inference and Uncongeniality
►Statistical Approaches to Protecting Confidentiality in Public Use Data

References and Further Reading

Aggarwal C, Yu P (2008) Privacy-preserving data mining: models and algorithms. Springer, New York

Backstrom L, Dwork C, Kleinberg J (2007) Wherefore art thou r3579x? Anonymized social networks, hidden patterns, and structural steganography. In: Proceedings of 16th international World Wide Web conference, Banff, Alberta, Canada

Coull S, Wright C, Monrose F, Keromytis A, Reiter M (2008) Taming the devil: techniques for evaluating anonymized network data. In: Fifteenth annual Network & Distributed System Security Symposium (NDSS), San Diego, CA, USA

Dalenius T (1977) Towards a methodology for statistical disclosure control. Stat Tidskr 5:35–64

Dobra A, Fienberg S, Rinaldo A, Slavković A, Zhou Y (2008) Algebraic statistics and contingency table problems: log-linear models, likelihood estimation, and disclosure limitation. In: Putinar M, Sullivant S (eds) Emerging applications of algebraic geometry. IMA series in applied mathematics. Springer, New York, pp 63–88

Doyle P, Lane J, Theeuwes J, Zayatz L (2001) Confidentiality, disclosure and data access. Theory and applications for statistical agencies. Elsevier, New York

Duncan G, Stokes L (2009) Data masking for disclosure limitation. Wiley Interdiscip Rev: Comput Stat 1(1):83–92

Dwork C, McSherry F, Nissim K, Smith A, March (2006) Calibrating noise to sensitivity in private data analysis. Proceedings of The theory of cryptography conference (TCC 2006). Springer-Verlag LNCS 3876(1):265–284. New York, NY

Federal Committee on Statistical Methodology (2005) Report on statistical disclosure limitation methodology. Statistical policy working paper 22. http://ntl.bts.gov/docs/wp22.html

Fienberg S, Slavković A (2005) Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. Data Mining and Knowledge Discovery, 11(2):155–180

Ganta SR, Kasiviswanathan SP, Smith A (2008) Composition attacks and auxiliary information in data privacy. In: Proceedings of SIG-KDD, pp 265–273

Ghosh J, Reiter J, Karr A (2007) Secure computation with horizontally partitioned data using adaptive regression splines. Comput Stat & Data Anal, 51(12):5813–5820

Lindell Y, Pinkas B (2009) Secure multiparty computation for privacy-preserving data mining. J Priv and Confidentiality, 1:59–98

Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L (2008) Privacy: Theory meets practice on the map. In: IEEE 24th international conference on data engineering, pp 277–286

Narayanan A, Shmatikov V (2006) How to break anonymity of the netflix prize dataset. ArXiV report cs.CR/0610105, http://arxiv.org

Reiter J (2005) Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. J R Stat Soc Ser A 168(1):185–205

Slavkovic AB, Lee J (2010) Synthetic two-way contingency tables that preserve conditional frequencies. Stat Meth, 7(3):225–239

Smith A (2008) Efficient, differentially private point estimators. Preprint arXiv:0809.4794v1

Trottini M, Fienberg S (2002) Modelling user uncertainty for disclosure risk and data utility. Int J of Uncert, Fuzzi and Know-Based Sys 10(5):511–527

Vu D, Slavković A (2009) Differential privacy for clinical trial data: preliminary evaluations. In: ICDMW '09: Proceedings of the 2009 IEEE international conference on data mining workshops, Miami, FL, USA. IEEE Computer Society, Washington, DC, pp 138–143

Wasserman L, Zhou S (2008) A statistical framework for differential privacy. URL http://www.citebase.org/abstract?id=oai:arXiv.org:0811.2501

Data Quality (Poor Quality Data: The Fly in the Data Analytics Ointment)

Frank M. Guess[1], Thomas C. Redman[2]
[1]Professor
University of Tennessee, Knoxville, TN, USA
[2]President of Navesink Consulting Group
Navesink Consulting Group, Rumson, NJ, USA

Introduction and Summary

In recent years a powerful combination of database technologies, data mining techniques (see ▶Data Mining) and analytics software have created vast new opportunities for data analysts and statisticians. For example, corporations have duly stored the results of their customer transactions in corporate databases for over a generation. There are, quite literally millions of records. Massively parallel engines can examine these data in heretofore unimagined ways. The potentials to understand customer profitability, develop better understandings of customers' past needs and predict future ones, and to use those insights to develop new product niches are enormous.

Yet all is not well in the world of data analytics. Unlocking the mysteries data have to offer is difficult at best. And putting the discoveries to work can be even harder. One major reason is poor quality data. Bad data camouflage the hidden nuggets in data or, worse, send an analysis in the wrong direction altogether. Some years ago George Box observed that "...all models are wrong, but some are useful". Compare Box (1976) and see Box and Draper (1987). Many decision makers seem to intuitively grasp the "all models are wrong" portion and so are skeptical of data mining. Poor data quality exacerbates the problem. Indeed most decision makers are well aware that poor data cause problems in operations. How can things be any better when bad data are combined with incorrect models?!

What is a poor data analyst to do? Our advice: "Think and act like an experimenter!" Since time immemorial, experimenters have understood the importance of high-quality data – data that are relevant to the subject area, are clearly defined, have minimal or no bias and high precision. And they invest considerable time and energy to achieve this goal. The following summarizes some of these investments and how they apply to analytics.

Experimenters plan and design their experiments carefully: Perhaps because their data are so dear, experimenters take time to understand what is already known and to carefully define their objectives, the populations and sub-populations of interest, and the hypotheses of primary interest. In contrast, it is too easy to turn an analytics tool loose, in hope that the computer can make interesting discoveries on its own. Much is lost, especially the abilities to interpret results and identify spurious correlations. So data analysts are well-advised to narrow their foci and develop working knowledge of their subject areas. Then they should aim their analytic tools rather more like a high-powered rifle than a shotgun.

Some may object that doing so limits opportunities for serendipity. In our experience just the opposite is true. Serendipitous discoveries are more likely when the current level of understanding does not explain the data and piques the experimenter's curiosity. That cannot happen without adequate background.

Similarly, experimenters *design* their experiments, seeking exactly the right data. Data analysts should be just as persistent in creating or searching for the data that best meets their needs, not the data that are easy to find. They should confirm interesting results based on large-scale analyzes with small, designed studies.

Experimenters get very close to their data: First, they develop deep understanding of their data collection

devices. Analysts can do the same thing by working backwards, identifying and understanding the business processes that created their data. Business people are often more casual than experimenters about data definition, so it is critical that data analysts understand the nuance in their data.

Second, experimenters build controls into their data collection processes. The simplest example is calibrating their equipment. Data analysts do not own these processes, but they must understand existing controls and recommend others where needed to prevent poor data, whenever possible.

Third, experimenters search for and eliminate ►outliers. When one of us (Redman) worked at Bell Labs in the 80s and 90s, we used the expression "rinse, wash, scrub" for increasing efforts to identify and eliminate suspect data. It was intense, manual effort, certainly not feasible for enormous databases. But data analysts can certainly rinse, wash, and scrub a small sample. Doing so provides a basis for evaluating the quality of the entire database. And if the results of a large analysis are confirmed on the scrubbed validating subset, one can proceed with greater confidence.

Experimenters are transparent in discussing strengths and weaknesses in their data collection processes: Data analysts must do so as well. It is the only way to understand the limitations on what they've learned.

Experimenters recognize they are part of an ongoing process: For an experimenter, there is no "ultimate experiment." A good experiment increases the body of knowledge (even if by saying "there's nothing of interest here") and leads to another experiment. Most business data analysts are not engaged in formal science, *per se*. But they may well be part of an end-to-end innovation process. Such data analysts should develop an understanding of where they fit and how they can make product developers, marketers, and others more effective. (Note: if your organization doesn't have such a process, data analysts are well-advised to behave as though it did.)

Final Remarks: Taken together, relatively simple actions, quite natural to experimenters, can assist data analysts to better understand their data, know how good their data really are, and improve their overall effectiveness. For learning more about the area of data quality, we recommend these books, among many out there, English (1999), Huang et al. (1999), and Redman (2008, 2001).

Cross References

►Fraud in Statistics
►Misuse of Statistics
►Multi-Party Inference and Uncongeniality

References and Further Reading

Box GEP (1976) Science and statistics. J Am Stat Assoc 71:791–799
Box GEP, Draper NR (1987) Empirical model-building and response surfaces. Wiley, New York
English LP (1999) Improving data warehouse and business information quality: methods for reducing costs and increasing profits. Wiley, New York
Huang K-T, Lee YL, Wang RY (1999) Quality information and knowledge. Prentice Hall, New York
Redman TC (2001) Data quality: the field guide. Butterworth-Heinemann Digital Press, Boston, MA
Redman TC (2008) Data driven: profiting from your most important business asset. Harvard Business School Press, Boston, MA

Decision Theory: An Introduction

Martin Peterson
Associate Professor
Eindhoven University of Technology, Eindhoven, Netherlands

Decision theory (see also ►Decision Theory: An Overview) is the theory of rational decision making. This is an interdisciplinary field to which philosophers, economists, psychologists, computer scientists and statisticians contribute their expertise. It is common to distinguish between *normative* and *descriptive* decision theory. Normative decision theory seeks to yield prescriptions about what decision makers are *rationally required* – or *ought* – to do. Descriptive decision theories seek to explain and predict how people *actually* make decisions. Descriptive decision theory is thus an empirical discipline, which has its roots in experimental psychology. Descriptive and normative decision theory are, thus, two separate fields of inquiry, which may or may not be studied independently of each other.

In decision theory, a *decision problem* is situation in which a *decision maker*, (a person, a company, or a society) chooses what to do from a set of *alternative acts*, where the *outcome* of the decision depends on which *state of the world* turns out to be the actual one. Decision problem are classified as decisions under *risk* or *ignorance* (or *uncertainty)* depending on the information available to the agent at the time at which he makes his choice. In decisions under risk the decision maker knows the probability of the possible outcomes, whereas in decisions under ignorance the probabilities are either unknown or non-existent. The term uncertainty is either used as a synonym for ignorance, or as a broader term referring to both risk and ignorance.

Let us consider an example. Suppose that you are thinking about taking out insurance against theft of your

new sports car. Perhaps it costs $300 to take out insurance on a car worth $90,000, and you ask: Is it worth it? In case you know that the probability that the car is stolen is, say, 1 in 1,000, then you are clearly facing a decision under risk. However, in case you are not able to assess the probability of theft then you are on the contrary facing a decision under uncertainty or ignorance.

The most widely applied decision rule for making decisions under risk is to apply the principle of maximizing expected value (or utility). According to this decision rule, the total value of an act equals the sum of the values of its possible outcomes weighted by the probability for each outcome. Hence, the expected values of not taking out fire insurance is 1/1,000 times the value of losing $200,000, whereas the expected value of taking out insurance equals the value of losing $200 (since the insurance guarantees that you will suffer no financial loss in case of a fire).

Modern decision theory is dominated by attempts to axiomatise the principles of rational decision making, and in particular the principle of maximizing expected utility. (The term "utility" refers to a technically precise notion of value.) The first axiomatisation was presented by Ramsey in his paper *Truth and Probability*, written in 1926 but published posthumously in 1931. Ramsey was a philosopher working at Cambridge together with Russell, Moore, and Wittgenstein. In his paper, Ramsey proposed a set of eight axioms for how rational decision makers ought to choose among uncertain prospects. He pointed out that every decision maker behaving in accordance with these axioms will act in a way that is *compatible* with the principle of maximizing expected value, by implicitly assigning numerical probabilities and values to outcomes. However, it does not follow that the decision maker's choices were *actually trigged* by these implicit probabilities and utilities. This way of thinking about rational decision making is very influential in the modern literature, and similar ideas were put forward by Savage in *The Foundations of Statistics* about two decades later.

Another important point of departure in modern decision theory is von Neumann and Morgenstern's book *Theory of Games and Economic Behavior*. Von Neumann and Morgenstern showed how a linear measure of value for outcomes (i.e., a utility function) can be generated from a set of axioms dealing with how rational decision makers ought to choose among lotteries. For von Neumann and Morgenstern a lottery is a probabilistic mixture of outcomes; for example, "a fifty-fifty chance of winning either $100 or a trip to London" is a lottery. They showed that every decision maker behaving in accordance with their axioms implicitly behaves in accordance with the principle of maximizing expected utility, and implicitly assigns numerical utilities to outcomes. The main difference compared to Ramsey's axiomatisation is that von Neumann and Morgenstern presented no novel theory of probability.

The use of the principle of maximizing expected utility has been criticized by some decision theorists. The most famous counter argument was proposed by Allais in the 1950s. Consider the following lotteries in which exactly one winning ticket will be drawn at random.

	Ticket no. 1	Ticket no. 2–11	Ticket no. 12–100
Gamble 1	$1 Million	$1 Million	$1 Million
Gamble 2	$0	$5 Million	$1 Million
Gamble 3	$1 Million	$1 Million	$0
Gamble 4	$0	$5 Million	$0

In a choice between Gamble 1 and Gamble 2 it seems reasonable to choose Gamble 1 since it gives the decision maker $1 Million for sure, whereas in a choice between Gamble 3 and Gamble 4 many people would feel that it makes sense to trade a ten-in-hundred chance of getting $5 Million, against a one-in-hundred risk of getting nothing, and consequently choose Gamble 4. Several empirical studies have confirmed that most people reason in this way. However, no matter what utility one assigns to money, the principle of maximizing expected utility recommends that the decision maker prefers Gamble 1 to Gamble 2 if and only if Gamble 3 is preferred to Gamble 4. There is simply no utility function such that the principle of maximizing utility is consistent with a preference for Gamble 1 to Gamble 2 and a preference for Gamble 4 to Gamble 3. To see why this is so, we calculate the *difference* in expected utility between the two pairs of gambles. Note that the probability that ticket 1 will be drawn is 0.01, and the probability that one of tickets numbered 2–11 will be drawn is 0.1; hence, the probability that one of tickets numbered 12–100 will be drawn is 0.89. This gives the following equations:

$$\begin{aligned} u(\mathrm{G1}) - u(\mathrm{G2}) &= u(1\mathrm{M}) - [0.01u(0\mathrm{M}) \\ &\quad + 0.1u(5\mathrm{M}) + 0.89u(1\mathrm{M})] \\ &= 0.11u(1\mathrm{M}) - [0.01u(0) \\ &\quad + 0.1u(5\mathrm{M})] \end{aligned} \tag{1}$$

$$\begin{aligned} u(\mathrm{G3}) - u(\mathrm{G4}) &= [0.11u(1\mathrm{M}) + 0.89u(0)] \\ &\quad - [0.9u(0\mathrm{M}) + 0.1u(5\mathrm{M})] \\ &= 0.11u(1\mathrm{M}) - [0.01u(0) + 0.1u(5\mathrm{M})] \end{aligned} \tag{2}$$

Equations 1 and 2 show that the difference in expected utility between G1 and G2 is precisely the same as the difference between G3 and G4. Hence, no matter what the decision maker's utility for money is, it is impossible to simultaneously prefer G1 to G2 and to prefer G4 to G3 without violating the expected utility principle. However, since many people who have thought very hard about this example still feel it would be rational to stick to the problematic preference pattern described above, there seems to be something wrong with the expected utility principle.

Let us now move on to the other type of decision problems mentioned above, viz., decisions under ignorance (or uncertainty). There is no single decision rule for decision making under ignorance that is currently widely accepted by decision theorists. However, the *maximin* rule and the *principle of insufficient reason* are two of the most influential rules, which have been widely discussed in the literature. The maximin rule, famously adopted by Rawls, focuses on the worst possible outcome of each alternative. According to this principle, one should *maximize* the *minimal* value obtainable with each act. If the worst possible outcome of one alternative is better than that of another, then the former should be chosen. According to the principle of insufficient reason, adopted by e.g., Bernoulli and Laplace, it holds that if one has *no* reason to think that one state of the world is more probable than another, then all states should be assigned *equal* probability. A well-known objection to the principle of insufficient reason is that that it seems completely arbitrary to infer that all states are equally probable. If one has *no* reason to think that one state is more probable than another, it seems strange to conclude anything at all about probabilities. Or, alternatively put, if one has *no* reason to think that some state is *twice* as probable as another, why not then reason as if that state is twice as probable as the other? Every possible distribution of probabilities seems to be equally justified.

However, not every decision problem can be classified as being either a decision under risk or a decision under ignorance. A major sub-field of modern decision theory is so-called multi-attribute approaches to decision theory. The difference between single- and multi-attribute approaches is that in a single-attribute approach, all outcomes are compared on a single utility scale. For example, in a decision between saving a group of fishermen from a sinking ship at a cost of one million dollars or letting the fishermen die and save the money, the value of a human life will be directly compared with monetary outcomes on a single scale. However, many authors think that such direct comparisons between the value of a human life and money makes no sense – i.e., that human life and money are incommensurable.

The multi-attribute approach seeks to avoid the criticism that money and human welfare are incommensurable by giving up the assumption that all outcomes have to be compared on a common scale. In a multi-attribute approach, each type of attribute is measured in the unit deemed to be most suitable for that attribute. Perhaps money is the right unit to use for measuring financial costs, whereas the number of lives saved is the right unit to use for measuring human welfare. The total value of an alternative is thereafter be determined by aggregating the attributes, e.g., money and lives, into an overall ranking of the available alternatives.

Here is an example. Mary has somehow divided the relevant objectives of her decision problem into a list of attributes. For illustrative purposes, we assume that the attributes are (a) the number of lives saved, (b) the financial aspects of the decision, (c) the political implications of the decision, and (d) the legal aspects of the decision. Now, to make a decision, Mary has to gather information about the degree to which each attribute can be realized by each alternative. Consider the following table, in which we list four attributes and three alternatives.

	Attribute 1	Attribute 2	Attribute 3	Attribute 4
Alt. a_1	1	3	1	2
Alt. a_2	3	1	3	1
Alt. a_3	2	2	2	2

The numbers represent the degree to which each attribute is fulfilled by the corresponding alternative. For example, in the leftmost column the numbers show that the second alternative fulfills the first attribute to a higher degree than the first alternative, and so on. So far the ranking is ordinal, so nothing follows about the "distance" in value between the numbers. However, in many applications of the multi-attribute approach it is of course natural to assume that the numbers represent more than an ordinal ranking. The number of people saved from a sinking ship can, for instance, be measured on a ratio scale. This also holds true of the amount of money saved by not rescuing the fishermen. In this case, nothing prevents the advocate of the multi-attribute approach to use a ratio or interval scale if one so wishes.

Several criteria have been proposed for choosing among alternatives with multiple attributes. It is common to distinguish between additive and non-additive criteria. Additive criteria assign weights to each attribute, and rank

alternatives according to the weighted sum calculated by multiplying the weight of each attribute with its value. The weights are real numbers between zero and one, which together sum up to one. Obviously, this type of criterion makes sense only if the degree to which each alternative satisfies any given attribute can be represented at least on an interval scale, i.e., if it makes sense to measure value in quantitative terms. Let us, for the sake of the argument, suppose that this is the case for the numbers in table above, and suppose that all attributes are assigned equal weights, i.e., 1/4. This implies that the value of alternative a_1 is $1/4 \cdot 1 + 1/4 \cdot 3 + 1/4 \cdot 1 + 1/4 \cdot 2 = 7/4$. Analogous calculations show that the value of a_2 is 2, while that of a_3 is also 2. Since we defined the ranking by stipulating that a higher number is better than a lower, it follows that a_2 and a_3 are better than a_1.

Another major sub-field of contemporary decision theory is *social choice theory*. Social choice theory seeks to analyze collective decision problems: How should a group aggregate the preferences of its individual members into a joint preference ordering? In this context, a group could be any constellation of individuals, such as a married couple, a number of friends, the members of a club, the citizens of a state, or even all conscious beings in universe. A *social choice problem* is any decision problem faced by a group, in which each individual is willing to state at least ordinal preferences over outcomes. Once all individuals have stated such ordinal preferences we have a set of *individual preference orderings*. The challenge faced by the social decision theorist is to somehow combine the individual preference ordering into a *social preference ordering*, that is, a preference ordering that reflects the preferences of the group. A *social state* is the state of the world that includes everything that individuals care about, and the term *social welfare function* (SWF) refers to any decision rule that aggregates a set of individual preference orderings over social states into a social preference ordering over those states. The majority rule used in democratic elections is an example of a SWF.

The most famous technical result in social choice theory is Arrow's impossibility theorem, according to which there is no SWF that meets a set of relatively weak normative conditions. A natural interpretation is that social decisions can never be rationally justified, simply because every possible mechanism for generating a social preference ordering – including the majority rule – is certain to violate at least one of Arrow's conditions. This result received massive attention in academic circles, and in the 1960s and 70s, many people took the theorem to prove that "democracy is impossible". However, the present view is that the situation is not that bad. By giving up or modifying some of Arrow's conditions one can formulate coherent SWFs that are not vulnerable to his impossibility result. Today, the theorem is interesting mainly because it opened up an entirely new field of inquiry.

Cross References

►Bayesian Statistics
►Decision Theory: An Overview
►Imprecise Probability
►Loss Function
►Multicriteria Decision Analysis
►Multiple Statistical Decision Theory
►Philosophical Foundations of Statistics
►Statistics and Gambling

References and Further Reading

Allais M (1953) Le comportement de l'homme rationnel devant le risque: critique des postulates et axiomes de l'ecole Américaine. Econometrica 21:503–546

Arrow KJ (1951) Social choice and individual values, 2nd edn (1963). Wiley, New York

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decisions under risk. Econometrica 47:263–291

Peterson M (2009) An introduction to decision theory. Cambridge University Press, Cambridge

Ramsey FP (1926) Truth and Probability, Ramsey, 1931. In: Braithwaite RB (ed) The foundations of mathematics and other logical essays, Ch. VII. Kegan Paul, London, pp 156–198

Savage LJ (1954) The foundations of statistics, 2nd edn (1972, Dover). Wiley, New York

Trench, Trubner & Co., Harcourt, Brace and Company, New York

von Neumann J, Morgenstern O (1947) Theory of games and economic behavior, 2nd edn. Princeton University Press (1st edn without utility theory)

Decision Theory: An Overview

Sven Ove Hansson
Professor and Head
Royal Institute of Technology, Stockholm, Sweden

Decision theory (see also ►Decision Theory: An Introduction) is the systematic study of *goal-directed behavior under conditions when different courses of action (options) can be chosen*. The focus in decision theory is usually on the outcome of decisions as judged by pre-determined criteria or, in other words, on means-ends rationality. Decision theory has developed since the middle of the twentieth century through contributions from several academic disciplines. In this overview over fundamental decision theory, the

focus will be on how decisions are represented and on decision rules intended to provide guidance for decision-making. Finally two paradoxes will be presented in order to exemplify the types of issues that are discussed in modern decision theory.

The Representation of Decisions

The standard representation of a decision problem requires that we specify the alternatives available to the decision-maker, the possible outcomes of the decision, the values of these outcomes, and the factors that have an influence on the outcome.

Alternatives

To decide means to choose among different alternatives (options). In some decision problems, the set of alternatives is *open* in the sense that new alternatives can be invented or discovered by the decision-maker. A typical example is your decision how to spend tomorrow evening. In other decision problems, the set of alternatives is *closed* so that no new alternatives can be added. Your decision how to vote in the upcoming elections will probably be an example of this. There will be a limited number of alternatives (candidates or parties) between which you can choose.

In real life, many if not most decisions come with an open set of alternatives. In decision theory, however, alternative sets are commonly assumed to be closed. The reason for this is that closed decision problems are much more accessible to theoretical treatment. If the alternative set is open, a definitive solution to a decision problem is not in general available.

In informal deliberations about decisions, we often refer to alternatives that can be combined with each other. Hence, when deciding how to spend tomorrow evening you may begin by choosing between eating out and going to the cinema, but end up deciding to do both. In decision theory, the alternatives are assumed to be *mutually exclusive*, i.e., no two of them can both be realized. However, this difference is not very important since you can always convert a set of compatible alternatives to a set of mutually exclusive ones. This is done by listing all the possible combinations (in this example: eating out and not going to the cinema, eating out and going to the cinema, etc.).

States of Nature

The effects of a decision depend not only on what choice the decision-maker makes, but also on various factors beyond the decision-maker's control. Some of these extraneous factors constitute *background information* that the decision-maker has access to. Others are unknown. They may depend, for instance, on the actions of other persons and various natural events.

In decision theory, it is common to summarize the various unknown extraneous factors into a number of cases, called *states of nature*. The states of nature include decisions by other persons. This is a major difference between decision theory and game theory. In game theory, decisions by several persons that may compete or cooperate are treated on a par with each other in the formal representation. In decision theory, the focus is on one decision-maker, and the actions and choices of others are treated differently, namely in the same way as natural events.

As an example, consider a young boy, Peter, who makes up his mind whether or not to go to the local soccer ground to see if there is any soccer going on that he can join. The effect of that decision depends on whether there are any soccer players present. In decision theory, this situation is described in terms of two states of nature, "players present" and "no players present."

Outcomes

The possible *outcomes* of a decision are determined by the combined effects of a chosen alternative and the state of nature that materializes. Hence, if Peter goes to the soccer ground and there are no players present, then the outcome can be summarized as "walk and no soccer," if he goes and there are players present then the outcome is "walk and soccer," and if he does not go then the outcome is "no walk and no soccer."

Decision Matrices

The alternatives, the states of nature, and the resulting outcomes in a decision can be represented in a *decision matrix*. A decision matrix is a table in which the alternatives are represented by rows and the states of nature by columns. For each alternative and each state of nature, the decision matrix assigns an outcome (such as "walk, no soccer" in our example). The decision matrix for Peter's decision is as follows:

	No soccer players	Soccer players
Go to soccer ground	Walk, no soccer	Walk, soccer
Stay home	No walk, no soccer	No walk, no soccer

Such a matrix provides a clear presentation of the decision, but it does not contain all the information that the decision-maker needs to make the decision. The most important missing information concerns how the outcomes are valued.

Value Representation

When we make decisions, or choose among options, we try to obtain as good an outcome as possible, according to some standard of what is good or bad. The choice of a value-standard for decision-making is usually not considered to fall within the subject matter of decision theory. Instead, decision theory assumes that such a standard is available from other sources, perhaps from moral philosophy.

There are two major ways to express our evaluations of outcomes. One of these is *relational representation*. It is expressed in terms of the three comparative value notions, namely "better than" (*strong preference*, >), "equal in value to" (*indifference*, ≡), and "at least as good as" (*weak preference*, ≥). These three notions are interconnected according to the following two rules:

1. A is better than B if and only if A is at least as good as B but B is not at least as good as A. (A > B if and only if A ≥ B and not B ≥ A.)
2. A is equally good as B if and only if A is at least as good as B and B is at least as good as A. (A ≡ B if and only if A ≥ B and B ≥ A.)

The other major method to express our evaluations of outcomes is *numerical representation*. It consists in assigning numbers to the possible outcomes; such that an outcome has a higher number than another if and only if it is preferred to it. In an economic context, willingness to pay is often used as a measure of value. If a person is prepared to pay, say $500 for a certain used car and $250 for another, then these sums can be used to express her (economic) valuation of the two vehicles.

However, not all values are monetary. According to some moral theorists, all values can instead be reduced to one unit of measurement, *utility*. This entity may or may not be identified with units of human happiness. According to utilitarian moral theory, decision-makers should, at least in principle, always (try to) maximize total utility.

Decision theorists often use numerical values as abstract tools in the analysis of decisions. These values may be taken to represent utilities, but only in a rather abstract sense since they are not based on any method to measure utilities.

Once we have a numerical representation of value, we can replace the verbal descriptions of outcomes in a decision matrix by these values. In our example, suppose that Peter likes to play soccer but does not like walking to the soccer ground and back home. Then his utilities may be representable as follows:

	No soccer players	Soccer players
Go to soccer ground	0	10
Stay home	3	3

Mainstream decision theory is almost exclusively devoted to problems that can be expressed in matrices of this type, *utility matrices* (payoff matrices).

Probability or Uncertainty

Decisions are often categorized according to how much the decision-maker knows beforehand about what state of nature will in fact take place. In an extreme case, the decision-maker knows for sure which state of nature will obtain. If, in the above example, Peter knows with certainty that there are players at the soccer ground, then this makes his decision very simple. The same applies if he knows that there are no players. Cases like these, when only one state of nature needs to be taken into account, are called *decision-making under certainty*.

Non-certainty is usually divided into two categories, called *risk* and *uncertainty*. A decision is made under risk if it is based on exact probabilities that have been assigned to the relevant states of nature; otherwise it is made under uncertainty. Decisions at the roulette table are good examples of decisions under risk since the probabilities are known (although some players do not pay much attention to them). A decision whether to marry is a good example of a decision under uncertainty. There is no way to determine the probability that a marriage will be successful, and presumably few prospective brides or grooms would wish to base their decision on precise probability estimates of marital success or failure.

In some cases, we do not even have a full list of the relevant states of affairs. Hence, decisions on the introduction of a new technology have to be made without full knowledge of the possible future social states in which the new technology will be used. Such cases are referred to as decision-making under *great uncertainty*, or *ignorance*. This adds up to the following scale of knowledge situations in decision problems:

Certainty	It is known what states of nature will occur
Risk	The states of nature and their probabilities are known
Uncertainty	The states of nature are known but not their probabilities
Great uncertainty, ignorance	Not even the states of nature are known

The probabilities referred to in decision theory may be either objective or subjective. In some applications, reliable estimates of probabilities can be based on empirically known frequencies. As one example, death rates at high exposures to asbestos are known from epidemiological studies. In most cases, however, the basis for probability estimates is much less secure. This applies for instance to failures of a new, as yet untried technological device. In such cases we have to resort to subjective estimates of the objective probabilities. Some decision theorists deny the existence of true objective probabilities and regard all probabilities as expressions of degrees of belief, which are of course strictly subjective.

In cases when exact probabilities are not known, uncertainty can be expressed with various, more complex measures.

Binary measures: The probability values are divided into two groups, possible and impossible values (or attention-worthy and negligible values). Usually, the former form a single interval. Then the uncertainty can be expressed in terms of an interval, for instance: "The probability of a nuclear war in the next thirty years is between 10 and 25 per cent."

Multivalued measures: A numerical measure is used to distribute plausibility over the possible probability values. This measure may (but need not) be a (second-order) probability measure. Then, instead of just saying that the probability is between 10% and 25%, we can say that there is a 5% probability that the probability is between 17% and 18%, a 4% probability that it is between 18% and 19%, etc.

Robustness measures: The more certain we are about a probability, the less willing we are to change our estimate of it. Therefore, willingness to change one's estimate of a probability when new information arrives can be used as a measure of uncertainty.

Decision Rules

Decision theorists have developed a series of decision rules, intended to ensure that decisions are made in a systematic and rational way.

The Maximin Rule

Among the decision rules that are applicable without numerical information, the *maximin* rule is probably the most important one. For each alternative, we define its *security level* as the worst possible outcome with that alternative. The maximin rule urges us to choose the alternative that has the highest security level. In other words, we *maxi*mize the *mini*mal outcome.

The maximin rule has often, and quite accurately, been described as a cautious rule. It has also been described as pessimistic, but that is an unfortunate terminology, since caution and pessimism are quite independent of each other.

As an example of the maximin rule, consider the following variant of the soccer example from above:

	No soccer players	Soccer players
Go to soccer ground	Walk, no soccer	Walk, soccer
Stay home	No walk, no soccer	No walk, no soccer

The preferences are:

Walk, soccer
is better than
No walk, no soccer
is better than
Walk, no soccer

The security level of Stay home is "no walk, no soccer" whereas that of Go to soccer ground is "walk, no soccer". Since the former is better than the latter, in order to maximize the security level, Peter would have to stay at home. Consequently, this is what the maximin rule recommends him to do.

Even though the maximin rule can be applied to relational value information as above, it is easier to apply if the value information is presented in numerical form. Again, consider the following utility matrix:

	No soccer players	Soccer players
Go to soccer ground	0	10
Stay home	3	3

Here, the security level of Stay home is 3 whereas that of Go to soccer ground is 0. Since 3 is larger than 0, the maximin rule recommends Peter to stay at home, just as in the relational presentation of the same example.

The Maximax Rule

The best level that we can at all obtain if we choose a certain alternative is called its *hope level*. According to the *maxi-max* rule, we should choose the alternative whose hope level (best possible outcome) is best. Just like the maxi-min rule, the maxi-max rule can be applied even if we only have relational (non-numerical) value information. Consider again the soccer example. The hope level of

Stay home is "no walk, no soccer," and that of Go to soccer ground is "walk, soccer" that Peter values higher. Hence, the maximax rule urges Peter to go to the soccer ground. Similarly, in the numerical representation, Stay home has the hope level 3 and Go to soccer ground has 10; hence again Peter is advised to go to the soccer ground.

The maximax rule has seldom been promoted. Contrary to the maximin rule, it is often conceived as irrational or as an expression of wishful thinking. It is indeed hardly commendable as a general rule for decision-making. However, in certain subareas of life, taking chances may be beneficial, and in such areas behavior corresponding to the maximax rule may not be irrational. Life would probably be much duller unless at least some of us were maximaxers on at least some occasions.

The Cautiousness Index

There is an obvious need for a decision criterion that does not force us into the extreme cautiousness of the maximin rule or the extreme incautiousness of the maximax rule. A middle road is available, but only if we have access to numerical information. We can then calculate a weighted average between the security level and the hope level, and use this weighted average to rank the alternatives. Let us again consider the numerical presentation of the soccer example:

	No soccer players	Soccer players
Go to soccer ground	0	10
Stay home	3	3

For each alternative A, let *min*(A) be its security level and *max*(A) its hope level.

In our example, *min*(Go to soccer ground) = 0 and *max*(Go to soccer ground) = 10. If we choose to assign equal weight to the security level and the hope level, then the weighted value of Go to soccer ground is $0.5 \times 0 + 0.5 \times 10 = 5$. Since *min*(Stay home) = *max*(Stay home) = 3, the weighted average value of Stay home is 3. Hence, with these relative weights, Peter is recommended to go to the soccer ground. More generally speaking, each alternative A is assigned a value according to the following formula:

$$\alpha \times min\,(\mathrm{A}) + (1 - \alpha) \times max\,(\mathrm{A})$$

If $\alpha = 1$, then this rule reduces to the maximin criterion and if $\alpha = 0$, then it reduces to the maximax criterion. The index α is often called the *Hurwicz α index*, after economist Leonid Hurwicz who proposed it in 1950. It is also often called the *optimism-pessimism index*, but the latter terminology should be avoided since the index represents the degree of (un)cautiousness rather than that of optimism. It can more appropriately be called the *cautiousness index*.

Minimax Regret

Utility information also allows for another decision criterion that puts focus on how an outcome differs from other outcomes that might have been obtained under the same state of affairs, if the decision-maker had chosen another alternative. In our example, if Peter stays home and there are players at the soccer ground, then he has made a loss that may give rise to considerable regret. If he goes to the soccer ground and there is no one there to play with him, then he has also made a loss, but a smaller one. The decision rule based on these considerations is usually called the *minimax regret* criterion. It also has other names, such as *minimax risk*, *minimax loss*, and *minimax*.

In this decision rule the degree of regret is measured as the difference between the utility obtained and the highest utility level that could have been obtained (in the same state of the world) if another alternative had been chosen. A *regret matrix* can quite easily be derived from a utility matrix: Replace each entry by the number obtained by subtracting it from the highest utility in its column. In our example, the regret matrix will be as follows:

	No soccer players	Soccer players
Go to soccer ground	3	0
Stay home	0	7

The minimax regret criterion advices the decision-maker to choose the option with the lowest maximal regret (to *mini*mize *maxi*mal *regret*). In this case it recommends Peter to go to the soccer ground.

Just like the maximin rule, the minimax regret rule can be described as cautious, but they apply cautiousness to different aspects of the decision (the value of the actual outcome respectively its regrettableness). As this example shows, they do not always yield the same recommendation.

Expected Utility

None of the above decision rules requires or makes use of probabilistic information. When probabilities are available, the dominating approach is to maximize expected utility (EU). Then to each alternative is assigned a weighted average of its utility values under the different states of nature, with the probabilities of these states used as weights.

In the above example, suppose that based on previous experience Peter believes the probability to be 0.4 that there are players at the soccer ground. We can enter the probabilistic information into the column headings of the utility matrix as follows:

	No soccer players Probability 0.6	Soccer players Probability 0.4
Go to soccer ground	0	10
Stay home	3	3

The expected (probability-weighted) utility of going to the soccer ground is $0.6 \times 0 + 0.4 \times 10 = 4$, and that of staying at home is $0.6 \times 3 + 0.4 \times 3 = 3$. If Peter wants to maximize expected utility then he should, in this case, go to the soccer ground. Obviously, the recommendation would be different with other probabilities.

For a general formula representing expected utility, let there be n outcomes, to each of which is associated a utility and a probability. The outcomes are numbered, so that the first outcome has utility u_1 and probability p_1, the second has utility u_2 and probability p_2, etc. Then the expected utility is defined as follows:

$$p_1 \times u_1 + p_2 \times u_2 + \ldots + p_n \times u_n$$

Expected utility maximization based on subjective probabilities is commonly called *Bayesian decision theory*, or Bayesianism. (The name derives from Thomas Bayes, 1702–1761, who provided much of the mathematical foundations for probability theory). According to Bayesianism, a rational decision-maker should have a complete set of probabilistic beliefs (or at least behave as if she had one) and all her decisions should take the form of choosing the option with the highest expected utility.

Two Paradoxes of Decision Theory

Much of the modern discussion on decision theory has been driven by the presentation of paradoxes, i.e., situations in which decision-making criteria that seem to epitomize rationality nevertheless give rise to decisions that are contrary to most people's intuitions. The following two decision paradoxes serve to exemplify the kinds of philosophical problems that are discussed in the decision-theoretical research literature.

Ellsberg's Paradox

Daniel Ellsberg has presented the following decision problem: We have an urn that contains 30 red ball and 60 balls that are either black or yellow. The distribution between the latter two colors is unknown. A ball is going to be drawn at random from the urn. Before that is done you are offered bets by two persons.

Anne offers you to bet either on red or on black. If you bet on red, then you will receive € 100 if the drawn ball is red and nothing if it is either black or yellow. Similarly, if you bet on black, then you will get € 100 if the ball is black, and nothing if it is red or yellow.

Betty offers you to bet either on red-or-yellow or on black-or-yellow. If you bet on red-or-yellow, then you will get € 100 if the drawn ball is either red or yellow, but nothing if it is black. If you bet on black-or-yellow, then you will get € 100 if the drawn ball is either black or yellow, but nothing if it is red.

Most people, it turns out, prefer betting red to betting black, but they prefer betting black-or-yellow to betting red-or-yellow. It is fairly easy to show that this pattern is at variance with expected utility maximization, i.e., there is no way to assign utilities that would make this pattern compatible with the maximization of expected utility. Ellsberg's own conclusion was that decision-making must take into account factors not covered by probabilities and utilities, in particular the degree of uncertainty of the various probability estimates.

Another problem with this pattern is that it violates the *sure-thing principle* that is a much acclaimed rationality criterion for decisions. To introduce the principle, let A and B be two alternatives, and let S be a state of nature such that the outcome of A in S is the same as that of B. In other words, the outcome in case of S is a "sure thing," not influenced by the choice between A and B. The sure-thing principle says that if the "sure thing" (i.e., the common outcome in case of S) is changed, but nothing else is changed, then the choice between A and B is not affected.

As an example, suppose that a whimsical host wants to choose a dessert by tossing a coin. You are invited to choose between alternatives A and B. In alternative A, you will have fruit in case of heads and nothing in case of tails. In alternative B you will have pie in case of heads and nothing in case of tails. The decision matrix is as follows:

	Heads	Tails
A	Fruit	Nothing
B	Pie	Nothing

When you have made up your mind and announced which of the two alternatives you prefer, the whimsical host suddenly remembers that he has some ice cream, and

changes the options so that the decision matrix is now as follows:

	Heads	Tails
A	Fruit	Ice cream
B	Pie	Ice cream

Since only a "sure thing" (an outcome that is common to the two alternatives) has changed between the two decision problems, the sure-thing principle demands that you do not change your choice between *A* and *B* when the decision problem is revised in this fashion. If, for instance, you chose alternative *A* in the first decision problem, then you are bound to do so in the second problem as well.

In this example, the sure-thing principle appears rational enough, and it would seem natural to endorse it as a general principle for decision-making. Ellsberg's paradox shows that is not quite as self-evident as it may seem to be at first sight.

Newcomb's Paradox

The following paradox was proposed by the physicist William Newcomb: In front of you are two boxes. One of them is transparent, and you can see that it contains $1,000. The other is covered, so that you cannot see its contents. It contains either $1,000,000 or nothing. You have two options to choose between. One is to take both boxes, and the other is to take only the covered box. A predictor who has infallible (or almost infallible) knowledge about your psyche has put the million in the covered box if he predicted that you will only take that box. Otherwise, it is empty.

Let us apply maximized expected utility to the problem. If you decide to take both boxes, then the predictor has almost certainly foreseen this and put nothing in the covered box. Your gain is $1,000. If, on the other hand, you decide to take only one box, then the predictor has foreseen this and put the million in the box, so that your gain is $1,000,000. In other words, maximization of expected utility urges you to take only the covered box.

There is, however, another plausible approach to the problem that leads to a different conclusion. If the predictor has put nothing in the covered box, then it is better to take both boxes than to take only one, since you will gain $1,000 instead of nothing. If he has put the million in the box, then too it is better to take both boxes, since you will gain $1,001,000 instead of $1,000,000. Thus, taking both boxes is better in all states of nature. (It is a *dominating* option.) It seems to follow that you should take both boxes, contrary to the rule of maximizing expected utilities.

The two-box strategy in Newcomb's problem maximizes the "real gain" of having chosen an option, whereas the one-box strategy maximizes the "news value" of having chosen an option. The very fact that a certain decision has been made in a certain way changes the probabilities that have to be taken into account in that decision.

In *causal decision theory*, expected utility calculations are modified so that they refer to real value rather than news value. This is done by replacing standard probabilities by some formal means for evaluating the causal implications of the different options. Since there are several competing philosophical views of causality, there are also several formulations of causal decision theory.

About the Author

Sven Ove Hansson is professor in philosophy and Head of the division of Philosophy, Royal Institute of Technology, Stockholm. He is Editor-in-Chief of *Theoria*. His research areas include value theory, decision theory, epistemology, belief dynamics and the philosophy of probability. He is the author of well over 200 articles in refereed journals. His books include *A Textbook of Belief Dynamics. Theory Change and Database Updating* (Kluwer 1999) and *The Structures of Values and Norms* (CUP 2001). He is a member of the Royal Swedish Academy of Engineering Sciences (IVA).

Cross References

►Bayesian Statistics
►Decision Theory: An Introduction
►Imprecise Probability
►Loss Function
►Multicriteria Decision Analysis
►Multiple Statistical Decision Theory
►Philosophical Foundations of Statistics
►Statistical Inference for Quantum Systems
►Statistics and Gambling

References and Further Reading

Gärdenfors P, Sahlin NE (eds) (1988) Decision, probability, and utility. Cambridge University Press, Cambridge

Hansson SO (2005) Decision theory. A brief introduction. http://www.infra.kth.se/~soh/decisiontheory.pdf

Luce RD, Raiffa H (1957) Games and decisions: introduction and critical survey. Wiley, New York

Resnik MD (1987) Choices – an introduction to decision theory. University of Minnesota Press, Minneapolis

Decision Trees for the Teaching of Statistical Estimation

Carlos Eduardo Valdivieso Taborga
Professor of Statistics
Universidad Privada Boliviana, Cochabamba, Bolivia

Introduction

In many experiments or research, researchers want to know whether the difference in means, proportions, or variances between two populations is significant, or sometimes if the average, proportion, or variance are near to a standard value. They also will be interested in an interval which is expected to contain the value of the difference in means, proportions, and variances of the two populations, or the average, proportion, or variance of a population. In all these cases, researchers have to rely on statistical estimation and choose the most appropriate procedure.

Since many different confidence intervals have been proposed in the literature, the choice can sometimes be difficult and confusing, especially for students. In this chapter we will show how decision trees can help in finding the most appropriate ▶confidence interval.

Types of Estimations

An estimator is a statistic of a sample that is used to estimate a population parameter such as mean, proportion, or variance. An estimate is a specific value of the observed statistics. In real life the parameter values of the populations studied are unknown. To obtain the best achievable information for the parameters we rely on the sample data and apply a procedure that is called statistical estimation. For this purpose we can use one of the following:

- *Point estimator* as is a single value that estimates the value of the parameter. Two estimators are, for example, the sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ (as estimator for the population mean), and sample variance: $s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$ (as estimator for the population variance).
- *Interval estimation*, which provides a numerical range in which it is intended to find the value of the parameter under study.

One disadvantage of point estimators is that they do not provide information on how close they are to the true value of the parameter. To incorporate some measure of the accuracy of an estimate, we determine a range (confidence interval). This range includes the value of the parameter with a certain predetermined probability $1 - \alpha$, which is called the confidence level, or sometimes confidence coefficient (and is typically taken to be 0.95 or 0.99).

Decision Trees for Choosing the Proper Confidence Interval

Because in statistics there are many different confidence interval expressions we'll attempt to show here how one can choose the appropriate one, in other words, the one that is best suited for the data at hand. Choice of the interval depends on several conditions:

- Which population parameter has to be estimated?
- How many populations are under investigation?
- With one population, does the sample come from a normal population or not, do we know the population variance or not, and also is the sample size small (less than 30) or large?
- In the case of two populations, are the samples independent or dependent and are population variances equal or not?

To facilitate making the proper decision regarding an adequate confidence interval a researcher may use the decision trees, discussed in the next section. Their use is simple; however, we will explain the procedure and give an example to provide more clarity.

Example

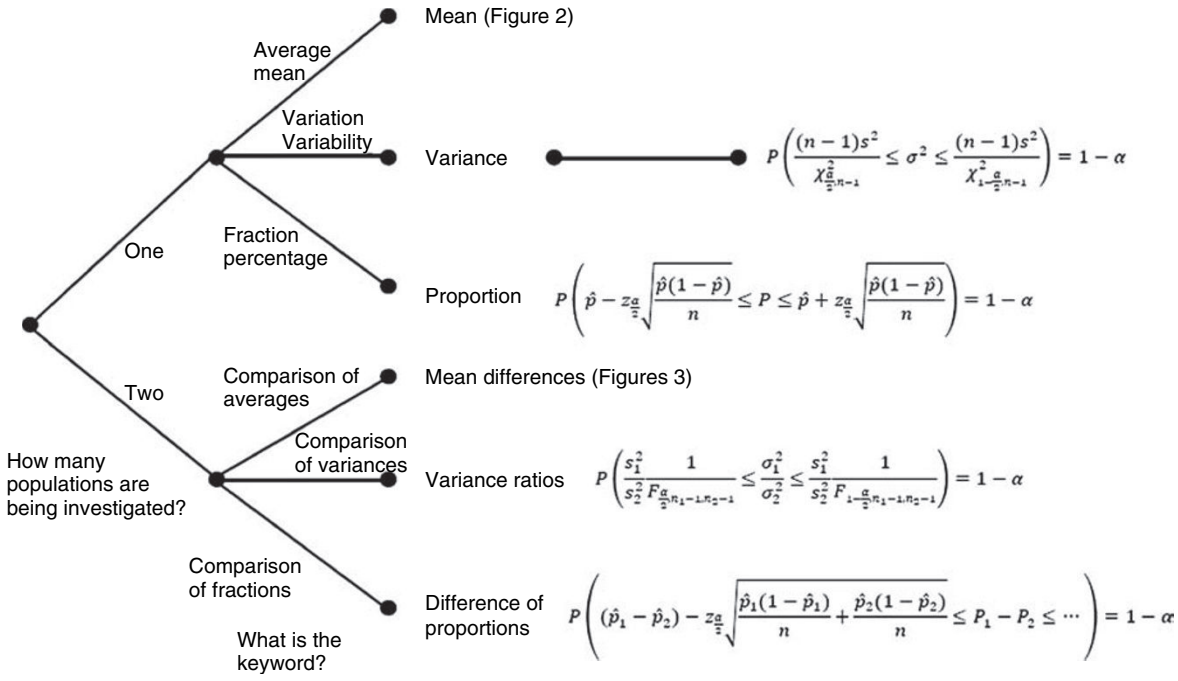
A new filtering device is installed at a chemical plant. Random samples yield the following information of percentage of impurity before and after installation:

Sample/Statistics	Mean	Variance	Sample size
Before	12.5	101.17	8
After	10.2	94.73	9

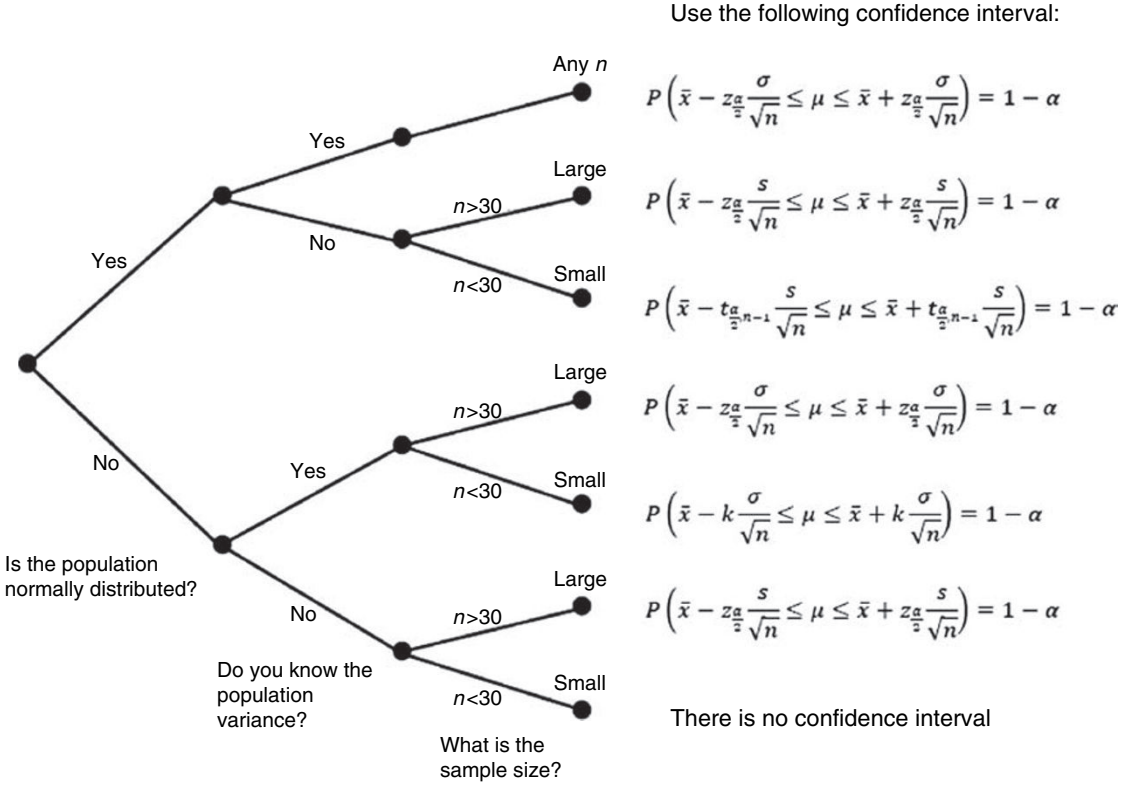
(a) Establish a way to estimate if the impurity has the same variability before and after installing the new filter device. Use $\alpha = 0.05$.
(b) Does the filtering device have reduced the average impurity significantly? Use $\alpha = 0.05$.
(c) If the average percentage of impurities allowed in the chemical plant is 5%, is the goal reached?

(a) From Fig. 1, we get the following information:

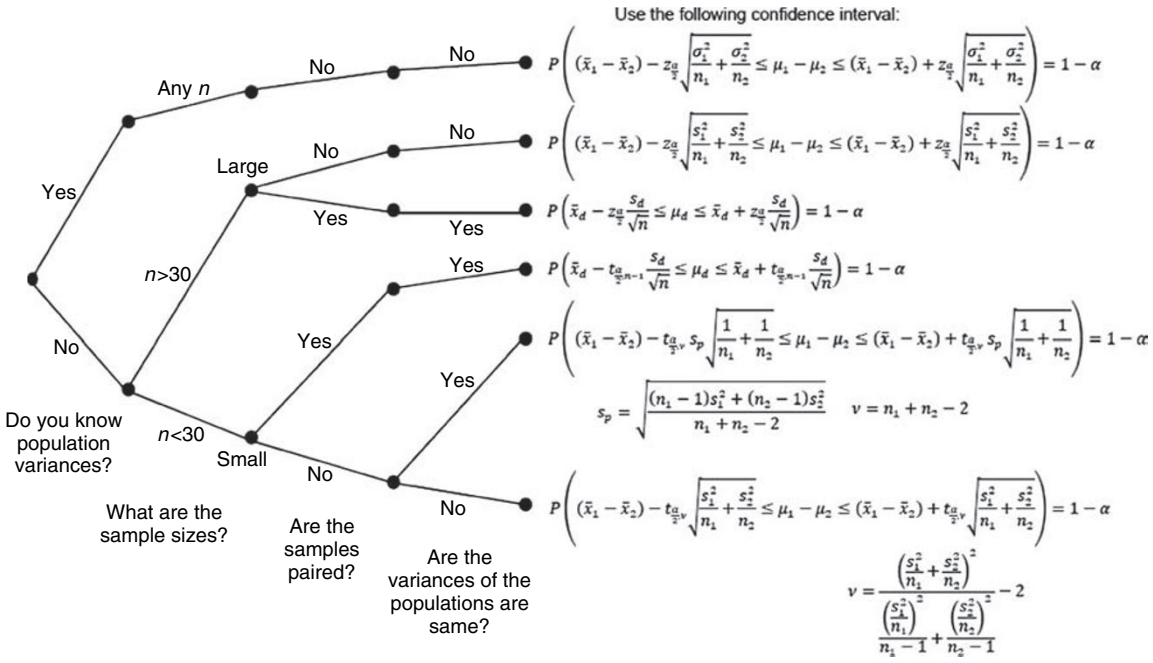
- How many populations are being investigated? *Since we have a before-and-after case, obviously there are two populations.*
- What is the keyword? *Comparison of variances.*



Decision Trees for the Teaching of Statistical Estimation. Fig. 1 Decision tree for choosing the proper confidence interval



Decision Trees for the Teaching of Statistical Estimation. Fig. 2 Decision tree for choosing the proper confidence interval for the mean of the population



Decision Trees for the Teaching of Statistical Estimation. Fig. 3 Decision tree for choosing the proper confidence interval for the difference of means of normal populations

- Therefore, we choose the interval for the variance ratio

$$P\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2,n_1-1,n_2-1}}\le\frac{{\sigma_1}^2}{{\sigma_2}^2}\le\frac{s_1^2}{s_2^2}\frac{1}{F_{1-\alpha/2,n_1-1,n_2-1}}\right)=1-\alpha.$$

Replacing the sample information and critical points of the F distribution we obtain:

$$P\left(\frac{101.17}{94.73}\frac{1}{4.53}\le\frac{{\sigma_1}^2}{{\sigma_2}^2}\le\frac{101.17}{94.73}\frac{1}{0.20}\right)$$
$$=95\%\;P\left(0.24\le\frac{{\sigma_1}^2}{{\sigma_2}^2}\le5.23\right)=95\%.$$

We conclude that there is no significant difference between the impurity variability before and after installing the new filter device (since the interval contains 1).

(b) From Fig. 1, we acquire answers to the following questions:

- How many populations are being investigated? *Two populations, with the same argument as above.*
- What is the keyword? *Comparison of averages.* We will make use of Fig. 3.
- Do we know population variances? *No.*
- What are the sample sizes? *Less than 30.*
- Are the variances of the populations the same? *Yes* (Result of part a).

Therefore, the appropriate confidence interval is for mean difference:

$$P\left((\bar{x}_1-\bar{x}_2)-t_{\alpha/2,v}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\le\mu_1-\mu_2\le(\bar{x}_1-\bar{x}_2)\right.$$
$$\left.+t_{\alpha/2,v}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

with

$$s_p=\sqrt{\frac{(n_1-1)\,{s_1}^2+(n_2-1){s_2}^2}{n_1+n_2-2}}$$

By including the sample information and critical points of the t distribution we obtain

$$s_p=\sqrt{\frac{(8-1)101.17+(9-1)94.73}{8+9-2}}=9.89.$$

$$P\left((12.5-10.2)-(2.13)9.89\sqrt{\frac{1}{8}+\frac{1}{9}}\le\mu_1-\mu_2\le\right.$$
$$\left.(12.5-10.2)+(2.13)9.89\sqrt{\frac{1}{8}+\frac{1}{9}}\right)=0.95$$
$$P(-7.94\le\mu_1-\mu_2\le12.54)=0.95$$

The filtering device has not been effective because the average impurity before and after is the same (since the confidence interval encompasses 0).

(c) From Fig. 1, we get the following information:

- How many populations are being investigated? *Since only want to see if the new device has helped to achieve the goal: one.*
- What is the keyword? *Mean*. We will use Fig. 2.
- Is the population normally distributed? *Let us assume that the sample came from a normal population.*
- Do we know the population variance? *No*.
- What is the sample size? *Small (less than 30)*.

Thus, the appropriate confidence interval for the mean is:

$$P\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Replacing the sample information and critical points of the t distribution:

$$P\left(10.2 - 2.31\frac{9.73}{\sqrt{9}} \le \mu \le 10.2 + 2.31\frac{9.73}{\sqrt{9}}\right) = 95\% \quad P(2.71 \le \mu \le 17.69) = 95\%$$

The impurity is within the limit allowed by the plant.

References and Further Reading

Freund JE, Simon GA (1994) Estadística Elemental, Octava Edición. Prentice Hall, México

Levin RI, Rubin DS (1996) Estadística para Administradores, Sexta Edición. Prentice Hall Hispanoamericana S.A., México

Mason R, Lind D (1995) Estadística para Administración y Economía, Séptima Edición. Alfaomega, México

Mendenhal W (1990) Estadística para Administradores, Segunda Edición. Grupo Editorial Iberoamérica, México

Miller I, Freund J, Jonson R (1992) Probabilidad y Estadística para Ingenieros, Cuarta Edición. Editorial Prentice Hall Hispanoamericana S.A

Degradation Models in Reliability and Survival Analysis

Mikhail Nikulin
Professor
Université Victor Segalen, Bordeaux, France

Introduction

Traditionally the failure time data are usually used for product reliability estimation. Failures of highly reliable units are rare and other information should be used in addition to censored failure time data. One way of obtaining complementary reliability information is to use higher levels of experimental factors or covariates to increase the number of failures and, hence, to obtain reliability information quickly. The *accelerated life testing* (ALT, see ►Accelerated Lifetime Testing) of technical, biological or biotechnical systems is an important practical method of estimation of the reliability of new systems without having to wait through the operating life of them. It is evident that the *extrapolating reliability* from ALT always carries the risk that the accelerating stresses do not properly excite the failure mechanism which dominates at operating (*usual, normal, standard*) stresses. Another way of obtaining this complementary reliability information is to measure some parameters (covariates) that characterize the aging and the degradation of the product in time. In analysis of *longevity* of highly reliable complex industrial or biological systems, the degradation processes provide additional information about the *aging, degradation, fatigue, internal wear* and *deterioration* of systems, and from this point of view the degradation data are really a very rich source of additional information and often offer many advantages over failure time data. Degradation is the natural response for some tests, and it is also natural that with degradation data it is possible to make useful reliability and statistical inference, even with no failures. It is evident that it may be difficult or costly to collect degradation measures from some components or materials. Sometimes it is possible to apply the expert's estimation of the level of degradation or fatigue, etc. Sometimes it is possible to construct degradation models in which degradation is measured *without errors*, but there are also situations when the degradation data are obtained with measurement errors.

Dynamic Regression Models

Statistical inference from ALT is possible if failure time regression models relating failure time distribution with explanatory variables (covariates, stresses), influencing the reliability are well chosen. Statistical inference from failure time-degradation data with covariates needs even more complicated models relating failure time distribution not only with external but also with internal explanatory variables (degradation, wear) which explain the state of units before the failures. In the last case models for degradation process distribution are needed, too. Here we discuss only several most used failure time-degradation regression models.

Denote by T the random time-to-failure of a unit (or system). We say also that T is the time of *hard* or *traumatic*

failure. Let S be the survival function and λ be the hazard rate of T:

$$S(t) = P\{T > t\}, \quad \lambda(t) = \lim_{h \to 0} \frac{1}{h} P\{t < T \le t + h | T > t\}, \tag{1}$$

from where it follows that $S(\cdot)$ can be written as

$$S(t) = e^{-\Lambda(t)}, \quad \text{where} \quad \Lambda(t) = \int_0^t \lambda(s) ds$$

is the cumulative hazard function of T. Denote $F(\cdot) = 1 - S(\cdot)$ the cumulative distribution function of T. In Survival Analysis and Reliability the models are often formulated in terms of cumulative hazard and hazard rate functions. The most common shapes of hazard rates are monotone, $\cup$- shaped or $\cap$- shaped, see, for example, Meeker and Escobar (1998), Lawless (2003), Zacks (2004).

Let suppose that any explanatory variable is a deterministic time function

$$x = x(\cdot) = (x_1(\cdot), \ldots, x_m(\cdot))^T : [0, \infty[\to B \in R^m,$$

which is a vector of covariates itself or a realization of a stochastic process $X(\cdot)$, which is called also the *covariate process*, $X(\cdot) = (X_1(\cdot), \ldots, X_m(\cdot))^T$. We denote by E the set of all possible (admissible) covariates. We do not discuss here the questions of choice of X_i and m, but they are very important for the organization (design) of the experiments and for statistical inference. The covariates can be interpreted as the control, so we may consider models of aging in terms of the optimal control theory. We may say also that we consider statistical modeling with dynamic design or in dynamic environments.

In accordance with (1) the survival, the hazard rate, the cumulative hazard and the distribution functions given covariate x are:

$$\begin{aligned} S(t|x) &= P(T > t | x(s); 0 \le s \le t), \\ \lambda(t|x) &= -\frac{S'(t|x)}{S(t|x)}, \\ \Lambda(t|x) &= -\ln[S(t|x)], \\ F(t|x) &= P(T \le t | x(s); 0 \le s \le t) \\ &= 1 - S(t|x), \quad x \in E, \end{aligned} \tag{2}$$

from where one can see their dependence on the life-history up to time t. On any family E of admissible stresses, we may consider a class $\{S(\cdot|x), x \in E\}$ of survival functions which could be very rich. Failure is affected by time-dependent covariates (loads, stress, usage rate) which describe heterogeneous and dynamic operating conditions.

We say that the time $f(t|x)$ under the stress x_0 is *equivalent* to the time t under the stress x if the probability that a unit used under the stress x would survive till the moment t is *equal* to the probability that a unit used under the stress x_0 would survive till the moment $f(t|x)$:

$$\begin{aligned} S(t|x) = P\{T > t | x(s); 0 \le s \le t\} = P\{T > f(t|x) | x_0(s); \\ 0 \le s \le f(t|x)\} = S(f(t|x)|x_0). \end{aligned}$$

It implies that

$$f(t|x) = S^{-1}[S(t|x)|x_0], \quad x \in E. \tag{3}$$

Let x and y be two admissible stresses: $x, y \in E$. We say that a stress y is *accelerated* with respect to x, if $S(t|x) \ge S(t|y), \quad \forall\, t \ge 0$.

The accelerated failure time (AFT) model is more adapted for failure time regression analysis, see Meeker and Escobar (1998), Bagdonavičius and Nikulin (2002), Lawless (2003), Nelson (2004).

We say that AFT model holds on E if there exists a positive continuous function $r : E \to R^1$ such that for any $x \in E$ the survival and the cumulative hazard functions under a covariate realization x are given by formulas:

$$\begin{aligned} S(t|x) &= G\left(\int_0^t r[x(s)] ds\right) \quad \text{and} \\ \Lambda(t|x) &= \Lambda_0\left(\int_0^t r[x(s)] ds\right), \quad x \in E, \end{aligned} \tag{4}$$

respectively, where $G(t) = S(t|x_0)$, $\Lambda_0(t) = -\ln G(t)$, x_0 is a given (usual) stress, $x_0 \in E$. The function r changes locally the time scale. From the definition of $f(t|x)$ (cf. (3)) it follows that for the AFT model on E

$$f(t|x) = \int_0^t r[x(s)] ds, \quad \text{hence} \quad \frac{\partial f(t|x)}{\partial t} = r(x(t))$$
$$\text{at the continuity points of } r[x(\cdot)]. \tag{5}$$

Note that the model can be considered as parametric (r and G belong to parametric families) semiparametric (one of these functions is unknown, other belongs to a parametric family) or nonparametric (both are unknown).

Modeling of Degradation Process

In reliability there is considerable interest in modeling covariate processes $Z(t), t \ge 0$, with some properties, depending on the phenomena in consideration, which describe the *real* process of wear or the usage history up to time t, indicate the level of fatigue, degradation, and deterioration of a system, and may influence the rate of degradation, risk of failure, and reliability of the system. Statistical modeling of *observed degradation* processes can help to understand different real physical, chemical, medical, biological, physiological, or social degradation processes of aging. Information about *real degradation* processes help us to construct degradation models, which permit us to

predict the cumulative damage, and so on. According to the principal idea of degradation models, a *soft failure* is observed if the degradation process reaches a *critical threshold* z_0. A soft failure caused by degradation occurs when $Z(t), t \geq 0$, reaches the value z_0. The *moment* T^0 of soft failure is defined by relation

$$T^0 = \sup\{t : Z(t) < z_0\} = \inf\{t : Z(t) \geq z_0\}. \quad (6)$$

So it is reasonable to construct the so-called degradation models, based on some suppositions about the mathematical properties of the degradation process $Z(t), t \geq 0$, in accordance with observed *longitudinal* data in the experiment. Both considered methods may be combined to construct the so-called joint degradation models. For this, it is enough to define a failure of the system when its degradation (*internal wear*) reaches a critical value or a traumatic event occurs. The joint degradation models form the class of models with *competing risk*, since for any item we consider two competing causes of failure: degradation, reaching a *threshold* (soft failure), and occurrence of a *hard* or *traumatic failure*. Let T be the moment of traumatic failure of a unit. In the considered class of models, the *failure time* τ is the *minimum* of the moments of traumatic and soft failures,

$$\tau = min(T^0, T) = T^0 \wedge T.$$

These models are also called *degradation-threshold-shock models*. For such models the degradation process $Z(t)$ can be considered as an additional time-depending covariable, which describes the process of wear or the usage history up to time t. The degradation models with covariates are used to estimate reliability when the environment is dynamic (see Singpurwalla 1995; Bagdonavičius and Nikulin 2002). The covariates cannot be controlled by an experimenter in such a case. For example, the tire wear rate depends on the quality of roads, the temperature, and other factors.

In Meeker and Escobar (1998) the use of so-called path models is proposed for construction the degradation process. The authors describe many different applications and models for *accelerated degradation* and Arrhenius analysis for data involving a destructive testing. Meeker and Escobar (1998) used *convex and concave degradation models* to study the growth of fatigue cracks, the degradation of components in electronic circuits.

The most applied ▶stochastic processes describing degradation are general path models and time scaled stochastic processes with stationary and independent increments such as the gamma process, shock processes and Wiener process with a drift.

Example 1 *General degradation path model:* The *general degradation path model* was considered by Meeker and Escobar (1998), according to which

$$Z(t) = g(t, A, \theta) \quad (7)$$

where $A = (A_1, \ldots, A_r)$ is a finite dimensional random vector with positive components and the distribution function π of A, θ is a finite dimensional non-random parameter, and g is a specified continuously differentiable in t function, which increases from 0 to $+\infty$ when t increases from 0 to $+\infty$. The form of the function g is suggested by the form of individual degradation curves, obtained from the experiments. For example, $g(t, a) = t/a$ (linear degradation, $r = 1$) or $g(t, a) = (t/a_1)^{a_2}$ (convex or concave degradation, $r = 2$). It is also supposed that, for each $t > 0$, the degradation path $(g(s, a) \mid 0 < s \leq t)$ determines in the unique way the value a of the random vector A. Degradation in these models is modeled by the process $Z(t, A)$, where t is time and A is some possibly multidimensional random variable. The *linear degradation models* are used often to study the increase in a resistance measurement over time, and the *convex degradation models* and *concave degradation models* are used to study the growth of fatigue cracks.

In many different applications and models for *accelerated degradation* for data involving a destructive testing are described. Influence of covariates on degradation is also modeled in Bagdonavičius and Nikulin (2002), Doksum and Normand (1995), Padgett and Tomlinson (2005), etc., to estimate reliability when the environment is dynamic. The semiparametric analysis of several new degradation and failure time regression models without and with covariables is described in Bagdonavičius and Nikulin (2001, 2004), Yashin (2004), Bagdonavičius et al. (2007), and Zacks (2004). The degradation under the covariate x is modeled by

$$Z(t|x) = g(f(t, x, \beta), A), \quad m(t|x) = Eg(f(t, x, \beta), A).$$

About the methods of estimation one can see also in Meeker and Escobar (1998).

Example 2 *Time scaled gamma process:* $Z(t) = \sigma^2 \gamma(t)$, where $\gamma(t)$ is a process with independent increments such that for any fixed $t > 0$

$$\gamma(t) \sim G\big(1, \nu(t)\big), \quad \nu(t) = \frac{m(t)}{\sigma^2},$$

i.e., $\gamma(t)$ has the ▶gamma distribution with the density

$$p_{\gamma(t)}(x) = \frac{x^{\nu(t)-1}}{\Gamma\big(\nu(t)\big)} e^{-x}, \quad x \geq 0,$$

where $m(t)$ is an increasing function. Then

$$Z(t|x) = \sigma^2\, \gamma(f(t,x,\beta)).$$

The mean degradation and the covariances under the covariate x are

$$m(t|x) = E(Z(t|x)) = m\left(f(t,x,\beta)\right),$$
$$Cov(Z(s|x), Z(t|x)) = \sigma^2 m(f(s \wedge t, x, \beta)).$$

Bagdonavičius and Nikulin (2001) considered estimation from failure time-degradation data, Lawless (2003), considered estimation from degradation data.

Example 3 *Time scaled Wiener process with a drift:* $Z(t) = m(t) + \sigma W(m(t))$, where W denotes the standard Wiener motion, i.e., a process with independent increments such that $W(t) \sim N(0,t)$. Then

$$Z(t|x) = m(f(t,x,\beta)) + \sigma W(m(f(t,x,\beta))).$$

The mean degradation and the covariances under the covariate x are

$$m(t|x) = m\left(f(t,x,\beta)\right), \quad Cov(Z(s|x), Z(t|x)) = \sigma^2 m(f(s \wedge t, x, \beta)).$$

Doksum and Normand (1995), Lehmann (2001, 2004, 2005) Whitmore and Schenkelberg (1997) considered estimation from degradation data.

Example 4 *Shock processes:* Assume that degradation results from shocks, each of them leading to an increment of degradation. Let T_n, $(n \geq 1)$ be the time of the nth shock and X_n the nth increment of the degradation level. Denote by $N(t)$ the number of shocks in the interval $[0,t]$. Set $X_0 = 0$. The degradation process is given by

$$Z(t) = \sum_{n=1}^{\infty} 1\{T_n \leq t\} X_n = \sum_{n=0}^{N(t)} X_n.$$

Kahle and Wendt (2006) model T_n as the moments of transition of the doubly stochastic Poisson process, i.e., they suppose that the distribution of the number of shocks up to time t is given by

$$P\{N(t) = k\} = E\left\{\frac{(Y\eta(t))^k}{k!} \exp\{-Y\eta(t)\}\right\},$$

where $\eta(t)$ is a deterministic function and Y is a non-negative random variable with finite expectation. If Y is non-random, N is non-homogenous Poisson process, in particular, when $\eta(t) = \lambda t$, N is homogenous Poisson process. If $\eta(t) = t$, then N is a mixed Poisson process. Other models for η may be used, for example, $\eta(t) = t^{\alpha}$, $\alpha > 0$. The random variable Y is taken from some parametric class of distributions. Wendt and Kahle (2004), Kahle and Wendt (2006) considered parametric estimation from degradation data. Lehmann (2005, 2006) considered estimation from failure time-degradation data.

Example 5 *Degradation model with noise:* suppose that the lifetime of a unit is determined by the degradation process $Z(t)$ and the moment of its potential traumatic failure is T. Denote by T^0 the moment at which the degradation attains some critical value z_0. Then the moment of the unit's failure is $\tau = T^0 \wedge T$. To model the degradation-failure time process, we suppose that the real degradation process is modeled by the *general path model*:

$$Z_r(t) = g(t, A) \tag{8}$$

where $A = (A_1, \ldots, A_r)$ is a random positive vector, and g is continuously differentiable increasing in t function. As remarked earlier, the real degradation process $Z_r(t)$ is often not observed, and we have to measure (to estimate) it. In this case, the *observed degradation process* $Z(t)$ is different from the real degradation process $Z_r(t)$. We supposed that we have the *degradation model with noise* according to which the *observed degradation process* is

$$Z(t) = Z_r(t)U(t), \quad t > 0 \tag{9}$$

where $\ln U(t) = V(t) = \sigma W(c(t))$, W is the standard *Wiener process* independent of A, and c is a specified continuous increasing function, with $c(0) = 0$. For any $t > 0$ the *median* of $U(t)$ is 1.

Using this model it is easy to construct a *degradation model with noise*. An important class of models based on degradation processes was developed recently by Bagdonavičius et al. (2007). Wulfsohn and Tsiatis (1997) proposed considered the so-called *joint model* for survival and longitudinal data measured with error, given by

$$\lambda_T(t|A) = \lambda_0(t) e^{\beta(A_1 + A_2 t)} \tag{10}$$

where $A = (A_1, A_2)$ follows bivariate normal distribution. On the other hand Bagdonavičius and Nikulin (2004) proposed the model in terms of a conditional survival function of T given the real degradation process:

$$S_T(t|A) = P\{T > t | g(s,A), 0 \leq s \leq t\} = \exp\left\{-\int_0^t \lambda_0(s,\theta)\lambda(g(s,A))ds\right\} \tag{11}$$

where λ is the unknown intensity function, $\lambda_0(s,\theta)$ is from a parametric family of hazard functions. The distribution of A is not specified. This model states that the conditional hazard rate $\lambda_T(t|A)$ at the moment t given the degradation

$g(s, A), 0 \le s \le t$, has the multiplicative form as in the famous Cox model:

$$\lambda_T(t|A) = \lambda_0(s, \theta)\lambda(g(s, A)) \quad (12)$$

If for example, $\lambda_0(s, \theta) = (1 + t)^{\theta}$ or $\lambda_0(s, \theta) = e^{t\theta}$, then $\theta = 0$ corresponds to the case when the hazard rate at any moment t is a function of the degradation level at this moment. One can note that in the second model the function λ, characterizing the influence of degradation on the hazard rate, is nonparametric.

About the Author

For biography *see* the entry ►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements.

Cross References

►Accelerated Lifetime Testing
►Modeling Survival Data
►Parametric and Nonparametric Reliability Analysis
►Survival Data

References and Further Reading

Bagdonavičius V, Nikulin M (2001) Estimation in degradation models with explanatory variables. Lifetime Data Anal 7:85–103

Bagdonavičius V, Nikulin M (2002) Accelerated life models. Chapman & Hall/CRC, Boca Raton

Bagdonavičius V, Nikulin M (2004) Semiparametric analysis of degradation and failure time data with covariates. In: Nikulin M, Balakrishnan N, Mesbah M, Limnios N (eds) Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life. Birkhauser, Boston, pp 41–64

Bagdonavičius V, Bikelis A, Kazakevičius V, Nikulin M (2007) Non-parametric estimation from simultaneous renewal-failure-degradation data with competing risks. J Stat Plan Inference 137:2191–2207

Doksum KA, Normand SLT (1995) Gaussian models for degradation processes – part I: methods for the analysis of biomarker data. Lifetime Data Anal 1:131–144

Kahle W, Wendt H (2006) Statistical analysis of some parametric degradation models. In: Nikulin M, Commenges D, Huber C (eds) Probability, statistics and modelling in public health. Springer, New York, pp 266–279

Lawless JF (2003) Statistical models and methods for lifetime data. Wiley, New York

Lehmann A (2001) A Wiener process based model for failure and degradation data in dynamic environments. Drexdner Schriften zur Mathemat Stochastik 4:35–40

Lehmann A (2004) On degradation-failure models for repairable items. In: Nikulin M, Balakrishnan N, Mesbah M, Limnios N (eds) Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life. Birkhauser, Boston, pp 65–80

Lehmann A (2005) Joint models for degradation and failure time data. In: Proceedings of the internationaol workshop "Statistical modelling and inference in life Sciences", September 1–4, Potsdam, pp. 91–94

Lehmann A (2006) Degradation-threshold-shock models. In: Nikulin M, Commenges D, Huber C (eds) Probability, statistics and modelling in public health. Springer, New York, pp 286–298

Meeker W, Escobar L (1998) Statistical methods for reliability data. Wiley, New York

Nelson WB (2004) Accelerated testing: Statistical models, test plans, and data analysis, 2nd edn. Wiley, NY

Nikulin M, Limnios N, Balakrishnan N, Kahle W, Huber C (eds) (2010) Advances in degradation modeling, Birkhauser: Boston

Padgett WJ, Tomlinson MA (2005) Accelerated degradation models for failure based on geometric Brownian motion and gamma processes. Lifetime Data Anal 11:511–527

Singpurwalla N (1995) Survival in dynamic environments. Stat Sci 1:86–103

Wendt H, Kahle W (2004) On parametric estimation for a position-dependent marking of a doubly stochastic Poisson process. In: Nikulin M, Balakrishnan N, Mesbah M, Limnios N (eds) Parametric and semiparametric models with applications to reliability, survival analysis, and quality of life. Birkhauser, Boston, pp 473–486

Whitmore GA, Schenkelberg F (1997) Modelling accelerated degradation data using Wiener diffusion with a time scale transformation. Lifetime Data Anal 3:27–45

Wulfsohn M, Tsiatis A (1997) A joint model for survival and longitudinal data measured with error. Biometrics 53: 330–339

Yashin AI (2004) Semiparametric models in the studies of aging and longevity. In: Nikulin M, Balakrishnan N, Limnios N, Mesbah M (eds) Parametric and semiparametric models with applications for reliability, survival analysis, and quality of life. Birkhauser, Boston, pp 149–166

Zacks Sh (2004) Failure distributions associated with general compound renewal damage processes. In: Antonov V, Huber C, Nikulin M, Polischook V (eds) Longevity, aging and degradation models in reliability, public health, medicine and biology, vol 2. St. Petersburg State Polytechnical University, St. Petersburg, pp 336–344

Degrees of Freedom

Chong Ho Yu
Arizona State University, Tempe, AZ, USA

Many elementary statistics textbooks introduce the concept of degrees of freedom (df) in terms of the number scores that are "free to vary." However, this explanation cannot clearly show the purpose of df. There are many other approaches to present the concept of degrees of freedom. Two of the most meaningful ways are to illustrate df

in terms of sample size and dimensionality. Both represent the number of pieces of *useful information.*

DF in Terms of Sample Size

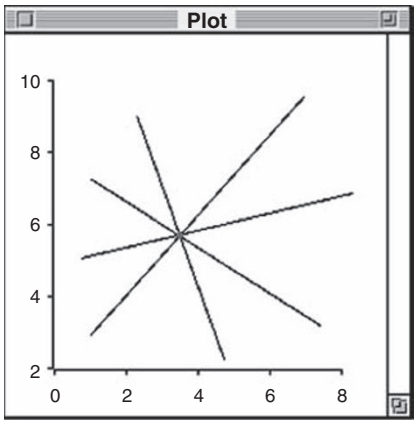
Toothaker (1986) explained *df* as the number of independent components minus the number of parameters estimated. This approach is based upon the definition provided by Walker (1940): the number of observations minus the number of necessary relations, which is obtainable from the observations ($df = n - r$). Although Good (1973) criticized that Walker's approach is not obvious in the meaning of necessary relations, the number of necessary relationships is indeed intuitive when there are just a few variables. "Necessary relationship" can be defined as the relationship between a dependent variable (Y) and each independent variable (X) in the research. Please note that this illustration is simplified for conceptual clarity. Although Walker regards the preceding equation as a universal rule, $df = n - r$ might not be applicable to all situations.

No Degree of Freedom and Effective Sample

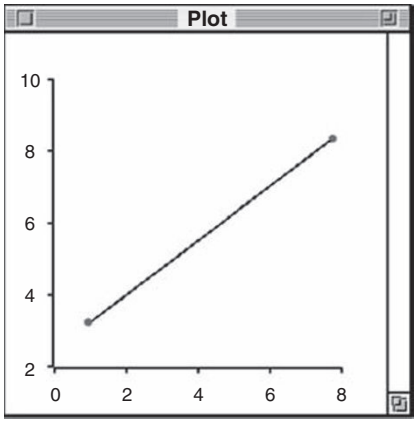
Figure 1 shows that there is one relationship under investigation ($r = 1$) when there are two variables. In the scatterplot there is only one datum point. The analyst cannot do any estimation of the regression line because the line can go in any direction, as shown in Fig. 1. In other words, there isn't any useful information. When the degree of freedom is zero ($df = n - r = 1 - 1 = 0$), it is impossible to affirm or reject the model. In this sense, the data have no "freedom" to vary. Bluntly stated, one subject is basically useless, and obviously, *df* defines the *effective* sample size (Eisenhauer 2008). The effective sample size is smaller than the actual sample size when *df* is taken into account.

Perfect Fitting and Overfitting

In order to plot a regression line, one must have at least two data points as indicated in Fig. 2. In this case, there is one degree of freedom for estimation ($n - 1 = 1$, where $n = 2$). When there are two data points only, one can always join them to be a straight regression line and get a perfect correlation ($r = 1.00$). This "perfect-fit" results from the lack of useful information. Since the data do not have much "freedom" to vary, no alternate models could be explored. In addition, when there are too many variables in a regression model, i.e., the number of parameters to be estimated is larger than the number of observations, this model is said to lacking degrees of freedom and thus is over-fit.



Degrees of Freedom. Fig. 1 No degree of freedom with one datum point



Degrees of Freedom. Fig. 2 Perfect fit with two data points

DF in Terms of Dimensions and Parameters

In this section, degrees of freedom are illustrated in terms of dimensionality and parameters. According to Good (1973), degrees of freedom can be expressed as

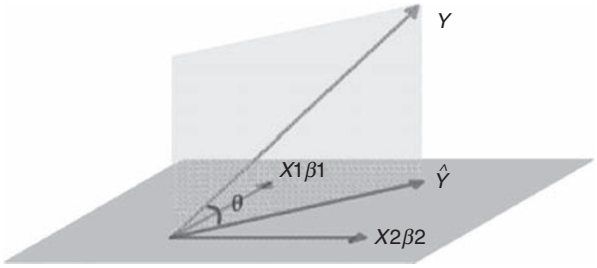
$$D(K) - D(H),$$

whereas

$D(K)$ = the dimensionality of a broader hypothesis, such as a full model in regression

$D(H)$ = the dimensionality of the null hypothesis, such as a restricted or null model.

In Fig. 3 vectors (variables) in hyperspace (Saville and Wood 1991; Wickens 1995) are used for illustrating a regression model. It is important to point out that the illustration is only a metaphor to make comprehension easier. Vectors do not behave literally as shown. In hyperspace, Vector Y represents the dimension of the outcome variable



Degrees of Freedom. Fig. 3 Vectors in hyperspace

whereas Vector $X1\beta1$ and $X2\beta2$ denote the dimensions of the predictors with two estimated parameters (βs). Vector $\hat{Y}$ is the dimension of Y expected and $\cos(\theta)$ equals R (relationship among $X1, X2$, and Y). In this example the intercept is ignored.

What is (are) the degree(s) of freedom when there is one variable (vector) in a regression model? First, we need to find out the number of parameter(s) in a one-predictor model. Since only one predictor is present, there is only one beta weight to be estimated. The answer is straightforward: There is one parameter to be estimated. How about a null model? In a null model, the number of parameters is set to zero. The expected Y score is equal to the mean of Y and there is no beta weight to be estimated. Based upon $df = D(K) - D(H)$, when there is only one predictor, the degree of freedom is just one $(1-0 = 1)$. It means that there is only one piece of useful information for estimation. In this case, the model is not well-supported. As you notice, a 2-predictor model $(df = 2 - 0 = 2)$ is better-supported than the 1-predictor model $(df = 1 - 0 = 1)$. When the number of *orthogonal* vectors increases, we have more pieces of independent information to predict Y and the model tends to be more stable. In short, the degree of freedom can be defined in the context of *dimensionality*, which conveys the amount of *useful information*. However, it is important to note that some regression methods, such as ridge regression (see ►Ridge and Surrogate Ridge Regressions), linear smoothers and ►smoothing splines are not based on ►least squares, and so *df* defined in terms of dimensionality is not applicable to these modeling.

Putting Both Together

The above illustrations (Walker's *df* and Good's *df*) compartmentalize *df* in terms of sample size and *df* in terms of dimensionality (variables). However, observations (n) and parameters (k), in the context of *df*, must be taken into consideration together. For instance, in regression, the *working definition* of degrees of freedom involves the information of both observations and dimensionality: $df = n - k - 1$ whereas n = sample size and k = the number of variables. Take the 3-observation and 2-variable case as an example. In this case, $df = 3 - 2 - 1 = 0$. Readers who are interested in *df* are referred to the online tutorial posted on http://www.creative-wisdom.com/pub/df/default.htm

About the Author

Chong Ho Yu is the Director of Testing, Evaluation, Assessment, and Research group at Applied Learning Technologies Institute, ASU.

Cross References

►Degrees of Freedom in Statistical Inference

References and Further Reading

Eisenhauer JG (2008) Degrees of freedom. Teach Stat 30(3):75–78
Good IJ (1973) What are degrees of freedom? Am Stat 27:227–228
Saville D, Wood GR (1991) Statistical methods: the geometric approach. Springer, New York
Toothaker LE, Miller L (1996) Introductory statistics for the behavioral sciences, 2nd edn. Brooks/Cole, Pacific Grove
Walker HW (1940) Degrees of freedom. J Educ Psychol 31:253–269
Wickens T (1995) The geometry of multivariate statistics. Lawrence Erlbaum, Hillsdale

Degrees of Freedom in Statistical Inference

Joseph G. Eisenhauer
Professor and Dean
University of Detroit Mercy, Detroit, MI, USA

The term *degrees of freedom* refers to the number of items that can be freely varied in calculating a statistic without violating any constraints. Such items typically include observations, categories of data, frequencies, or independent variables. Because the estimation of parameters imposes constraints on a data set, a degree of freedom is generally sacrificed for each parameter that must be estimated from sample data before the desired statistic can be calculated.

Consider first a simple case. A sample of size n initially has n degrees of freedom, in the sense that any or all of the observations could be freely discarded and replaced by others drawn from the population. However, once their sum has been calculated, only $n - 1$ observations are free to vary, the final one being determined by default. Equivalently, once the sample mean $\bar{x}$ has been calculated, n deviations from the mean can be found but they must sum to zero, so only $n - 1$ deviations are free to vary. In letters to W.S. Gosset around 1912, Sir Ronald Fisher first

showed that if the deviations are squared and averaged over $n-1$ rather than n, the resulting sample variance s^2 is an unbiased estimator of the population variance σ^2.

This basic notion recurs throughout inferential statistics. With a normal distribution, testing a hypothesis regarding σ^2 proceeds by way of a chi-square statistic, $\chi^2 = (n-1)s^2/\sigma^2$. Notice that $E(s^2) = \sigma^2$ implies $E(\chi^2) = n-1$. As this suggests, there exists an entire family of chi-square distributions, the mean of any one of which is its degrees of freedom. Moreover, because χ^2 takes only nonnegative values, the distribution is positively skewed when its mean is near zero, and it becomes increasingly symmetric as the degrees of freedom rise. Consequently, the probability that χ^2 exceeds a given value increases with the degrees of freedom. Put differently, the critical value for a chi-square test at a given level of significance increases with its degrees of freedom. Student's t statistic for testing a normal population mean likewise has the $n-1$ degrees of freedom from the sample variance. As the degrees of freedom rise, the shape of the t distribution approaches that of the standard normal; its variance decreases, and thus, the probability that t exceeds a given value diminishes. To compare the means from two normal populations with $\sigma_1^2 = \sigma_2^2$, the sample variances are pooled and the resulting t statistic then has $(n_1-1)+(n_2-1)$ degrees of freedom. For normal populations with $\sigma_1^2 \neq \sigma_2^2$, Satterthwaite (1946) proposed the approximation $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{a_1 + a_2}$ with $(a_1+a_2)^2/\left[\left(a_1^2/\nu_1\right) + \left(a_2^2/\nu_2\right)\right]$ degrees of freedom, where $a_i = s_i^2/n_i$ and $\nu_i = n_i - 1$; in this case, the degrees of freedom will generally not be an integer. To test for equal variances, the F ratio (so named for Fisher), $F = s_1^2/s_2^2$, is used with $n_1 - 1$ numerator degrees and $n_2 - 1$ denominator degrees of freedom. If F is significantly larger or smaller than unity, the null hypothesis of equal variances can be rejected. Of course, because the designation of populations 1 and 2 is arbitrary, we could just as well invert the F ratio and reverse its degrees of freedom. Therefore, commonly available tables of the F distribution report only the right-hand critical value for an F test or confidence interval; to obtain the left-hand critical value, the numerator and denominator degrees of freedom are reversed, and the resulting table value is inverted.

The decomposition of a sample variance into explained and unexplained components via ▶analysis of variance (ANOVA) also relies on degrees of freedom. The sum of squared deviations from the sample mean, or sum of squares total (SST), has $n-1$ degrees of freedom as explained above. Similarly, with k treatment categories, the sum of squares due to treatment ($SSTR$) has $k-1$ degrees, and the residual, the sum of squares due to error (SSE), has $(n-1)-(k-1) = (n-k)$ degrees of freedom. Dividing the sums of squares by their respective degrees of freedom, the mean square due to treatment is $MSTR = SSTR/(k-1)$ and the mean square due to error is $MSE = SSE/(n-k)$. Because each of these is a type of variance, we compare them using the test statistic $F = MSTR/MSE$ with $k-1$ numerator degrees and $n-k$ denominator degrees of freedom. A related application involves measuring the strength, or explanatory power, of a multiple linear regression, where k is reinterpreted as the number of regression coefficients (including the intercept) and $SSTR$ denotes the sum of squares due to regression. Increasing the number of independent variables $(k-1)$ tends to alter the composition of SST by raising $SSTR$ and reducing SSE, though the increase in explained variation may be somewhat spurious; the coefficient of determination, $R^2 = 1 - (SSE/SST)$, can thereby become inflated. The customary correction is to divide SSE and SST by their respective degrees of freedom, yielding the adjusted coefficient of determination $\bar{R}^2 = 1 - \{[SSE/(n-k)]/[SST/(n-1)]\}$ as the measure of explanatory power.

Algebraically, a system of j simultaneous equations in k unknowns has $k-j$ degrees of freedom. Consider, for example, a goodness-of-fit test to determine whether the variable x, which takes the values $0, 1, 2 \ldots (k-1)$, follows a truncated Poisson distribution with a given mean. There are k expected frequencies, but since they must sum to the sample size n, one can be determined from the others. Thus, only $k-1$ are free to vary, i.e., there are $k-1$ degrees of freedom. Moreover, if it is necessary to first estimate the population mean from the sample data in order to obtain the Poisson distribution, an extra constraint is imposed, leaving $k-2$ degrees of freedom; see Eisenhauer (2008) for a numerical illustration. By the same reasoning, a goodness-of-fit test for a normal distribution segmented into k intervals has $k-1$ degrees of freedom if the mean and standard deviation of the population are known; but if both parameters need to be estimated, two more degrees are sacrificed and $k-3$ degrees of freedom remain.

Fisher (1922) originally coined the phrase *degrees of freedom* in correcting Karl Pearson's test of independence. A contingency table consisting of r rows and c columns into which frequencies are entered has rc cells, but given row and column totals, the entries for the final row and final column – totaling $r+c-1$ cells – can be determined from the others. Thus, there are only $rc-(r+c-1) = (r-1)(c-1)$ degrees of freedom for the chi-square test statistic.

About the Author

Dr. Joseph G. Eisenhauer is Dean of the College of Business Administration at the University of Detroit Mercy. His prior positions include Professor and Chair of Economics

at Wright State University, and Professor of Economics and Finance at Canisius College. He has been a postdoctoral fellow at the University of Pennsylvania, a Visiting Scholar at the Catholic University of America, and a Visiting Professor at the Universita di Roma, La Sapienza, Italy. He is a past President and Distinguished Fellow of the New York State Economics Association, and has published more than 80 articles and reviews in scholarly journals. He recently won the C. Oswald George Prize (2008) for the best article, "Degrees of Freedom" in *Teaching Statistics*.

Cross References

►Degrees of Freedom

References and Further Reading

Eisenhauer JG (2008) Degrees of freedom. Teach Stat 30(3):75–78

Fisher RA (1922) On the interpretation of χ^2 from contingency tables and the calculation of P. J Roy Stat Soc 85(1):87–94

Satterthwaite FE (1946) An approximate distribution of estimates of variance components. Biometrics Bull 2(6):110–114

Demographic Analysis: A Stochastic Approach

Krishnan Namboodiri
Professor Emeritus
Ohio State University, Columbus, OH, USA

Introduction

Demographers study population dynamics: changes in population size and structure resulting from fertility (childbearing performance), mortality (deaths), and spatial and social mobility. The focus may be the world population or a part of it, such as the residents of a country or the patients of a hospital. Giving birth, dying, shifting usual place of residence, and trait changes (e.g., getting married) are called events. Each event involves transition from one "state" to another (e.g., from never-married state to married state). A person is said to be "at risk" or "exposed to the risk" of experiencing an event, if for that person the probability of that experience is greater than zero. The traits influencing the probability of experiencing an event are called the risk factors of that event (e.g., high blood pressure, in the case of ischemic heart disease).

Demographic data are based on censuses (see ►Census), sample surveys, and information reported to offices set up for continuously recording demographic events. Some observational studies can be viewed as random experiments. For an individual selected at random from a population at time t, the value of the variable $y_{t+\theta}$, denoting whether that individual will be alive as of a subsequent moment $t + \theta$ is unpredictable. This unpredictability of the value of $y_{t+\theta}$ qualifies the observational study involving observations at times t and $t + \theta$ to be considered as a random experiment and $y_{t+\theta}$ as a random variable, defined by a set of possible values it may take (e.g., 1 if alive at time $t + \theta$, and 0, otherwise) and an associated probability function (Kendall and Buckland 1971). The interval between a fixed date and a subsequent event is a random variable, in the above-mentioned sense.

The term rate is used in ►demography for the number of events (e.g., deaths) expressed per unit of some other quantity, such as person-years at risk (often expressed per 1,000). For example, the *crude death rate* (annual number of deaths expressed per 1,000 mid-year population) in Japan in 2007 was 9. The mid-year population in such calculations is an approximation to the sum of the person-years lived by the members of the population involved during the specified year. Death rates calculated for subpopulations, homogeneous, to some degree, with respect to one or more relevant risk factors are called specific death rates. Examples are age-specific and age-sex specific death rates.

A *life-table* (see ►Life Table) shows the life-and-death history of a group of persons, called a *cohort*, born at the same time (e.g., a year), as the cohort members survive to successive ages or die in the intervals, subject to the mortality conditions portrayed in a schedule of age-specific death rates. An account of the origin, nature, and uses of life tables is available in P. R. Cox (1975). Life tables have become powerful tools for the analysis of non-renewable (non-repeatable) processes. If a repeatable process, such as giving births, can be split into its non-renewable components (e.g., births by birth order) then each component can be studied, using the life-table method. The term: *survival analysis* is applied to the study of non-renewable processes, in general. Associated with the survival rate is the hazard rate, representing the instantaneous rate of failure (to survive). Hazard rate corresponds to the instantaneous death rate or force of mortality, as used in connection with life tables.

Macro-Level Focus

A great deal of demographic research is linked directly or indirectly to model construction and validation, viewing observations as outcomes of random experiments. Birth-and-death process (see Kendall 1948; Bhat 1984), is a continuous time, integer valued, counting process, in which

population size at time t, remains constant, increases by one unit (a birth), or decreases by 1 unit (a death), over the period: t to $t+\Delta t$. Time-trend in population size is studied using branching processes, in a simple version of which, each member of each generation produces offspring, in accordance with a fixed probability law common to all members (see, e.g., Grimmett and Stirzaker 1992 for a discussion of simple as well as complex models of branching processes). The logistic process for population growth of the "birth-and-death" type views the instantaneous rates of birth and death for each individual alive at a given moment as linear functions of population size (see Brillinger 1981; Goel and Dyn 1979; Mollison 1995). For compositional analysis, one may apply an appropriate log-ratio transformation to the composition of interest, and treat the resulting values as a random vector from a multivariate normal distribution (see Aitchison 1986; Namboodiri 1991).

Using the component model (see Keyfitz 1971) of population projection, one obtains internally consistent estimates of the size and age-sex composition of populations as of future years by combining hypothesized patterns of change in fertility, mortality, and migration. On the basis of such projections, issues such as the following can be examined: (1) Reduction in population growth rate resulting from the elimination of deaths due to a specific cause, e.g., heart disease; (2) Relative impact on the age-composition, in the long-run, of different combinations of population-change components (e.g., fertility and mortality); and (3) tendency of populations to "forget" the past features (e.g., age composition) if the components of population dynamics were to continue to operate without change over a sufficiently long time.

To estimate and communicate the uncertainty of population projections, the practitioners have been combining "high," "medium," and "low" scenarios for the components of population change in various ways (e.g., "high" fertility combined with "low" mortality to produce "high" population projection) to show different possibilities regarding future population size and composition. Since such demonstrations of uncertainties have no probabilistic interpretations, Lee and Tuljapurkar, among others, have pioneered efforts to develop and popularize the use of stochastic population projections (see Lee 2004). Lee and Tuljapurkar (1994) demonstrated, for example, how to forecast births and deaths, from time-series analyses of fertility and mortality data for the United States, and then combine the results with deterministically estimated migration to forecast population size and composition. They used in the demonstration, products of stochastic matrices.

Comparison of the simple non-stochastic trend model: $y_t = \beta_0 + \beta_1(t) + e_t$, with the stochastic (random-walk with a drift) model: $y_t = \alpha_0 + y_{t-1} + e_t$, where e_t's are $NID(0, \sigma_e)$ for all t, shows that even when the error terms have equal variance $\left(\sigma_e^2\right)$ in the two models, the prediction intervals for the latter are wider than those of the former: For a forecast horizon H, the variance of the forecast error (the departure of the forecast from the actual) in the case of $y_t = \beta_0 + \beta_1(t) + e_t$ is σ_e^2, while the corresponding quantity is $H\sigma_e^2$, in the case of $y_t = \alpha_0 + y_{t-1} + e_t$.

Micro-Level Processes

At the micro level, one focuses on events (such as giving birth to the first child, dying, recovering from illness, and so on) experienced by individuals. In event histories, points of time at which transitions occur (e.g., from not in labor force to employed) are represented by a sequence of non-negative random variables: $(T_1, T_2, \ldots)$, and the differences: $V_k = T_k - T_{k-1}, k = 2, 3, \ldots$, are commonly referred to as waiting times. Comprehensive discussions of waiting times are available, for example, in: Cleves et al. (2004); Collett (2003); Elandt-Johnson and Johnson (1980/1999); and Lawless (1982/2003).

D. R. Cox (1972) introduced, what has come to be known as, the proportional hazards model: $\lambda(t) = \lambda_0(t)\psi(z_1, z_2, \ldots, z_k)$, where "$t$" represents time, and the multiplier, $\psi(z_1, z_2, \ldots, z_k)$, is positive and time-independent. A special form of the model is: $\lambda(t) = \lambda_0(t)\exp(\Sigma\beta_j z_j)$, in which $\{\beta_j\}$ are unknown regression coefficients.

An important feature of waiting time is heterogeneity (variation among individuals) in the hazard rate (see Sheps and Menken 1973; Vaupel et al. 1979; Heckman and Singer 1982). Heterogeneity is incorporated often as a multiplier in the Cox proportional hazards model. For example, the hazard function for the ith individual may be specified as: $\lambda_0(t)v_i\exp(\Sigma_j\beta_j Z_{ij})$, representing an individual-specific, unobserved heterogeneity factor by v_i. Vaupel et al. (1979) called such models: "frailty" models (see ▶Frailty Model).

Heckman and Singer (1982) suggested the specification of the unobserved heterogeneity factor in $\lambda(t) = \lambda_0(t)v_i\exp(\Sigma_j\beta_j Z_{ij})$, as a K-category discrete random variable. Thus the ith individual is presumed to belong to one of K groups. The value of K is determined empirically so as to maximize the likelihood of the sample on hand, under a specified (e.g., the exponential or Weibull) form for $\lambda_0(t)$. In the presence of heterogeneity, inference becomes sensitive to the form assumed for the hazard function (see, e.g., Trussell and Richards 1985).

As Sheps and Perin (1963) and Menken (1975), among others, have pointed out, simplified models, unrealistic though they may be, have proved useful in gaining insights such as that a highly effective contraceptive used by a rather small proportion of a population reduces birth rates more

than does a less effective contraceptive used by a large proportion of the population.

Some fertility researchers have been modeling parts rather than the whole of the reproductive process. The components of birth intervals have been examined, with emphasis on the physiological and behavioral determinants of fertility (see Leridon 1977). Another focus has been abortions, induced and spontaneous (see Abramson 1973; Potter et al. 1975; Michels and Willett 1996). Fecundability investigations have been yet another focus (see Menken 1975; Wood et al. 1994). Menken (1975) alerts researchers to the impossibility of reliably estimating fecundability from survey data. The North Carolina Fertility Study referred to in Dunson and Zhou (2000) is of interest in this connection: In that study couples were followed up from the time they discontinued birth control in order to attempt pregnancy. The enrolled couples provided base-line data and then information regarding ovulation in each menstrual cycle, day-by-day reports on intercourse, first morning urine samples, and the like. Dunson and Zhou present a Bayesian Model and Wood et al. (1996) present a multistate model for the analysis of fecundability and sterility.

To deal with problems too complex to be addressed using analytic models, researchers have frequently been adopting the simulation strategy, involving computer-based sampling and analysis at the disaggregated (e.g., individual) level. See, for example, the study of (1) kinship-resources for the elderly by Murphy (2004); and Wachter (1997); (2) female family-headship by Moffit and Rendall (1995); (3) AIDs and the elderly by Wachter et al. (2002); and (4) the impact of heterogeneity on the dynamics of mortality by Vaupel and Yashin (1985); and Vaupel et al. (1979). Questions such as the following arise: Is it possible to reproduce by simulation the world-population dynamics, detailing the changes in the demographic-economic-spatial-social DESS) complex, over the period, say: 1900–2000? Obviously, in order to accomplish such a feat, one has to have a detailed causal model of the observed changes to be simulated. As of now no satisfactory model of that kind is available. Thinking along such lines demographers might begin to view micro-simulation as a challenge and an opportunity to delve into the details of population dynamics.

About the Author

Dr. Krishnan Namboodiri was Robert Lazarus Professor of Population Studies at the Ohio State University, Columbus, Ohio, USA, (1984–2000) and has been Professor Emeritus at the same institution since 2000. Before joining the Ohio State University, he was Assistant Professor, Associate Professor, Professor, and Chairman, Department of Sociology, University of North Carolina at Chapel Hill, USA, (1966–1984); Reader in Demography, University of Kerala, India, (1963–1966). Dr. Namboodiri was Editor of Demography (1976–1979), and Associate Editor of of a number of professional journals such as Mathematical Population Studies (1985–1989). He has authored or co-authored over 80 publications including 12 books. He is a Fellow of the American Statistical Association, and is a recipient of honors such as Lifetime Achievement Award from Kerala University, and has been consultant from time to time to Ford Foundation, World Bank, United Nations, and other organizations.

Cross References

►Demography
►Life Table
►Population Projections

References and Further Reading

Abramson FD (1973) High foetal mortality and birth intervals. Popul Stud 27:235–242
Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London
Bhat UN (1984) Elements of applied stochastic processes, 2nd edn. Wiley, New York
Brillinger DR (1981) Some aspects of modern population mathematics. Can J Stat 9:173–194
Cleves M, Gould WG, Gutierrez RG (2004) An introduction to survival analysis using stata, Revised edn. Stata Press, College Station
Collett D (2003) Modeling survival data in medical research, 2nd edn. Chapman and Hall/CRC, Boca Raton
Cox DR (1972) Regression models and life tables (with discussion). J Roy Stat Soc B 34:187–202
Cox PR (1975) Population trends, vols I–II. Her Majesty's Stationary Office, London
Dunson DB, Zhou H (2000) A Bayesian model for fecundability and sterility. J Am Stat Assoc 95(452):1054–1062
Elandt-Johnson R, Johnson N (1980/1999) Survival models and data analysis. Wiley, New York
Goel NS, Dyn NR (1979) Stochastic models in biology. Academic Press, New York
Grimmett GR, Stirzaker DR (1992) Probability and random processes, 2nd edn. Clarendon Press, Oxford
Heckman JJ, Singer B (1982) Population heterogeneity in demographic models. In: Land KC, Rogers A (eds) Multidimensional mathematical demography. Academic, New York, pp 567–599
Kendall DG (1948) A generalized birth and death process. Ann Math Stat 19:1–15
Kendall MG, Buckland WR (1971) Dictionary of statistical terms, 3rd edn. Hafner, New York
Keytz N (1971) Models. Demography 8:329–352
Lawless JF (1982/2003) Statistical models and methods for lifetime data, 2nd edn. Wiley, New York
Lee RD (2004) Quantifying our ignorance: stochastic forecasts of population and public budgets. In: Waite LJ (ed) Aging, health, and public policy: demographic and economic perspectives. A special supplement to vol 30 of population and development review, pp 153–175

Lee RD, Tuljapurkar S (1994) Stochastic population projections for the United States beyond high, medium, and low. J Am Stat Assoc 89(438):1175–1189
Leridon H (1977) Human fertility: the basic components (trans: Helzner JF). University of Chicago Press, Chicago
Menken J (1975) Biometric models of fertility. Soc Forces 54:52–65
Michels KB, Willett WC (1996) Does induced or spontaneous abortion affect the risk of cancer? Epidemiology 7:521–528
Moffit RA, Rendall MS (1995) Cohort trends in the lifetime distribution of family headship in the United States, 1968–1985. Demography 32:407–424
Mollison D (ed) (1995) Epidemic models: their structure and relation to data. Cambridge University Press, London
Murphy M (2004) Tracing very long-term kinship networks Using SOCSIM. Demogr Res 10:171–196
Namboodiri K (1991) Demographic analysis: a stochastic approach. Academic, San Diego/New York
Potter RG, Ford K, Boots B (1975) Competition between spontaneous and induced abortions. Demography 12:129–141
Sheps MC, Menken J (1973) Mathematical models of conception and birth. University of Chicago Press, Chicago
Sheps MC, Perin EB (1963) Changes in birth rates as a function of contraceptive effectiveness: some applications of a stochastic model. Am J Public Health 53:1031–1046
Trussell TJ, Richards T (1985) Correcting for unobserved heterogeneity in hazard models: an application of the Heckman-Singer model for demographic data. In: Tuma NB (ed) Sociological methodology. Jossey-Basspp, San Francisco, pp 242–276
Vaupel JW, Manton KG, Stallard E (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16:439–454
Vaupel JW, Yashin AJ (1985) Heterogeneity ruses: some surprising effects of selection in population dynamics. Am Stat 39:176–185
Wachter KW (1997) Kinship resources for the elderly. Philos T Roy Soc B 352:1811–1817
Wachter KW, Knodel JE, Vanlandingham M (2002) AIDs and the elderly of Thailand: projecting familial impacts. Demography 39:25–41
Wood JW, Holman DJ, Yashin AI, Peterson RJ, Weinstein M, Chang MC (1994) A multistate model of fecundability and sterility. Demography 31:403–426

Demography

Jan M. Hoem
Professor, Director Emeritus
Max Planck Institute for Demographic Research, Rostock, Germany

The Character of the Field of Demography

The Topics of Investigation

Demography is the statistical study of human populations, including their size, composition, and geographical distribution, and of the processes of change in these elements. Demographers focus on childbearing (fertility), death (mortality), and geographical moves (migration), but they also cover other processes, like the formation and dissolution of (marital and nonmarital) unions, the transition out of the parental home, and other transitions relevant to population structure and population trends. As a field of inquiry, Demography has roots going back to John Graunt's famous study of the bills of mortality (1662) and to T. R. Malthus's essay on the principle of population (1798). For an account of how the endeavors of demographers have developed into a discipline, see Hodgson (2001), Szreter (2001), and Caldwell, 2003). For more extensive accounts about the field of demography, see Rosenzweig and Stark (1997) and Demeny and McNicoll (2003).

The Discipline

Like all academic disciplines, demographers have formed national, regional, and international societies, such as the Population Association of America, the European Association for Population Studies, the International Union for the Scientific Study of Population, and many others. Its oldest professional journals are publishing their 65th volumes or so in 2009 (*Population Studies, Population, Genus*), and there are a large number of younger journals (*Demography*, the *Population and Development Review*, the *European Journal of Population*, *Demographic Research, Journal of Population Research, Mathematical Population Studies, Journal of Population Economics* and so on), some of which are recent start-ups (like the *Romanian Journal of Population Studies*).

Demography courses are given by some universities, often as a part of studies in other disciplines, such as geography, sociology, or statistics. More extensive teaching is organized through demographic research centers like INED (Institut National des Études Démographiques in Paris), NIDI (the Netherlands Interdisciplinary Demographic Institute in The Hague), and MPIDR (the Max Planck Institute for Demographic Research in Rostock), sometimes as a joint venture of demographic centers and university departments across Europe (like the European Doctoral School in Demography). Teaching, research, and instrumental advice are also offered by demographic centers established by the United Nations and others (e.g., the Population Council). North America has its own demographic centers. Ideally, teaching programs reflect the multidisciplinarity of demography (Caselli 2002).

Demographic Data

Since the beginning of public data collection, demographers have made use of several sources of official statistics. For instance, status data are collected in decennial

censuses, typically to provide information about population composition and about the size of local populations or other groups at census time. Vital statistics (i.e., data on births, deaths, migration, marriages, etc.) are published annually and normally provide aggregate data for an entire population or for large population units. Individual-level (micro) data are mostly collected in special sample surveys organized at the national (or sub-national) level or designed to provide internationally comparable individual-level data (the Fertility and Family Surveys, the Generations and Gender Surveys, the Demographic and Health Surveys, and so on). In the Nordic countries and in a few other countries, individual-level demographic data are organized in a continuous-register system and made available for research. Several other countries or organizations have started to make registers of specific events available for research (such as a birth register, pension registers, migration registers). Recently international databases have been established for scientific use, such as the Human Mortality Database and the Human Fertility Database, both at MPIDR, or the IPUMS-data (International Public Use Microdata Series at the Minnesota Population Center). For further insights about demographic data regimes, see Haines (2001).

Population Forecasts

Producing population forecasts is a practical activity that many demographers are engaged in. Colleagues in other disciplines often resort to population forecasts produced by demographers.

Demography at the Crossroad of many Disciplines

As a field of endeavor, demography is highly interdisciplinary. Today we would characterize as demography much of the early activity of academic statisticians, and there is a residue of common interest down to the present time. Historically, demography has also been allied particularly closely with actuarial mathematics, epidemiology, sociology, economics, geography, and history. More recently it has developed links with biology, political science, and anthropology. The overlap between scientific fields is often so great that individual scientists may find it difficult to pledge their allegiance solely to demography or to a neighboring discipline; they often see themselves as members of both. Even what initially looks like a purely demographic theory may really need an underpinning in other disciplines. One case in point is an understanding of the role that a reduction of mortality at young ages has in initiating a decline in fertility. To explain why fertility declines, why there is a link between mortality decline and fertility decline, and why the pattern of the latter differs between societies, it is natural to resort to economics, health science, and perhaps anthropology.

The following examples provide selected glimpses of some overlaps between disciplines: There is constantly a mutual enrichment between demographic studies of family dynamics and sociological theory about the family. A deep investigation of policy effects in fertility trends is unthinkable without the insight and knowledge of political scientists, and conversely the latter can benefit from demographic statisticians' understanding of measurement issues (Frejka et al. 2008, Overview Chapter 8). Studies of mortality differentials gain from a sociological understanding of class differences. Demographic items are central to some recent theories about violent conflicts internal to populations. Demographers use elements from spatial theory in their studies of childbearing patterns among immigrants. Patterns of union formation contain important signals about the integration of minority groups, including immigrants (Basu and Aaby 1998), signals that should be useful to anthropologists. On the other hand, the latter are not always happy about the lack of attention paid to anthropological insight in classical demographic explanations of fertility trends in developing countries (Greenhalgh 1996).

Demographic Methodology

Stocks and Flows; Timing and Quantum

Demography has a distinctive emphasis on stocks and flows. Conceptually, a population is subdivided into a number of subgroups between which individuals move. The subgroups may represent life statuses, like "married," "divorced," "childless," "at parity 1" (= mothers/fathers with one child), and individuals are seen as constantly exposed to the risk of transition from one subgroup (or status) into another. Deaths are counted among the flows, and a ▶life table can be seen as a tabular representation of attrition from the stock of the living at the various ages. The underlying risks determine the timing of the flows, and the percentage that ultimately participates in a flow is called its quantum.

Data Cleaning and Careful Description

Cleaning the data of incorrect and inaccurate records is part of a demographer's preparation for analysis (Booth 2001). Demographers have always put great store on bringing out the facts, and this continues to be important in demographic investigations, while deeper explanation is often left to members of other disciplines despite any aspiration demographers may have toward the use (and development) of substantive theory. In this vein, demographers have a long tradition of careful description of empirical

regularities, mostly based on macro-data, but lately also covering patterns of individual-level behavior.

New Statistical Methods

To analyze demographic (vital) processes as dependent "variables," demographers have developed or adopted a number of statistical methods also used in other disciplines, notably public health and epidemiology. One can get an overview of the range of such methods as seen by demographers by consulting the various entries under 'demographic techniques' in Smelser and Baltes (2001). Entries under 'population dynamics' in the same source cover stable population theory, which is much used in demography. It is essentially based on extensions of branching-process theory.

The advent of event-history techniques has recently induced a change in the way many demographers approach their investigations. Demographers have increasingly turned to individual-level analyses of multi-process models and to multi-level modeling. The latter has opened the way for the inclusion of population-level features as determinants in investigations of individual behavior. It has also become possible to incorporate a feature like unobserved heterogeneity and thus to pick up selectivity in demographic behavior. These new possibilities have strengthened the tendency to use probabilistic thinking in demographic investigations and thus to make the field more recognizable to mathematical statisticians.

Transition Rates of Type 1 and Type 2

To give an impression of how demographic methodology fits into the world of statistical theory but have a flavor of its own, let us first note that just like in any application of event-history analysis in a transition model with a piecewise-constant hazard specification, a count of event occurrences $\{D_j\}$ and exposures $\{R_j\}$ (time-units of exposure) for the various subgroups $\{j\}$ of a population can be seen as a statistically sufficient set of statistics for the model risk parameters $\{\mu_j\}$. In such a situation, the occurrence/exposure rate $\hat{\mu}_j = D_j/R_j$ is a maximum-likelihood estimator of the parameter μ_j, and demographers would call it a *rate of the first kind*. Suppose, for instance, that the subscript j stands for a combination of (discrete) age x attained and civil status c, where $c = 0$ stands for "never married," say. Then $\mu_{x,0}$ can be the "risk" of entering first marriage at age x (for someone who has never been married before that age) and the first-marriage rate of the first kind will be $\hat{\mu}_{x,0} = D_{x,0}/R_{x,0}$. When the occurrences can be fully sub-specified but the population cannot be subdivided by marital status but only by age attained, say, a first-marriage rate $\mu^*_x = D_{x,0}/R_x$ of *the second kind* can still be computed for a given birth cohort, where $R_x = \sum_c R_{x,c}$ is an aggregate over all marital statuses c. (Note that μ^*_x is not an occurrence/exposure rate, because the denominator R_x includes the person-years $\sum_{c\neq 0} R_{x,c}$ also of individuals who are not under risk of first marriage in addition to the exposures $R_{x,0}$ for those who *are* under the risk of first marriage.) Demographers put rates of the second kind to uses of their own, but to a statistician, the main advantage of such rates is probably that in *cohort* data they can be converted into estimates rates of the first kind by means of the following formula:

$$\hat{\mu}_{x,0} = \mu^*_x \Big/ \left(1 - \sum_{y<x} \mu^*_y\right).$$

For a proof, see the argument by Calot (2001) leading up to his formula (3). (Note that all quantities in this formula refer to the same birth cohort. No corresponding conversion formula seems to exist for period data.)

About the Author

Dr. Jan Hoem is a Past President of the Scandinavian Demographic Society (1973–1974). He was Professor of Insurance Mathematics at Copenhagen University (1974–1981), Professor of Statistical Demography, Stockholm (1981–1999), where he is Guest Professor since 2009. He was Director of the Max Planck Institute for Demographic Research (1999–2007), and is Director Emeritus since 2007. Dr. Hoem was Section Editor for Demography of the International Encyclopedia of the Social and Behavioral Sciences (1998–2001). He was Editor of the electronic journal *Demographic Research* (1999–2006). Professor Hoem is Laureate of the International Union for the Scientific Study of Population (2006). He has authored and co-authored well over a hundred publications. He has initiated the use of Markov chains in life insurance mathematics, and has been instrumental in the development of event-history analysis in demography.

Cross References

►Census
►Demographic Analysis: A Stochastic Approach
►Life Table
►Population Projections
►Sociology, Statistics in

References and Further Reading

Basu A, Aaby P (1998) The methods and the uses of anthropological demography. Clarendon Press, Oxford

Booth H (2001) Demographic techniques: data adjustment and correction. In: Smelser N, Baltes P (eds) International encyclopedia

of the social and behavioral sciences, vol 5. Pergamon, Oxford, pp 3452–3456
Caldwell JC (2003) Demography, history of. In: Demeny P, McNicoll G (eds) Encyclopedia of population, vol 1. Macmillan Reference USA/Thomson Gale, New York, pp 216–221
Calot G (2001) Demographic techniques: rates of the first and second kind. In: Smelser N, Baltes P (eds) International encyclopedia of the social and behavioral sciences, vol 5. Pergamon, Oxford, pp 3480–3483
Caselli G (2002) Teaching demography in the early 21st century. Genus 58(special issue):3–4
Demeny P, McNicoll G (eds) (2003) Encyclopedia of population. Macmillan Reference USA/Thomson Gale, New York
Frejka T, Sobotka T, Hoem JM, Toulemon L (2008) Childbearing trends and policies in Europe. Demogr Res 19, Art. 1–29
Greenhalgh S (1996) The social construction of population science: an intellectual, institutional, and political history of twentieth century demography. Soc Comp Study Soc Hist 38:26–66
Haines MR (2001) Demographic data regimes. In: Smelser N, Baltes P (eds) International encyclopedia of the social and behavioral sciences, vol 5. Pergamon, Oxford, pp 3432–3435
Hodgson D (2001) Demography, twentieth-century. In: smelser N, Baltes P (eds) International encyclopedia of the social and behavioral Sciences, vol 5. Pergaman, Oxford, PP 3493–3498
Rosenzweig, Mark R, Oded Stark (eds) (2001) Handbook of population and family economics, Amsterdom, Elsevier
Smelser, Neil J, Paul B, Baltes (eds) (2001) International Encyclopedia of the Social and behavioral sciences, Elsevier
Szsreter, Simon (2001) Demography, history of. In: Smelser N, Baltes P (eds) International encyclopedia of the social and Behavioral Sciences, vol 5. Pergamon, Oxford, pp 3488–3493

Density Ratio Model

Konstantinos Fokianos
Associate Professor
University of Cyprus, Nicosia, Cyprus

The problem of comparing two (or more samples) appears in several and diverse applications. The parametric theory resolves the problem by appealing to the well-known t-test. To carry out the t-test both samples are assumed to be normally distributed with common unknown variance and unknown means. The two-sample t-test enjoys several optimality properties, for instance, it is uniformly the most powerful unbiased test. Occasionally some (or all) of the needed assumptions fail; for instance, when there exists a group of observations with skewed distribution, then both assumptions of normality and equality of variances do not hold true. Hence, application of the ordinary two-sample t-test is questionable. The problem is usually bypassed after a suitable transformation but the comparison needs to be carried out in the transformed scale. Alternatively, we can appeal to the nonparametric theory, which approaches the problem of comparing two samples by the so-called Mann–Whitney–Wilcoxon test (see ►Wilcoxon–Mann–Whitney Test).

We consider a quite different approach to the two-sample comparison problem. The methodology is relatively new and depends on the so-called *density ratio model* for *semiparametric* comparison of two samples. To be more specific, assume that

$$\begin{aligned} X_1, \ldots, X_{n_0} &\sim f_0(x) \\ X_{n_0+1}, \ldots, X_n &\sim f_1(x) = \exp(\alpha + \beta h(x)) f_0(x). \end{aligned} \tag{1}$$

where $f_i(x)$, $i = 0, 1$ are probability densities, h is a *known* function, and α, β are two unknown parameters. In principle, $h(x)$ can be *multivariate* but we assume for simplicity that it is a univariate function.

Model (1) is motivated by means of the standard ►logistic regression and the equivalence between prospective and retrospective sampling, Prentice and Pyke (1979). Suppose that Y is a binary response variable and let X be a covariate. The simple logistic regression model is of the form

$$\mathrm{P}[Y = 1 \mid X] = \frac{\exp(\alpha^\star + \beta h(x))}{1 + \exp(\alpha^\star + \beta h(x))}, \tag{2}$$

where $\alpha^\star$ and β are scalar parameters. Notice that the marginal distribution of X is left completely unspecified. Assume that $X_1, \ldots, X_{n_0}$ is a random sample from $F(x \mid Y = 0)$. Independent of the X_i, assume that $X_{n_0+1}, \ldots, X_n$ is a random sample from $F(x \mid Y = 1)$, and let $n_1 = n - n_0$. Put $\pi = P(Y = 1) = 1 - P(Y = 0)$ and assume that $f(x \mid Y = i) = dF(x \mid Y = i)/dx$ exists and represents the conditional density function of X given $Y = i$ for $i = 0, 1$. A straightforward application of ►Bayes' theorem shows that

$$\frac{f(x \mid Y = 1)}{f(x \mid Y = 0)} = \exp(\alpha + \beta h(x))$$

with $\alpha = \alpha^* + \log[(1 - \pi)/\pi]$. In other words, model (2) is equivalent to (1) with $\alpha = \alpha^* + \log[(1 - \pi)/\pi]$.

We refer to (1) as the density ratio model since it specifies a parametric function of the log likelihood ratio of two densities without assuming any specific form about them. Hence, it is a semiparametric model and it is easy to see that, under the hypothesis $\beta = 0$, both of the distributions are identical. In other words, both density functions are assumed unknown but are related, however, through an exponential tilt–or distortion–which determines the difference between them. Notice that model (1) is quite general and includes examples such as the exponential and

partial exponential families of distributions. Model (1) provides a compromise between the fully parametric and nonparametric approaches to the problem of testing equality of two distribution, see Qin et al. (2002), Kedem et al. (2004), and Fokianos et al. (2005), among others, for applications of the density ratio model to real data problems.

It is also important to note that (1) is a biased sampling model with weights depending upon parameters. Inference regarding biased sampling models has been discussed by Vardi (1982, 1985), Gill et al. (1988), and Bickel et al. (1998), in the case of completely known weight functions, while Qin and Zhang (1997), Qin (1998), Gilbert et al. (1999), Gilbert (2000), and Fokianos et al. (2001) consider weight functions unknown up to a parameter.

It is easy to see that (1) generalizes to the m–samples comparison problem. Consider m unknown densities that are related by an exponential tilt of the following form

$$\begin{aligned} X_{11},\ldots,X_{1n_1} &\sim f_1(x) = \exp\left(\alpha_1 + \beta_1 h(x)\right) f_m(x),\\ X_{21},\ldots,X_{2n_2} &\sim f_2(x) = \exp\left(\alpha_2 + \beta_2 h(x)\right) f_m(x),\\ &\ldots \quad \ldots \quad \ldots \\ X_{m1},\ldots,X_{mn_m} &\sim f_m(x), \end{aligned} \tag{3}$$

where the notation follows (1). Estimation of $\beta_1,\ldots,\beta_{m-1}$ as well as inference regarding the cumulative distribution functions that correspond to $f_1(.),\ldots,f_m(.)$ has been considered by Fokianos et al. (2001), who also propose some test statistics for the hypotheses $\beta_1 = \ldots = \beta_{m-1} = 0$, that is, all the samples are identically distributed. In this sense, model (3) is also referred to as a *semiparametric one-way ANOVA*.

In conclusion, the density ratio model for two and m samples avoids the normal theory by specifying that the log ratio of two unknown densities is of some parametric form. Hence, it provides another way of testing the equality of several distributions without resorting to transformations or any other techniques. The last comment is particularly useful since there are examples of data that show that populations follow skewed distributions and therefore classical estimation theory might yield questionable results. The suggested model accommodates skewed data and provides desirable results such as consistent estimators of means, test statistics, and so on.

About the Author

Konstantinos Fokianos received his Ph.D. in Statistics from the University of Maryland at College Park in 1996 and is currently an Associate Professor with the Department of Mathematics and Statistics, University of Cyprus. His research interests include time series, bioinformatics, and semiparametric statistical inference. He has authored/coauthored 40 papers and the book *Regression Models for Time Series Analysis* (Wiley, 2002 with B. Kedem). He has been an elected member of the International Statistical Institute since 2005. He serves as an associate editor for the *Journal of Environmental Statistics, Statistics* and *Probability Letters* and *Computational Statistics & Data Analysis*.

Cross References

►Exponential Family Models
►Logistic Regression
►Student's *t*-Tests

References and Further Reading

Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1998) Efficient and adaptive estimation for semiparametric models. Springer-Verlag, New York. Reprint of the 1993 original

Fokianos K, Kedem B, Qin J, Short DA (2001) A semiparametric approach to the one-way layout. Technometrics 43:56–64

Fokianos K, Sarrou I, Pashalidis I (2005) A two-sample model for the comparison of radiation doses. Chemom Intell Lab Syst 79:1–9

Gilbert PB (2000) Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. Ann Stat 28:151–194

Gilbert PB, Lele SR, Vardi Y (1999) Maximum likelihood estimation in semiparametric selection bias models with application to AIDS vaccine trials. Biometrika 86:27–43

Gill RD, Vardi Y, Wellner JA (1988) Large sample theory of empirical distributions in biased sampling models. Ann Stat 16:1069–1112

Kedem B, Wolff D, Fokianos K (2004) Statistical comparisons of algorithms. IEEE Trans Instrum Meas 53:770–776

Prentice RL, Pyke R (1979) Logistic disease incidence models and case–control studies. Biometrika 66:403–411

Qin J (1998) Inferences for case–control data and semiparametric two-sample density ratio models. Biometrika 85:619–630

Qin J, Zhang B (1997) A goodness of fit test for the logistic regression model based on case–control data. Biometrika 84:609–618

Qin J, Barwick M, Ashbolt R, Dwyer T (2002) Quantifying the change of melanoma incidence by Breslow thickness. Biometrics 58:665–670

Vardi Y (1982) Nonparametric estimation in the presence of length bias. The Annals of Statistics, 10:616–620

Vardi Y (1985) Empirical distribution in selection bias models. The Annals of Statistics, 13:178–203

Design for Six Sigma

Rick L. Edgeman
Professor and Chair & Six Sigma Black Belt
University of Idaho, Moscow, ID, USA

►Six Sigma can be defined as *a highly structured strategy for acquiring, assessing, and applying customer, competitor, and enterprise intelligence in order to produce superior product, system or enterprise innovation and designs*

(Klefsjö et al. 2006). Six Sigma originated approximately three decades ago as a means of generating near-perfect products via focus on associated manufacturing processes and while initially applied almost exclusively in manufacturing environments, its inherent sensibilities and organization facilitated migration to service operations. Similarly, while Six Sigma was at the outset used to generate significant innovation in and improvement of existing products, those same sensibilities led to its adaptation to new product and process design environments and it is on use of Six Sigma in design applications that the present contribution is focused.

There is a distinction between using Six Sigma principles in design or innovation applications versus a process operating at a level of six sigma. In terms of performance, a process operating at a "true" six sigma level produces an average of only 3.4 defects per million opportunities (DPMO) for defects where this figure is associated with a process with a 12 standard deviation spread between lower and upper specification limits, but wherein the 3.4 DPMO figure is based on allowance for a 1.5 standard deviation non-centrality factor or shift away from "perfect centering" so that, in essence, one specification limit is 4.5 standard deviations away from the targeted or ideal performance level whereas the other specification limit is 7.5 standard deviations away from that performance level. In practice, of course, a process may operate (typically) at lower or (rarely) higher sigma levels – that is, with less or more spread in standard deviation units between specification limits.

Herein we are not focused on the "sigma level" per se, but rather on the design approach leading to the process or product in question that operates at some sigma level. This approach is called *Design for Six Sigma* (DFSS) and is conducive to higher sigma levels – that is, nearer to perfect results – while at the same time aligning with customer demands and desires or, in more customary language, the Voice of the Customer (VOC) expressed as "customer needs and wants".

While multiple DFSS approaches exist, a few similar ones dominate the application arena with the two most prevalent ones being referred to as I^2DOV (Innovation, Invention, Design, Optimization and Verification) and DMADV (Define, Measure, Analyze, Design, Verify) Algorithm with DMADV being more commonly applied of the two and hence emphasized herein.

It should be emphasized that whichever DFSS approach is used, whether DMADV, I^2DOV, or another, that the approach provides freedom within structure rather than rigidity. That is to say that each phase in the chosen approach has a particular intent, that the phases are generally sequential and linked, and that together they are complete, but that within a given phase many and differing tools and techniques can be brought to bear with those used potentially differing substantially from one application to the next.

Turning now to the DMADV approach to DFSS we can provide the following brief descriptions of each portion of the algorithm.

Define (D): A primary goal of the *Define* phase of DFSS is to acquire and access the VOC and subsequently align goals for the product, process, or service with the VOC. We note here that the customers considered should be both internal and external ones, as applicable. Among methods for acquiring the VOC are focus groups, sample surveys, and examination of customer complaints. Another method of value is to directly observe customer use of similar to products, processes or services so that unspoken, more implicit information can be gathered. In goal-setting it is recommended that these be so-called "SMART" goals, where SMART is an acronym for Specific, Measurable, Attainable, Relevant, and Time-bound. Further, while these goals should be attainable, they should not be "easily" attained, but should rather represent stretch goals, ones that are more likely to position the product, process, or service at the leading edge.

Measure (M): In the DFSS context this requires that we measure and match performance to customer requirements. Fundamentally this is a *quantification* of the VOC and the alignment of this quantification with organizational and management goals.

Analyze (A): This phase demands that the *design* for any existing relevant product, process or service be analyzed and assessed to determine its suitability, performance, error or defect sources, and any corrective or innovative actions that may be taken. Various tools are of potential value in this phase, including *Design Failure Modes and Effects Analysis* (DFMEA), a tool whose name belies its intent. Other useful tools and methods include *Concept Generation and Selection*, and the *Theory of Inventive Problem Solving* (TRIZ).

Design (D): In this phase the array of corrective or innovative actions identified in the analyze phase are embedded in the design and subsequent deployment of new processes required to activate the VOC while simultaneously fulfilling organizational and management goals. While various tools may be of value here, a few of the more advanced approaches that are useful include many from *experimental design and response surface analysis* (Myers et al. 2009) along with more rigorous quality oriented approaches such as *Quality Function Deployment* (QFD). As a way of relating and integrating these latter approaches various customer needs and wants (the VOC) that are critical to QFD and its so-called *House of Quality*

(HOQ) can be regarded as response variables (Y's) whose selected or joint optimization is attained through deployment of identified product or process design attributes are controllable variables $X_1, X_2, \ldots X_P$ so that we have

$$Y = f(X_1, X_2, \ldots X_P) + \varepsilon$$

Where the optimal combination of settings of $X_1, X_2, \ldots X_P$ – called "engineering attributes" in the parlance of QFD – can be determined through use of, e.g., response surface methods, steepest ascent methods, and evolutionary operations or EVOP (Myers et al. 2009). It is important to note that it is not sufficient to simply identify the optimal combination of these variables as it is their specific means of deployment – the process – that ultimately actualizes the VOC.

Verify (V): In the *Verify* phase the objective is to assess performance of the design via such means as prototyping, simulation, or direct observation of the designed product or process in use prior to marketplace deployment. In this way design performance is verified.

From the description of DMADV it is easily concluded that a variety of statistical and other methods can be used to support its effectiveness and a few of these have been suggested herein. That said, it should be evident that the specific methods applied are almost boundless, being limited only as they are primarily by the knowledge and imagination of the design team. In all DMADV offers a logical and highly structured, yet versatile approach to product, process, or service design.

About the Author

Rick Edgeman in Professor and Chair of Statistics at the University of ldaho and Professor in the Aarhus University (Denmark) Summer University and also serves on the Advisory Board of Hamdan Bin Mohammed E-University of Dubai. He has more than 100 publications in leading journals to his credit with much of his work in the areas of Six Sigma, Sustainability, Innovation, Quality Management, Leadership, and Statistical Applications in Quality and Reliablity Engineering. In 2000 he was cited in Quality Progress as one of 21 Voices of Quality for the 21st Century, one of only six academics worldwide so identified.

Cross References

►Business Statistics
►Industrial Statistics
►SIPOC and COPIS: Business Flow–Business Optimization Connection in a Six Sigma Context
►Six Sigma

References and Further Reading

Klefsjö B, Bergquist B, Edgeman R (2006) Six sigma and total quality management: different day, same soup? Six Sigma & Competitive Advantage 2(2):162–178

Myers RH, Montgomery DC, Anderson-Cook CM (2009) Response surface methodology: process and product optimization using designed experiments, 3rd edn. Wiley, New York

Design of Experiments: A Pattern of Progress

D. R. Cox
Honorary Fellow
Nuffield College, Oxford, UK

The article **Experimental Design, Introduction to** (Hinkelman 2010) sets out the basic statistical principles of experimental design. This supplementary note comments on the historical development of the subject.

Careful experimentation has a long history, perhaps especially in the physical sciences. There is, however, a long history also of experimentation in fields as diverse as agriculture and clinical medicine. The first systematic discussion of experimental design in the presence of substantial haphazard variation seems to be that of Fisher (1926), later developed in his book (Fisher 1935). He set out four principles:

- error control, for example by some form of matching to compare like with like
- independent replication to improve precision and allow its estimation
- randomization to achieve a number of aims, notably avoidance of selection biases
- factorial design to improve the efficiency of experimentation and to allow the exploration of interactions.

The relative importance of these four principles varies between subject-matter fields. This accounts to some extent for differences in how the subject has developed when directed toward, say, agricultural field trials as contrasted with some other fields of application.

In the period up to 1939 these ideas were extensively developed, notably by Fisher's friend and colleague, F. Yates, whose monograph (Yates 1937) is a little known masterpiece of the statistical literature. The focus and impact of this work was primarily but by no means exclusively agricultural.

In the 1950's a strand of new ideas entered from the chemical process industries notably with the work of G.E.P.

Box and his associates; see, for example, Box and Draper (1987). The differences were not so much that factors in a factorial experiment mostly had quantitative levels as that emphasis shifted from the estimation of factorial effects to the response surface of mean outcome considered as a smooth function of the factor level. This led to a richer family of designs and to an emphasis on exploring the form of the response surface in the neighborhood of an optimum. In some of the applications error was relatively small and experiments could be completed quite quickly allowing for a developing sequence of designs.

Carefully designed clinical trials have a long history but the period from about 1970 onward saw their application and development on a large scale. Here typically there a very small number of possible treatments, often just two, and error is large, so that substantial replication is essential. Avoidance of various forms of selection bias by concealment achieved by ►randomization is often crucial. See Piantadosi (1997) for a systematic account.

Two more recent areas of development are applications to matters of public policy, for example in education and criminology. Here an issue concerns the extent to which a parallel with randomized clinical trials is appropriate. The second and very different application concerns the systematic sensitivity analysis of complex computer models involving many adjustable parameters and computationally highly intensive individual runs. Models of climate change are an example.

In all these areas choice of a specific design in a particular context typically involves largely qualitative considerations of balancing the primary requirements of achieving precision and clarity and security of interpretation with practical constraints that are always present, although taking different forms in different fields. There may also be a number of distinct somewhat conflicting objectives. This often makes formal theoretical analysis of design choice difficult. Nevertheless a formal theory of design choice as an optimization problems has appeal both in the sense of showing the formal optimality of standard designs in specific circumstances and in guiding how to proceed in unusual situations, as for example there are specific technical constraints on the factor combinations that may be studied in a complex factorial experiment. A landmark result in such a theory of optimal design is the General Equivalence Theorem of Kiefer and Wolfowitz (1959).

About the Author

For biography *see* the entry ►Survival Data.

Cross References

►Clinical Trials: An Overview
►Experimental Design: An Introduction
►Factorial Experiments
►Optimum Experimental Design
►Randomization
►Statistical Design of Experiments (DOE)
►Uniform Experimental Design

References and Further Reading

Box GEP, Draper NR (1987) Empirical model building and response surfaces. Wiley, New York
Fisher RA (1926) The arrangement of field experiments. J Minist Agric 13:311–320
Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh (and subsequent editions)
Hinkelman K (2010) This publication
Kiefer J, Wolfowitz J (1959) Optimal designs in regression problems. Ann Math Stat 30:271–294
Piantadosi S (1997) Clinical trials. Wiley, New York
Yates F (1937) The design and analysis of factorial experiments. Imperial Bureau of Soil Science, Harpenden

Designs for Generalized Linear Models

André I. Khuri
Professor Emeritus
University of Florida, Gainesville, FL, USA

Introduction

►Generalized linear models (GLMs) represent an extension of the class of linear models. They are used to fit models in general situations where the response data under consideration can be discrete or continuous. Thus, GLMs provide a unified approach for the analysis of such data. Nelder and Wedderburn (1972) are credited for having introduced these models.

Three components are needed to define GLMs. These components are:

(a) The response data are values of a random variable, denoted by y, whose distribution is of the exponential type. Its density function (or probability mass function for a discrete distribution) is of the form

$$\ell(y,\theta,\phi) = \exp\left[\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)\right], \qquad (1)$$

where $a(.)$, $b(.)$, and $c(.)$ are known functions, θ is a so-called *canonical parameter*, and ϕ is a *dispersion parameter*.

(b) The mean response, μ, is related to the so-called *linear predictor*, η, through a function, denoted by h, such that $\mu = h(\eta)$, which is assumed to be monotone and differentiable. The inverse function of h, denoted by g, is called the *link function*. Thus,

$$\eta = g(\mu). \tag{2}$$

The value of the mean response at a point, $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)'$, in a k-dimensional Euclidean space is denoted by $\mu(\boldsymbol{x})$, and the corresponding value of η is denoted by $\eta(\boldsymbol{x})$. Here, $x_1, x_2, \ldots, x_k$ represent control variables that affect the response variable y.

(c) The linear predictor is represented in terms of a linear model of the form

$$\eta(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\beta, \quad \boldsymbol{x} \in \mathcal{R}, \tag{3}$$

where $\mathcal{R}$ is a certain region of interest in the k-dimensional space, $\boldsymbol{f}(\boldsymbol{x})$ is a known function of $\boldsymbol{x}$, and β is a vector of p unknown parameters. It follows that the mean response value at $\boldsymbol{x}$ is given by

$$\begin{aligned}\mu(\boldsymbol{x}) &= h[\eta(\boldsymbol{x})] \\ &= h[\boldsymbol{f}'(\boldsymbol{x})\beta].\end{aligned} \tag{4}$$

Given a set of n independent observations on the response y, namely $y_1, y_2, \ldots, y_n$, at n distinct locations in $\mathcal{R}$, an estimate of β in (3) is obtained by using the method of maximum likelihood. This method maximizes the likelihood function given by

$$\mathcal{L}(\theta, \phi, \boldsymbol{y}) = \prod_{i=1}^{n} \ell(y_i, \theta_i, \phi), \tag{5}$$

with respect to β, where θ_i is a canonical parameter corresponding to y_i $(i = 1, 2, \ldots, n)$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)'$, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$. The dispersion parameter, ϕ, is considered to have a fixed value that does not change over the values of y_i $(i = 1, 2, \ldots, n)$. The *maximum likelihood estimate* (MLE) of β is denoted by $\hat{\beta}$. Details about the computation of $\hat{\beta}$ can be found in McCullagh and Nelder (1989, Chap. 2), McCulloch and Searle (2001, Chap. 5), and Dobson (2002, Chap. 4).

The variance–covariance matrix of $\hat{\beta}$ is approximately equal to (see, for example, McCulloch and Searle 2001:143)

$$\mathrm{Var}(\hat{\beta}) \approx (\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}, \tag{6}$$

where $\boldsymbol{X}$ is the model matrix for the linear predictor in (3) based on the design matrix used to generate the response data, $y_1, y_2, \ldots, y_n$, and $\boldsymbol{W}$ is the diagonal matrix,

$$\boldsymbol{W} = \bigoplus_{i=1}^{n} \{[g'(\mu_i)]^2 a(\phi) b''(\theta_i)\}. \tag{7}$$

In formula (7), $\oplus$ is the direct sum notation, $g'(\mu_i)$ is the derivative of $g(\mu)$ with respect to μ evaluated at μ_i, and $b''(\theta_i)$ is the second derivative of $b(\theta)$ with respect to θ evaluated at θ_i $(i = 1, 2, \ldots, n)$.

An estimate of the linear predictor, $\eta(\boldsymbol{x})$, in (3) is given by

$$\hat{\eta}(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\hat{\beta}. \tag{8}$$

Using (6), the variance of $\hat{\eta}(\boldsymbol{x})$ is approximately equal to

$$\mathrm{Var}[\hat{\eta}(\boldsymbol{x})] \approx \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{f}(\boldsymbol{x}). \tag{9}$$

The mean response, $\mu(\boldsymbol{x})$, in (4) can then be estimated by using the expression,

$$\hat{\mu}(\boldsymbol{x}) = h[\boldsymbol{f}'(\boldsymbol{x})\hat{\beta}]. \tag{10}$$

This estimate is called the *predicted response* at $\boldsymbol{x}$. From (9) and (10) it follows that the variance of $\hat{\mu}(\boldsymbol{x})$ is approximately equal to

$$\mathrm{Var}[\hat{\mu}(\boldsymbol{x})] \approx \{h'[\eta(\boldsymbol{x})]\}^2 \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{f}(\boldsymbol{x}), \tag{11}$$

where $h'[\eta(\boldsymbol{x})]$ is the derivative of h with respect to η evaluated at $\boldsymbol{x}$. Formula (11) results from taking the variance of the first-order Taylor's series approximation of $h[\hat{\eta}(\boldsymbol{x})]$ in a neighborhood of $\eta(\boldsymbol{x})$. The expression on the right-hand side of (11) is called the *prediction variance*.

It should be noted that $\hat{\mu}(\boldsymbol{x})$ is a biased estimator of $\mu(\boldsymbol{x})$. A measure of closeness of $\hat{\mu}(\boldsymbol{x})$ to $\mu(\boldsymbol{x})$ is given by its mean-squared error, namely,

$$\mathrm{MSE}[\hat{\mu}(\boldsymbol{x})] = \mathrm{E}[\hat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})]^2,$$

which can be partitioned as

$$\mathrm{MSE}[\hat{\mu}(\boldsymbol{x})] = \mathrm{Var}[\hat{\mu}(\boldsymbol{x})] + \{\mathrm{E}[\hat{\mu}(\boldsymbol{x})] - \mu(\boldsymbol{x})\}^2. \tag{12}$$

The second expression, $\mathrm{E}[\hat{\mu}(\boldsymbol{x})] - \mu(\boldsymbol{x})$, on the right-hand side of (12) is called the *bias* associated with estimating $\mu(\boldsymbol{x})$, and is denoted by $\mathrm{Bias}[\hat{\mu}(\boldsymbol{x})]$. We thus have

$$\mathrm{MSE}[\hat{\mu}(\boldsymbol{x})] = \mathrm{Var}[\hat{\mu}(\boldsymbol{x})] + \{\mathrm{Bias}[\hat{\mu}(\boldsymbol{x})]\}^2. \tag{13}$$

This is called the *mean-squared error of prediction* (MSEP) evaluated at $\boldsymbol{x}$. A second-order approximation of $\mathrm{Bias}[\hat{\mu}(\boldsymbol{x})]$ is described in Robinson and Khuri (2003).

Choice of Design for GLMs

By a choice of design, it is meant the determination of the settings of the control variables, $x_1, x_2, \ldots, x_k$, that yield an estimated (or predicted) response with desirable properties. Desirability is assessed by having small values for the prediction variance in (11), or small values for the MSEP in (13). Thus, we can have designs that minimize the prediction variance, or designs that minimize the MSEP. The

former designs utilize criteria similar to those used in linear models, such as *D-optimality* and *G-optimality*, and are therefore referred to as *variance-related criteria*. However, because of the bias in estimating $\mu(\boldsymbol{x})$, it would be more appropriate to adopt a design criterion based on minimizing the MSEP in (13).

The Design Dependence Problem

One problem that faces the actual construction of a design for GLMs is the dependence of the design on β, the vector of unknown parameters in the linear predictor in (3). This is true since $\eta(\boldsymbol{x})$, and hence $\mu(\boldsymbol{x})$, depends on β. Consequently, the elements of the matrix $\boldsymbol{W}$ in (7), which is used in the formulation of both the prediction variance and the MSEP, also depend on β. Thus, to minimize any design criterion function, some knowledge of β is required. This is quite undesirable since the purpose of any design is to estimate β in order to estimate the mean response $\mu(\boldsymbol{x})$.

Common approaches to solving this design dependence problem include the following:

(a) The specification of "guessed", or initial, values of the unknown parameters involved. These values are used in the determination of the so-called *locally-optimal design*. Some references that discuss this approach include those by Mathew and Sinha (2001), Wu (1988), and Sitter and Wu (1993).

(b) The *sequential approach* which starts by using initial values for the unknown parameters. The design derived from these values is then utilized to obtain estimates of the parameters which are then used as updated values leading to another design, and so on. Sequential designs were proposed by Wu (1985), Sitter and Forbes (1997), and Sitter and Wu (1999).

(c) The *Bayesian approach* which assumes a prior distribution on the elements of β. This distribution is then incorporated into an appropriate design criterion by integrating it over the prior distribution. This approach was discussed by several authors. See, for example, Chaloner and Verdinelli (1995) and Atkinson and Haines (1996).

(d) The use of the *quantile dispersion graphs* approach. This more recent approach considers the MSEP in (13) as a criterion for comparing designs rather than selecting an optimal design. More specifically, quantiles of $\mathrm{MSE}[\hat{\mu}(\boldsymbol{x})]$ values in (13) are obtained on several concentric surfaces inside a region of interest, $\mathcal{R}$, which is a subset of the k-dimensional Euclidean space (recall that k is the number of control variables in the linear predictor model in (3)). Let $\mathcal{R}_\nu$ denote the surface of a region obtained by reducing $\mathcal{R}$ using a shrinkage factor, ν $(0 < \nu \leq 1)$. Furthermore for a given design, D, let $Q_D(p, \beta, \nu)$ denote the p^{th} quantile of the distribution of the values of $\mathrm{MSE}[\hat{\mu}(\boldsymbol{x})]$ on $\mathcal{R}_\nu$. Several concentric surfaces can be so obtained by varying the values of ν.

In order to assess the dependence of the design D on β, a certain parameter space, $\mathcal{C}$, to which β is assumed to belong, is specified. Then, the minimum and maximum values of $Q_D(p, \beta, \nu)$ with respect to β are computed over the parameter space $\mathcal{C}$. This results in the following extrema of $Q_D(p, \beta, \nu)$ for each ν and a given p:

$$Q_D^{min}(p, \nu) = \min_{\beta \in \mathcal{C}} \{Q_D(p, \beta, \nu)\}$$

$$Q_D^{max}(p, \nu) = \max_{\beta \in \mathcal{C}} \{Q_D(p, \beta, \nu)\}.$$

Plotting these values against p produces the so-called *quantile dispersion graphs* (QDGs) of the MSEP over the surface $\mathcal{R}_\nu$ for the design D. By repeating this process using several values of ν we obtain plots that depict the prediction capability of the design D throughout the region $\mathcal{R}$. Several plots can be so constructed for each of several candidate designs for the model in (3). A preferred design is one that has small values of Q_D^{min} and Q_D^{max} over the range of p $(0 \leq p \leq 1)$. More details concerning this approach with examples can be found in Khuri and Mukhopadhyay (2006). The approach itself was first introduced in Robinson and Khuri (2003).

The general problem of design dependence along with the aforementioned four approaches were discussed in detail by Khuri et al. (2006).

About the Author

For biography *see* the entry ▶Response Surface Methodology.

Cross References

▶Generalized Linear Models
▶Optimum Experimental Design

References and Further Reading

Atkinson AC, Haines LM (1996) Designs for nonlinear and generalized linear models. In: Ghosh S, Rao CR (eds) Design and analysis of experiments. North-Holland, Amsterdam, pp 437–475

Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. Stat Sci 10:273–304

Dobson AJ (2002) An introduction to generalized linear models, 2nd edn. Chapman & Hall/CRC, Boca Raton

Khuri AI, Mukhopadhyay S (2006) GLM designs: the dependence on unknown parameters dilemma. In: Khuri AI (ed)

Response surface methodology and related topics. World Scientific, Singapore, pp 203–223
Khuri AI, Mukherjee B, Sinha BK, Ghosh M (2006) Designs issues for generalized linear models: a review. Stat Sci 21:376–399
Mathew T, Sinha BK (2001) Optimal designs for binary data under logistic regression. J Stat Plan Inference 93:295–307
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London
McCulloch CE, Searle SR (2001) Generalized, linear and mixed models. Wiley, New York
Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc A 135:370–384
Robinson KS, Khuri AI (2003) Quantile dispersion graphs for evaluating and comparing designs for logistic regression models. Comput Stat Data Anal 43:47–62
Sitter RR, Forbes BE (1997) Optimal two-stage designs for binary response experiments. Stat Sin 7:941–955
Sitter RR, Wu CFJ (1993) Optimal designs for binary response experiments: Fieller, D, and A criteria. Scand J Stat 20:329–341
Sitter RR, Wu CFJ (1999) Two-stage design of quantal response studies. Biometrics 55:396–402
Wu CFJ (1985) Efficient sequential designs with binary data. J Am Stat Assoc 80:974–984
Wu CFJ (1988) Optimal design for percentile estimation of a quantal response curve. In: Dodge Y, Fedorov VV, Wynn HP (eds) Optimal design and analysis of experiments. North–Holland, Amsterdam, pp 213–223

Detecting Outliers in Time Series Using Simulation

Abdalla M. El-Habil
Head of the Department of Applied Statistics, Faculty of Economics and Administrative Sciences
Al-Azhar University, Gaza, Palestine

Introduction

►Outliers have recently been studied more in the statistical time series literature and this interest is also growing in econometrics. Usually, time series outliers are informally defined as somehow unexpected or surprising values in relation to the rest of the series.

Data of potential value in the formulation of public and private policy frequently occur in the form of time series. Most time series data are observational in nature. In addition to possible gross errors, time series data are often subject to the influence of some uncontrolled or unexpected interventions, for example, implementations of a new regulation, major changes in political or economic policy, or occurrence of a disaster. Consequently, discordant observations and various types of structural changes occur frequently in time series data. Whereas the usual time series model is designed to grasp the homogeneous memory pattern of a time series, the presence of outliers, depending on their nature, may have a moderate to substantial impact on the effectiveness of the standard methodology for time series analysis with respect to model identification, estimation, and forecasting. Therefore, there is a clear need to have available methods to detect or accommodate them.

Simulation data are derived from a sequence of pseudorandom numbers. These pseudorandom numbers are created by a random number generator. The generator requires an initial seed value from which to generate its first value. The random number generator creates both a random number and a new seed for the next value.

The SIMULATE paragraph in the Scientific Computing Associate Corporation (SCA) program may be used to estimate an ARIMA model or a transfer function model. The use of the SIMULATE paragraph for the estimation of a transfer function model is identical as its use for the estimation of an ARIMA model, except for the presence of input series. The SIMULATE paragraph will first generate a noise sequence using a pseudorandom number generator. This sequence is then used according to a transfer function model specified lately using the TSMODEL paragraph.

Detecting Outliers of a Simulated $AR(1)$ Time Series

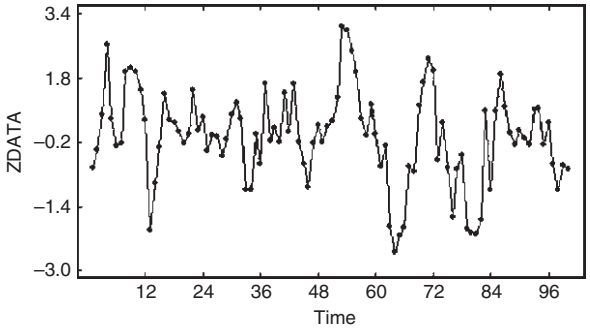
To facilitate our understanding of detecting outliers and their effects, for example, on the values of a simulated $AR(1)$ process, we will assume that the constant of the proposed model is equal to zero. For this purpose, 100 observations are simulated from the model $z_t = [1/(1-0.6B)]a_t$ with $\sigma_a = 1.0$. The data are shown in Fig. 1.

To illustrate, for example, the effect of an AO on the base $AR(1)$ model, we include an AO at time $t = 42$ with $\omega A = 6$ (the value ωA represents the amount of deviation from the "true" value of ZT). The new shape of data is shown in Fig. 2.

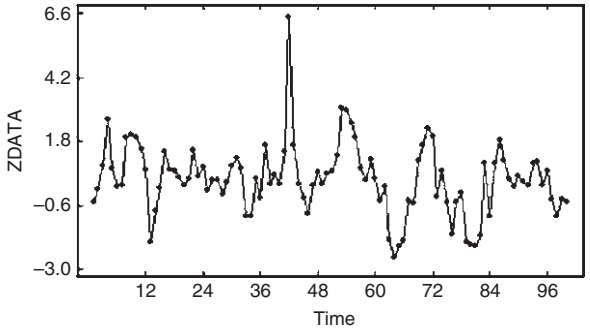
By using the SCA program, only AO has been detected at $t = 42$, and we obtain the estimation results for an $AR(1)$ fit of the simulated $AR(1)$ process as the following:

Case	φ estimate	S. E. of φ estimate	σ_a estimate
Without outlier	0.5921	0.0810	0.9783
AO at time $t = 42$	0.4683	0.0888	1.2177

From the table, with the additive outlier at time $t = 42$, we can see that the parameter estimate is decreased



Detecting Outliers in Time Series Using Simulation. Fig. 1
Zdata



Detecting Outliers in Time Series Using Simulation. Fig. 2
Zdata 1

by approximately 0.13, the estimated residual variance is inflated, and in consequence the prediction intervals can be too wide. In turn, it will affect the model identification, estimation, and forecasting.

Simulation of a Single-Equation Transfer Function Model (with Two-Input Variables)

To detect outliers and study their effects on the values of a simulated single-equation transfer function model (with two-input variables), 300 observations are simulated from the model

$$\text{zdata} = 12.0 + (0.6)\text{xdata} + (0.7)\text{ydata} + \text{at},$$

where the model of xdata is

$$(1 - 0.66B)\text{xdata} = 12.0 + \text{at},$$

and the model of ydata is

$$(1 - 0.7B)\text{ydata} = 11.0 + (1 - 0.6B)\text{at}, \text{ with } \sigma_a = 2.25.$$

We select only the last 250 values of xdata, ydata, and zdata to ensure that any potential irregularities in the beginning of the recursive computation of values are eliminated.

By using the SCA program, we estimated the model

$$\text{zdata} = 12.0 + (\omega 1)\text{xdata} + (\omega 2)\text{ydata} + \text{at},$$

AO has been detected at $t = 50, 82, 106$ and *TC* at $t = 160$. We obtain estimation results for a single-equation transfer function model (with two-input variables) fit of the simulated single-equation transfer function model (with two-input variables) process as the following:

Case	Estimate of $\omega 1$	S. E. of estimate of $\omega 1$	Estimate of $\omega 2$	S. E. of estimate of $\omega 2$	Estimate of σ_a
Without outlier	0.5741	0.0463	0.7416	0.0518	2.3085
AO at $t = 50$, 82, 106 and *TC* at $t = 160$	0.5494	0.0424	0.7448	0.0476	2.3096

As we see from the table, the parameter estimates are moderately changed, and the estimated residual variance is inflated. Thus, the presence of those extraordinary events could easily mislead the conventional time series analysis.

Simulation of Simultaneous Transfer Function (STF) Model

In order to detect outliers and study their effects on the values of a simulated simultaneous transfer function model, 150 observations are simulated from the models

$$\text{Z1data} = 17.0 + (1 - 0.5B)\text{at},$$
$$\text{Z2data} = 25.0 + (1 - 0.6B)\text{at},$$

with $\sigma_a = 2.25$.

By using the SCA program, we estimated the two models simultaneously using the STFMODEL JOINTMDL paragraph. *TC* has been detected at $t = 112$ and *IO* at $t = 126$. We get estimation results for simultaneous transfer

function model fit of the simulated simultaneous transfer function process as the following:

Estimation results for the simultaneous transfer function fit of the simulated transfer function process (first model)

Case	Estimate of z1data	S. E. of estimate of z1data	Estimate of z2data	S. E. of estimate of z2data
Without outlier	0.4579	0.0726	0.0702	0.0811
TC at $t = 112$	0.4786	0.0671	0.0786	0.0794

Estimation results for the simultaneous transfer function fit of the simulated transfer function process (second model)

Case	Estimate of z1data	S. E. of estimate of z1data	Estimate of z2data	S. E. of estimate of z2data
Without outlier	0.0061	0.0369	0.7262	0.0568
TC at $t = 126$	−0.0079	0.0706	0.7452	0.0553

As we see from the above two tables, the parameters estimates are changed, and the estimated residual variance is inflated. So, those outliers could easily mislead the conventional time series analysis.

Summary

In this entry, simulations for detecting outliers and studying their effects on the values of $AR(1)$ time series, transfer function model with one-input variable, transfer function model with two-input variables processes, and simultaneous transfer function (STF) are conducted using the STFMODEL JOINTMDL paragraph in the Scientific Computing Associate Corporation (SCA) program. The conclusion, which we come up with, is that the presence of outliers, depending on their nature, may have a moderate to substantial impact on the effectiveness of the standard methodology for time series analysis with respect to model identification, estimation, and forecasting.

About the Author

Dr. Abdalla El-Habil is Professor of Statistics, and Head of the Department of Applied Statistics at the Faculty of Economics & Admin. Sciences, Al-Azhar University – Gaza, Palestine. He is Past Dean of the Faculty of Economics & Admin. Sciences (2006–2008). He is a member of The International Association for Statistical Education and The International Biometrics Society. He is a member of The Consultative Council of Professional Statistics for Palestine. He was a Chairman of the Palestinian Economists and Statisticians Association (1993–2000).

Cross References

►Multivariate Outliers
►Outliers
►Time Series

References and Further Reading

Box GEP, Jenkins GM (1970) Time series analysis: forecasting and control. Holden Day, San Francisco

Chen C, Liu LM (1993) Joint estimation of model parameters and outlier effects in time series. J Am Stat Assoc 88:284–297

Lon-Mu L, Hudak GB (1992–2000) Forecasting and time series analysis using the SCA Statistical System, vols 1 and 2. Scientific Computing Associates, Dekalb

Tsay RS (1986) Time series model specification in the presence of outliers. J Am Stat Assoc 81:132–141

Detection of Turning Points in Business Cycles

Marianne Frisén
Professor, Statistical Research Unit
University of Gothenburg, Gothenburg, Sweden

Turns in Business Cycles

A turn in a business cycle is a change from a phase of expansion to one of recession (or vice versa). Both government and industry need to have systems for predicting the future state of the economy, for example in order to timely predict the shift from a period of expansion to one of recession. Warnings of a turn can be given by using information from one or several time series that are leading in relation to the actual business cycle. A system for detecting the turning points of a leading indicator can give us early indications on the future behavior of the business cycle.

As pointed out for example by Diebold and Rudebusch (1996), Kim and Nelson (1998) and Birchenhall et al. (1999), two distinct but related approaches to the characterisation and dating of the business cycle can be discerned. One approach emphasizes the common movements of several variables. This approach is pursued for example by Stock and Watson (1993). The other approach, the

regime shift, is the one pursued in the works by Neftci (1982), Diebold and Rudebusch (1989), Hamilton (1989), Jun and Joo (1993), Lahiri and Wang (1994), Layton (1998), Birchenhall et al. (1999), Koskinen and Öller (2003) and Andersson et al. (2005, 2006).

An important issue is which characteristic of the leading index best predicts a turn in the business cycle. The question remains whether the most useful predictor is the level, as in Birchenhall et al. (1999), the transition and level, as in Hamilton (1989) and Koskinen and Öller (2003), or the transition and change in monotonicity, as in Andersson et al. (2005, 2006).

The Detection Problem

There is a need for methods for early warning, i.e., methods for the timely detection of a regime shift in a leading index. For reviews and general discussions on the importance of timeliness in the detection of turning points, see for example Neftci (1982), Zarnowitz and Moore (1982), Hackl and Westlund (1989), Zellner et al. (1991), Li and Dorfman (1996) and Layton and Katsuura (2001).

The inference situation can be described as one of ►*surveillance*, since a time series is continuously observed with the goal of detecting the turning point in the underlying process as soon as possible. Repeated decisions are made, the sample size is increasing and no null hypothesis is ever accepted. Thus, the inference situation is different from that where we have a fixed number of observations.

The aim of statistical surveillance is the timely detection of important changes in the process that generates the data. A process X (a leading economic indicator) is under surveillance, where X is often measured monthly or quarterly. Based on the available observations, we decide whether the observations made so far indicate a turn.

Thus at every decision time s, we use the alarm system to decide whether there has been a turn or not. This can be formulated as discriminating between two events at each decision time: D = "the turn has not occurred yet" and C = "the turn has occurred".

Models

An important question is which assumptions can be made about the process.

Economic time series often exhibit seasonal variation. Unfortunately, most data-driven filters can seriously alter the turning point times.

Autocorrelation can be a problem when the sampling intervals are short (Luceno and Box (2000)). When this is expected to be a problem, methods that adjust for autocorrelation should be used. Ivanova et al. (2000) argue that the effect of the autoregressive parameters will largely be captured by the probabilities of remaining in the current state.

Many macroeconomic variables can be characterized as cyclical movements around a trend. In order to distinguish the movements and make the time series stationary in relation to the cycle it is sometimes judged necessary to adjust for the trend. Adjusting for the trend by data transformation may result in a distortion of the characteristics of the original series. Canova (1999) points out that the trend may interact with the cyclical component, and therefore it may be difficult to isolate.

In the common movement approach, a business cycle is characterized as the cyclical movement of several economic activities. This is one example of how multivariate data are used. The common movement approach demonstrates that important information is contained in the relation between the turns of various indices. This information can be utilized either by transforming the problem to a univariate one (by using a composite index of leading indicators) or by applying a special method for surveillance of multivariate data. See for example Ryan (2000), Frisén (2010) and Frisén et al. (2010).

When estimating the parameters of the monitoring system, historical data are often used. The user of a system for on-line detection is faced with the paradox that the parameters in the surveillance system may be estimated using previous data, which means that it is assumed that previous patterns will repeat themselves. However, the aim of the surveillance method is to detect changes, and by estimating parameters from previous data, the ability to detect changes in the current cycle might be diminished. Sarlan (2001) examines the change in intensity and duration of US business cycles and concludes that the modern business cycle is different from the historical one.

The approach described by Andersson et al. (2005, 2006) of a non-parametric method may be preferred in order to avoid the risk of misleading results. The non-parametric method does not assume that all phases are of the same type or have the same level and parametric shape. In practice, this varies a lot. At time τ there is a turn in the expected value, μ. Different assumptions can be made about μ, conditional on D (expansion) and C (recession), and at a turn. Instead of assuming that the parametric shape is known, we can use monotonicity restrictions to define μ under C and D. Then the aim (at the detection of a peak) is to discriminate between the following two events: $D(s): \mu(1) \leq \ldots \leq \mu(s)$ and $C(s)$: "$\mu(1) \leq \ldots \leq \mu(\tau-1)$ and $\mu(\tau-1) \geq \mu(\tau) \geq \ldots \geq \mu(s)$". The monotonicity restrictions for a trough are the opposite. In such situations the exact parametric shape of μ is unknown. We only know that μ is monotonic within each phase.

Methods

In recent years, methods based on likelihood have been in focus. The performance of three methods for turning point detection in leading indicators is compared in detail by Andersson et al. (2005, 2006). All three methods are based on likelihood, but there are differences in model specifications, the amount of information used and parameter estimation. The Hidden Markov Model (HMM) is suggested for business cycle modeling for example by Hamilton (1989), Lahiri and Moore (1991), Lahiri and Wang (1994), Layton (1996), Gregoir and Lenglart (2000) and Koskinen and Öller (2003). The HMlin method is based on regime switching and has a model which is piecewise linear. In many ways it is similar to, for example, the method presented by Koskinen and Öller (2003). The SRlin method is based on the Shiryaev-Roberts (SR) technique under the assumption of a piecewise linear model. The SRnp method is a generalized version of the SR method. It is a non-parametric version of the SRlin method with no parametric assumption on the shape of the curve. It uses only the monotonicity change and not the level. The safe way of the SRnp method was recommended since there is no major loss of efficiency.

Andersson et al. (2005, 2006) illustrated the methods by monitoring a period of Swedish industrial production. Evaluation measures that reflect timeliness were used.

About the Author

For biography *see* the entry ►Surveillance.

Cross References

►Business Statistics
►Economic Statistics
►Forecasting: An Overview
►Surveillance

References and Further Reading

Andersson E, Bock D, Frisén M (2005) Statistical surveillance of cyclical processes with application to turns in business cycles. J Forecasting 24:465–490

Andersson E, Bock D, Frisén M (2006) Some statistical aspects on methods for detection of turning points in business cycles. J Appl Stat 33:257–278

Birchenhall CR, Jessen H, Osborn DR, Simpson P (1999) Predicting us business-cycle regimes. J Bus Econ Stat 17:313–323

Canova F (1999) Does detrending matter for the determination of the reference cycle and the selection of turning points? Econ J 109:126–150

Diebold FX, Rudebusch GD (1989) Scoring the leading indicators. J Bus 62:369–391

Diebold FX, Rudebusch GD (1996) Measuring business cycles: a modern perspective. Rev Econ Stat 78:67–77

Frisén M (2010) Principles for multivariate surveillance. In: Lenz H-J, Wilrich P-T, Schmid W (eds) Frontiers in statistical quality control 9, pp 133–144

Frisén M, Andersson E, Schiöler L (2009) Evaluation of multivariate surveillance. J Appl Stat 37:12

Gregoir S, Lenglart F (2000) Measuring the probability of a business cycle turning point by using a multivariate qualitative hidden markov model. J Forecasting 19:81–102

Hackl P, Westlund AH (1989) Statistical analysis of "structural change", Empirical Econ 14:167–192

Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357–384

Ivanova D, Lahiri K, Seitz F (2000) Interest rate spreads as predictors of german inflation and business cycles. Int J Forecasting 16:39–58

Jun DB, Joo YJ (1993) Predicting turning points in business cycles by detection of slope changes in the leading composite index. J Forecasting 12:197–213

Kim C-J, Nelson CR (1998) Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. Rev Econ Stat 80:188–201

Koskinen L, Öller L-E (2003) A classifying procedure for signalling turning points. J Forecasting 23:197–214

Lahiri K, Moore G (1991) Leading economic indicators: new approaches and forecasting record. Cambridge University Press, Cambridge

Lahiri K, Wang JG (1994) Predicting cyclical turning points with a leading index in a markov switching model. J Forecasting 13:245–263

Layton AP (1996) Dating and predicting phase changes in the U.S. business cycle. Int J Forecasting 12:417–428

Layton AP (1998) A further test of the influence of leading indicators on the probability of us business cycle phase shifts. Int J Forecasting 14:63–70

Layton AP, Katsuura M (2001) Comparison of regime switching, probit and logit models in dating and forecasting us business cycles. Int J Forecasting 17:403–417

Li DT, Dorfman JH (1996) Predicting turning points through the integration of multiple models. J Bus Econ Stat 14:421–428

Luceno A, Box GEP (2000) Influence of the sampling interval, decision limit and autocorrelation on the average run length in cusum charts. J Appl Stat 27:177–183

Neftci S (1982) Optimal prediction of cyclical downturns. J Econ Dynam Control 4:225–241

Ryan TP (2000) Statistical methods for quality improvement. Wiley, New York

Sarlan H (2001) Cyclical aspects of business cycle turning points. Int J Forecasting 17:369–382

Stock JH, Watson MW (1993) A procedure for predicting recessions with leading indicators: econometric issues and recent experience. In: Stock JH, Watson MW (eds) Business cycles, indicators and forecasting, vol 28. University of Chicago Press, Chicago, pp 95–156

Zarnowitz V, Moore GH (1982) Sequential signals of recessions and recovery. J Bus 55:57–85

Zellner A, Hong C, Min C-K (1991) Forecasting turning points in international output growth rates using Bayesian exponentially weighted autoregression, time-varying parameter, and pooling techniques. J Econometrics 49:275–304

Dickey-Fuller Tests

David A. Dickey
William Neal Reynolds Professor
North Carolina State University, Raleigh, NC, USA

One of the most basic and useful of the time series models is the order 1 (1 lag) autoregressive model, denoted $AR(1)$ and given by $Y_t - \mu = \rho(Y_{t-1} - \mu) + e_t$ where Y_t is the observation at time t, μ is the long run mean of the time series and e_t is an independent sequence of random variables. We use this venerable model to illustrate the Dickey–Fuller test then mention that the results extend to a broader collection of models.

When written as $Y_t = \mu(1-\rho) + \rho Y_{t-1} + e_t$, or more convincingly as $Y_t = \lambda + \rho Y_{t-1} + e_t$, with e independent and identically distributed as $N(0, \sigma^2)$, the $AR(1)$ model looks like a regression with errors satisfying the usual assumptions. Indeed the least squares estimators of the coefficients are asymptotically unbiased and normally distributed under one key condition, namely that the true ρ satisfies $|\rho| < 1$. It appears that this assumption is quite often violated. Many prominent time series appear to have $\rho = 1$, in which case $Y_t - \mu = \rho(Y_{t-1} - \mu) + e_t$ becomes $Y_t = Y_{t-1} + e_t$ or $Y_t - Y_{t-1} = e_t$. That is to say there are many series whose first differences $Y_t - Y_{t-1}$ seem to form a sequence of independent shocks, as the e's are often called. For estimation of the parameters, only mild assumptions on e are required. Normality is not necessary as long as the sample size is reasonable and $|\rho| < 1$. Two things are worth noting. The first is that the long term mean μ has dropped from this equation, that is, there is no tendency to move toward a long term mean, no mean reversion. A series that is high at time t is just as likely to move up as to move down. There is nothing to be gained by assessing the distance from the historic mean. The second point is a recommendation to write the model as $Y_t - \mu = \rho(Y_{t-1} - \mu) + e_t$. The representation $Y_t = \lambda + \rho Y_{t-1} + e_t$ masks the relationship $\lambda = \mu(1-\rho)$. The uninformed analyst may not realize that the intercept disappears when $\rho = 1$. The case when $|\rho| < 1$ falls into the category of stationary autoregressive models whereas the model with $\rho = 1$ is an example of a nonstationary model. For reasons discussed later, the $\rho = 1$ case is also called a unit root model and sometimes an integrated model.

There are many series of economic interest for which a test of the hypothesis $H_0 : \rho = 1$ is desired. The interest in such tests can be explained with a hypothetical scenario. Suppose an autoregressive order 1 model $Y_t = 5.00 + 0.8(Y_{t-1} - 5.00) + e_t$ describes the price of a stock over some time index t. Now if the current price is $Y = 5.50$ then the expected value of the next observation is $5.00 + 0.8(5.50 - 5.00) + 0 = 5.40$, closer than the current 5.50 to the mean 5. Likewise we'd predict a rise to 4.60, were we currently at 4.50, that is, the strategy of selling high and buying low would make money in the long run. The ability to forecast the direction of the stock market would cast doubt on the assumption that it was a quickly responding market with fully informed participants so we think that 0.8 is an estimate of $\rho = 1$ and expect that the null hypothesis will not be rejected. In contrast, we might expect the ratio of short term to long term bond yields or interest rates to be stable, that is, we expect some long term equilibrium ratio and we expect movement of the ratio toward that number in the long run. Here we expect to *reject* $H_0 : \rho = 1$ in favor of $H_1 : \rho < 1$. On a logarithmic scale, we want to test $Y_t = \log(\text{short term rate}) - \log(\text{long term rate})$ for stationarity. When two individual series are nonstationary but their difference is stationary, the series are said to be "cointegrated." Some of the most interesting hypothesis tests in economics concern constructed variables, as does this test for cointegration.

Having established an interest in testing the hypothesis $H_0 : \rho = 1$, we return to the statement that the least squares estimator is not normal under this hypothesis. How can this be when the errors satisfy all the usual regression assumptions? The answer lies in the assumption on the independent (X) variables. In usual regression theory, these X variables are assumed fixed and measured without error. Extensions to random X cases typically involve a conditioning argument, that is, we consider what would happen if we only looked at repeated samples that have the same X values that we observed. There is no way to make that argument here, where the X variables are just the Y (dependent) variables at a different time.

When $\rho = 1$ and μ is known to be 0, the theory of weak convergence as in Billingsley (1968) implies that the least squares estimator $\hat{\rho}$ satisfies a limit in terms of a standard Wiener process $W(t)$:

$$n(\hat{\rho} - 1) = \frac{\sum_{t=2}^{n} Y_{t-1} e_t}{\sum_{t=2}^{n} Y_{t-1}^2} \xrightarrow{L} \frac{1}{2} \frac{W^2(1) - 1}{\int_0^1 W^2(t)dt}.$$

This is known as the "normalized bias" statistic. This is an appealing mathematical representation, but is nevertheless a random quantity. Knowing this expression does not lead directly to a distribution and resulting table of critical values for testing. With simulation, empirical percentiles for various sample sizes n can be computed, however it is not clear how large an n will get the simulations close

enough to the limit so that no further n values need be considered. In addition to running rather large simulations as just mentioned, Dickey and Fuller (1979) calculated the eigenvalues of the limits of the numerator and denominator quadratic forms in the above normalized bias and from those directly simulated the limit distribution. They were then able to see how large an n was required to get close to that limit. This resulted in critical values for the normalized bias $n(\hat{\rho}-1)$ for finite n and the limit. Statistical programs that do least squares regression also produce t tests. Dickey and Fuller looked at the t statistic associated with the 0 mean regression. Its limit is the same as that of $\left(\sigma^2\sum_{t=2}^{n}Y_{t-1}^2\right)^{-1/2}\sum_{t=2}^{n}Y_{t-1}e_t$. They gave percentiles for that statistic as well and called it τ to indicate that it did not behave like the usual t, even in the limit.

At this point two distributions, one for the normalized bias and one for τ have been discussed, but the underlying assumption that μ is 0 makes this result of little practical interest. Suppose you have a stationary ($|\rho| < 1$) $AR(1)$ time series whose values range between 1000 and 1100. With no intercept in the model, the estimator $\frac{\sum_{t=2}^{n}Y_{t-1}Y_t}{\sum_{t=2}^{n}Y_{t-1}^2} \approx 1-0.5\frac{\sum_{t=2}^{n}(Y_t-Y_{t-1})^2}{\sum_{t=2}^{n}Y_{t-1}^2}$, where the sum of squared differences can be no greater than $(n-1)(100^2)$ and the denominator is no less than $(n-1)(1000^2)$ so the estimate of ρ is between 0.995 and 1. Simulations (Dickey 1984) verify that, as the algebra suggests, stationary series with positive means give tests that rarely reject the (false) null hypothesis when no intercept is included in the regression. It thus becomes of interest to study the mean adjusted estimator $n(\hat{\rho}_\mu-1) = \frac{\sum_{t=2}^{n}(Y_{t-1}-\overline{Y})e_t}{\sum_{t=2}^{n}(Y_{t-1}-\overline{Y})^2}$, where $\overline{Y} = \sum_{t=1}^{n}\frac{Y_t}{n}$, and its associated studentized statistic τ_μ. An asymptotically equivalent estimator and τ_μ statistic are obtained by regressing Y_t on an intercept and Y_{t-1}. Dickey and Fuller show that the addition of an intercept changes even the asymptotic distribution rather dramatically. This can be demonstrated either through their quadratic form approach or the Wiener Process limit representation.

Applying the same logic to a series that appears to be moving upward or downward at a rather steady long run rate, a fair test for "trend stationarity" as the alternative hypothesis is called, one that gives a chance to both the null and alternative hypothesis, must use a model that can capture the trend under both hypotheses. A logical candidate is a model with linear trend and $AR(1)$ error. The model can be expressed in two ways:

$$Y_t-\alpha-\beta t=\rho(Y_{t-1}-\alpha-\beta(t-1))+e_t$$

or

$$Y_t=\alpha+\beta t+Z_t,\qquad Z_t=\rho Z_{t-1}+e_t.$$

The first of these models most easily demonstrates that under our null hypothesis $\rho = 1$, we have

$$Y_t=\beta+Y_{t-1}+e_t$$

a random walk with drift β. The former slope is now called a drift and algebraically appears as an intercept term but in reality, it represents the long term monotone trend regardless of the ρ value. It is, after all, the mean change in Y_t per unit change in t. Its inclusion accomplishes our purpose of allowing both the null and alternative model to capture the long term trend, thus separating the alternative hypothesis case of stationary errors around a linear trend from the null hypothesis case of random walk with drift. The alternative hypothesis defines the concept of trend stationarity. The model can be fit by regressing Y_t on an intercept, t, and Y_{t-1} or, in an asymptotically equivalent way, by regressing Y_t on 1 and t, obtaining residuals r_t, and then regressing r_t on r_{t-1} with no intercept. This second approach is motivated by the second form of the model above. As with the mean adjusted case, this results in further changes in the distributions of the estimate and its studentized statistic τ_τ, even asymptotically. Since this last model subsumes the others, why not always use an intercept and time trend? The answer lies in a substantial loss of power when unnecessary parameters are added to the model.

We now summarize a few numerical results from the literature. Tables of critical values for the normalized bias and studentized statistics are developed in Dickey and Fuller (1979, 1981) and reported in detail in Fuller (1996). The 1% and 5% critical values from Fuller's text are reported below for the smallest tabulated sample size $n = 25$ and for the limit distribution.

The studentized statistics (top 3 rows of Table 1) have the more stable percentiles over different sample sizes, especially at the 5% level and are more often used in practice. Dickey (1984, Table 6.1) shows what happens when there are nonzero intercept or trend parameters in the generated data that are not in the model. For even modest values of these omitted parameters, there are almost no rejections of the (false) null hypothesis when $|\rho| < 1$. In Table 6.2 of that paper we find empirical powers for 2000 replicates of these tests for data generated by $Y_t = \rho Y_{t-1} + e_t$, $t = 1, 2, \ldots, 50$. Substantial loss of power results when unneeded intercept or trend parameters are added to the model as in columns 3,4,6, and 7 of Table 2, which displays a few entries from the power study.

Clearly inclusion of either too many or too few deterministic terms in the model can cause power loss. Failing

Dickey-Fuller Tests. Table 1 Left tailed unit root test percentiles

Model	1%, $n = 25$	1%, limit	5%, $n = 25$	5%, limit
No intercept τ	−2.65	−2.58	−1.95	−1.95
Intercept τ_μ	−3.75	−3.42	−2.99	−2.86
Linear trend τ_τ	−4.38	−3.96	−3.60	−3.41
No intercept $n(\widehat{\rho}-1)$	−11.8	−13.7	−7.3	−8.1
Intercept $n(\widehat{\rho}_\mu-1)$	−17.2	−20.6	−12.5	−14.1
Linear trend $n(\widehat{\rho}_\tau-1)$	−22.5	−29.4	−17.9	21.7

Dickey-Fuller Tests. Table 2 Some empirical powers

True ρ	τ	τ_μ	τ_τ	$n(\widehat{\rho}-1)$	$n(\widehat{\rho}_\mu-1)$	$n(\widehat{\rho}_\tau-1)$
1.00	0.05	0.05	0.05	0.06	0.05	0.04
0.90	0.31	0.11	0.08	0.31	0.19	0.10
0.80	0.77	0.30	0.20	0.77	0.46	0.24
0.70	0.97	0.61	0.39	0.97	0.77	0.48
0.50	1.00	0.98	0.87	1.00	1.00	0.91

to reject unit roots when a model without trend is fit to trending data may do less damage than rejecting. Forecasting future values to be the same as the present value may be less damaging than giving mean reverting forecasts that go in the opposite direction of the trend.

The autoregressive order 1 model is common but is not rich enough to fit all time series. Consider an autoregressive model of order 3,

$$Y_t - \mu = 1.1(Y_{t-1} - \mu) - 0.3(Y_{t-2} - \mu) + 0.2(Y_{t-3} - \mu) + e_t$$

or, written in terms of the backshift operator $B(Y_t) = Y_{t-1}$,

$$(1 - 1.1B + 0.3B^2 - 0.2B^3)(Y_t - \mu) = e_t$$

Because $1 - 1.1 + 0.3 - 0.2 = 0$ and μ is constant over time, μ drops out of the equation. The factored form $(1 + 0.4B)(1 - 0.5B)(1 - B)(Y_t - \mu) = e_t$, shows that the first differences, $(1 - B)(Y_t - \mu) = Y_t - Y_{t-1}$, form a stationary process $(1 + 0.4B)(1 - 0.5B)(Y_t - Y_{t-1}) = e_t$. The characteristic polynomial is the backshift polynomial with B considered an algebraic variable. An autoregressive series is stationary if the roots of its characteristic polynomial all exceed 1 in magnitude. The order 3 autoregressive example here has a root 1, a "unit root." It is therefore nonstationary but its differences form a stationary (roots $1/0.5 = 2$ and $-1/0.40 = -2.25$) autoregressive order 2 series.

Dickey and Fuller proved that if the first difference $Y_t - Y_{t-1}$ of a unit root autoregressive process is regressed on the lagged level Y_{t-1} of the process (and possibly an intercept and trend) and enough lagged differences to produce the appropriate lag structure, the studentized statistic on the lag level term will have the same limit distribution as in the autoregressive order 1 case. The tables discussed above can thus be used. In contrast, the distribution of the normalized bias is altered by the presence of the lagged differences which is one reason the studentized statistics are preferred despite the power advantage of the normalized bias test in the simplest $AR(1)$ cases.

The characteristic polynomial here has the form $(1-a_1B)(1-a_2B)(1-a_3B)$ with roots $1/a_i$, $i = 1, 2, 3$. If (and only if) $B = 1$ is a root, we will have $(1-a_1)(1-a_2)(1-a_3) = 0$. These expressions are algebraically equivalent:

(I) $(1 - a_1B)(1 - a_2B)(1 - a_3B)(Y_t - \mu) = e_t$

(II) $Y_t - \mu = (a_1 + a_2 + a_3)(Y_{t-1} - \mu) - (a_1a_2 + a_2a_3 + a_1a_3)(Y_{t-2} - \mu) + a_1a_2a_3(Y_{t-3} - \mu) + e_t$

(III) $Y_t - Y_{t-1} = -(1-(a_1+a_2+a_3)+(a_1a_2+a_2a_3+a_1a_3)-a_1a_2a_3)(Y_{t-1} - \mu) + (a_1a_2 + a_2a_3 + a_1a_3 - a_1a_2a_3)(Y_{t-1} - Y_{t-2}) + a_1a_2a_3(Y_{t-2} - Y_{t-3}) + e_t$

Expression (III) is the motivation for the regression approach. The coefficient on $(Y_{t-1} - \mu)$ is $-(1 - a_1)(1 - a_2)(1 - a_3)$, the negative of the characteristic polynomial evaluated at $B = 1$ so a unit root (any $a_i = 1$) makes this coefficient 0 and removes μ from the model. There is, then, no mean reversion. The lagged differences are known as augmenting terms and the resulting tests are known as *ADF* or augmented Dickey Fuller tests. Focusing on one of the three factors in $(1-a_1)(1-a_2)(1-a_3)$, clearly it is multiplied by the other two. Under the null hypothesis that one a_i value is 1, the coefficient on Y_{t-1} in (III) is an estimate of 0 but is scaled by the other two factors. This affects the distribution of the coefficient, but not that of its τ statistic. As in any regression, division of the estimate by its standard error makes τ a self normalizing statistic. Some estimate of those other two factors is needed to scale the normalized bias so that it is asymptotically equivalent to the $AR(1)$ case.

Other results worth mentioning include the work of S. E. Said (Said and Dickey 1984). Said showed that even if a time series contained an invertible moving average part, one could use an augmented autoregressive model. Given a number of augmenting lagged differences that increases with n at the proper rate, the studentized statistic has the same limit distribution as in the autoregressive

order 1 model. One can simply fit a model with augmenting lagged differences to test for unit roots. Park and Fuller (1995) use a weighted regression resulting in considerable power improvement. Elliott et al. (1996) incorporate a generalized least squares trend and intercept estimator that results in a test with similar power increase. Pantula et al. (1994) compare a test of Gonzalez-Farias, based on maximization of the exact stationary likelihood, and these other two. They all have competitive power. Phillips and his students have made many contributions in the area. Among them, Phillips and Perron (1988) suggest a different adjustment for the higher order models and show that the tables described herein can be used for the test under weak assumptions on the errors.

About the Author

Professor David Dickey, with Wayne Fuller, developed a *Unit Root Test*, in 1979. His paper "opened-up a new way research agenda in time series econometrics, the investigation and identification of nonstationary processes" (Lex Oxley, The "Top 10" Papers in Econometrics, 1980–2000). He was Program Chair, Business and Economics section American Statistical Association (2007). He is an elected Fellow (2000) of the American Statistical Association. In addition to writing four books, Dr. Dickey has been published in numerous papers and has given over 50 presentations at a variety of professional events and organizations. He has also been recognized as a member of the Academy of Outstanding Teachers at North Carolina State University.

Cross References

►Autocorrelation in Regression
►Bayesian Approach of the Unit Root Test
►Seasonal Integration and Cointegration in Economic Time Series

References and Further Reading

Billingsley P (1968) Convergence of probability measures. Wiley, New York
Dickey DA, Fuller WA (1979) Distribution of the estimators for autoregressive time series with a unit root. J Am Stat Assoc 74:427–431
Dickey DA, Fuller WA (1981) Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49:1057–1072
Dickey DA (1984) Power of unit root tests. In: Proceedings of the business and economic statistics section. American Statistical Association, Philadelphia, Washington, DC, pp 489–493
Elliott G, Rothenberg TJ, Stock JH (1996) Efficient tests for an autoregressive unit root. Econometrica 64:813–836
Fuller WA (1996) Introduction to statistical time series. Wiley, New York
Pantula SG, Gonzalez-Farias G, Fuller WA (1994) A comparison of unit root criteria. J Bus Econ Stat 13:449–459
Park HJ, Fuller WA (1995) Alternative estimators and unit root tests for the autoregressive process. J Time Ser Anal 15:415–429
Phillips PCB, Perron P (1988) Testing for a unit root in time series regression. Biometrika 75:335–346
Said SE, Dickey DA (1984) Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika 71:599–607

Discriminant Analysis: An Overview

Barry J. Babin
Max P. Watson, Jr. Professor and Department Head
Louisiana Tech University, Ruston, LA, USA

Discriminant analysis is a multivariate technique that helps the user explain group membership as a function of multiple independent variables. In particular, discriminant analysis is appropriate when the user is faced with a situation involving a categorical (nominal or ordinal) dependent variable and independent variables that are primarily continuous (interval or ratio) although categorical independent variables can be included under some conditions. This would typically involve dummy coding of the categorical variables. In many cases, a dichotomous dependent variable is employed although it can be multichotomous.

The variate takes the familiar form:

$$V = w_1x_1 + w_2x_2 + \ldots + w_ix_i + e_i$$

where x represents the independent variables, w represents an empirically derived parameter estimate that weights the extent to which the independent variable determines group membership, and V represents the resulting variate value which determines group membership. The variate takes on the form of a Z score. This metric value is ultimately the key indicator of which group an observation should belong to based on the corresponding set of independent variable values. Groups are formed based on these Z-values by determining a cutting score which is used to separate observations into two groups. Discriminant analysis applies the fundamental logic that observations that produce similar Z-values (based on values for the set of independent variables) should be grouped together and observations with dissimilar Z-values are different and should not be grouped together. Thus, for a dichotomous dependent variable, the cutting score can be thought of as:

$$Z_{CS} = \frac{N_AZ_B + N_BZ_A}{N_A + N_B}$$

Where, Z_{cs} is the cutting score and N represents the sample size in groups A and B, respectively. Z_A and Z_B represent the centroids of groups A and B formed by splitting all observations into groups using the cutting score.

Considering the dichotomous case, the observed values for each independent variable can be used to predict whether each observed value for a dependent variable is a "1" or "0." The values for the independent variables are multiplied by derived parameter estimates just as in multiple regression. In other words, the analysis determines which group an observation should belong to based upon the calculated value? A classification matrix is created which shows how accurately the model can actually place these observations into categories. This matrix is formed by taking a conditional frequency of predicted group membership against actual observed group membership. The contingency table takes a form as follows:

Actual Group Membership	Predicted Group Membership	
	Group 1	Group 2
Group 1	Correctly classified	Incorrectly classified
Group 2	Incorrectly classified	Correctly classified

The percent of observations correctly classified becomes one measure of a discriminant function's effectiveness. Consider that if group 1 had two times as many respondents as group 2, correct classification could be 67 percent simply by placing all observations into group 1. Therefore, the discriminant analysis should be able to produce better than 67 percent correct classifications if it is to be useful. In fact, a t-test is available to determine the significance level for classification accuracy:

$$t = \frac{p - g}{\sqrt{\frac{g(1.0 - g)}{N}}}$$

Where, p = percent correctly classified, N = sample size, and g represents the proportion of observations actually observed in the largest group (0.67 in the example above and 0.5 if group sizes are equal, for instance).

The pattern of discriminant weights or discriminant loadings determines which variables most predict or explain group membership (i.e., via the Z-values). Discriminant loadings are sometimes called structure correlations and they are directly analogous to factor loadings in the multivariate technique known as factor analysis (see ►Factor Analysis and Latent Variable Modeling). They represent the correlation between an observation and the discriminant function (i.e., variate). Thus, the relative magnitude of a loading shows how influential a variable is to the grouping variable. If causal terms are appropriate, it would show relatively speaking, how much a variable caused observations to be in one group or another. The number of discriminant factors that can be recovered is equal to one less than the number of groups so that for a dichotomous dependent variable, only one discriminant function, or factor, is possible. In the three group case two factors are possible.

Thus, two key results are (1) how accurately a discriminant function can classify observations and (2) which variables are most responsible for the classifying. As the cross-classification suggests that the model predicts membership at better than a chance rate, the discriminant function has validity. As loadings are greater in magnitude, they are predictive of group membership. In a very real way, the mathematics of discriminant analysis mirrors that of MANOVA, another multivariate technique. Both MANOVA and discriminant analysis will each produce a Wilkes Λ that indicates whether or not the observed group membership accounts for differences in the means of the independent variables. Discriminant validity is often used in combination with other techniques like cluster analysis (see ►Cluster Analysis: An Introduction) and also is conceptually similar to techniques like ►logistic regression that also predict categorical dependent variables.

Here, a simple illustration of discriminant analysis is described. The observations represent records of academic journal submissions forming a sample of the larger population of submissions. The dependent variable is whether or not the submission was rejected. Thus, there are two groups of observations. One hundred and eighteen papers were rejected and 40 were not rejected. Group membership constitutes the dependent variable and was coded 0 = Rejected and 1 = Not Rejected. The independent variables represent the paper's review scores based on:

(a) topical relevance,
(b) theoretical development,
(c) research approach,
(d) analytical competence,
(e) writing quality,
(f) contribution potential.

Each independent variable was scored on a 5 point scale with higher scores being better. Here is a brief overview of the results. The Wilkes' Λ is 56.9 which suggests the means for the independent variables differ significantly based on group membership ($p < .001$). The resulting cross-classification matrix is shown below:

Actual Group Membership	Predicted Group Membership	
	Rejected	Not Rejected
Rejected	109	9
Not Rejected	16	24

133 of 158 observations, or 84.2 percent, are correctly classified. This can be compared to the 75 percent of observations that would be correctly classified if all 158 observations were predicted to be rejected. The discriminant analysis improves predictability significantly over chance ($p < .001$). Additionally, the discriminant loadings are as follows:

Variable	Function 1
Theoretical development	0.83
Research methods	0.76
Writing quality	0.71
Analytical competence	0.62
Contribution potential	0.62
Topical relevance	0.37

Thus, theoretical development is the most influential independent variable helping explain why papers get rejected from this particular journal. In contrast, topical relevance does very little to distinguish papers that are rejected from others. Perhaps the reason for this is that most submissions receive similar high scores for relevance. Practically, the results suggest that stronger theory is a route to avoiding rejection.

As you can see, discriminant analysis can be a very useful tool for explaining why observations end up in one group or another (Hair et al. 2010). Some applications of discriminant analysis include a study predicting cohort group membership as the dependent variable with personal value scores as a independent variables (Noble and Schewe 2003), a study trying to explain differences between small firms and firms that grow into large firms (Culpan 1989), and a study explaining groups of consumers based on how they respond to poor performance in a retail setting. This idea involves explaining why someone might be an irate consumer, an active consumer or a passive consumer, for instance, using a variety of demographic, lifestyle and situational independent variables (Singh 1990).

About the Author

Professor Babin is Past President, The Academy of Marketing Science (2006–2008) and Past President of The Society for Marketing Advances (2000–2001). He is a Distinguished Fellow of the Academy of Marketing Science and the Society for Marketing Advances. He was awarded the Omerre DeSerres Award for Outstanding Research in Retailing. Dr. Babin is a co-author (with J Hair, R. Anderson, R. Tatham, and W. Black) of the well known text *Multivariate Data Analysis* (Prentice Hall).

Cross References

►Canonical Correlation Analysis
►Discriminant Analysis: Issues and Problems
►Multivariate Analysis of Variance (MANOVA)
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Statistics: An Overview

References and Further Reading

Culpan R (1989) Export behavior of firms: relevance of firm size. J Bus Res 18:207–218 (March)
Hair J, Black WC, Babin BJ, Anderson R (2010) Multivariate data analysis, 7th edn. Prentice Hall, Upper Saddle River, NJ
Noble SM, Schewe CD (2003) Cohort segmentation: an exploration of its validity. J Bus Res 56:979–987 (December)
Singh J (1990) A typology of consumer dissatisfaction styles. J Retail 60:57–99 (Spring)

Discriminant Analysis: Issues and Problems

Carl J. Huberty
Professor Emeritus
University of Georgia, Athens, GA, USA

Introduction

It was in the mid 1930s when Sir Ronald A. Fisher (1890–1962) formally introduced the notion of "discriminant analysis" (DA) in writing. His introduction involved prediction of group membership in a two-group context – a predictive discriminant analysis (PDA). The notion of "discriminant analysis" became of interest to researchers in various areas of study in the 1950s and 1960s (e.g., Cooley and Lohnes 1962). That is when the variant which may be termed "descriptive discriminant analysis" (DDA) "caught on."

PDA Versus DDA

The mixing of PDA and DDA is fairly common in many books and journal articles. The distinction of PDA from

Discriminant Analysis: Issues and Problems. Table 1 PDA versus DDA

	PDA	DDA
Research context	Prediction of group membership	Description of group separation
Variable roles		
Predictor(s)	Response variables	Grouping variable(s)
Criterion(ia)	Grouping variable	Response variables
Response variable set	Hodgepodge	System
Response variable composite	LCF/QCF	LDF
Number of composites	k	$\min(p, k-1)$
Preliminary analysis concerns		
Equality of covariance matrices	Yes	Yes
MANOVA	No	Yes
Analysis aspects of typical concern		
Variable construct(s)	No	Yes(!)
Response variable deletion	Yes (!)	Maybe
Response variable ordering	Yes	Yes
Criterion for variable deletion/ordering	Classification accuracy	Group separation
Research purpose	Practical/theoretical	Theoretical

DDA is, to the current writer, fairly important for description, interpretation, and reporting purposes. A PDA is used to determine a "rule" for predicting membership in one of k groups of analysis units based on measures on p predictor variables. The rule, then, consists of k composites of the p predictors. These composites may be of linear form (if the k $p \times p$ covariance matrices are judged to be approximately equal) or of quadratic form (if the k covariance matrices are clearly unequal). For the former, one has k linear classification functions (LCFs). For the latter, one has k quadratic classification functions (QCFs). For details, see Huberty and Olejnik (2006, Chap. 13).

A DDA is considered in the context of comparing k group mean vectors of the p outcome variables. A DDA would be applicable when a ►multivariate analysis of variance (MANOVA) is conducted and the k mean vectors of the p outcome variables are judged to be unequal. [In a MANOVA context, it is assumed that the k covariance matrices are "in the same ballpark." See Huberty and Olejnik (2006, Chap. 3).] One obtains a set of $\min(p, k-1)$ linear discriminant functions (LDFs). These linear composites comprise the essence of a DDA. [Such composites are analogous to "factors" in a factor analysis.]

An issue/problem in DA is the lack of distinction between PDA and DDA. This lack results in the misuse of terms, and some misinterpretations of computer output. [The computer output often leads to misuse of terms.] See Table 1 for a summary of PDA versus DDA. In sum, it is classification (PDA) versus separation (DDA).

Terms

It should be clear from the stated purpose of an empirical study involving $k \geq 2$ groups of analysis units and $p \geq 2$ attributes/variables pertaining to the units, whether one is interested in group membership prediction or in group comparison description. For the former analysis (PDA), one would be interested in prediction, and, therefore, related term use should be utilized. Similarly for applications of DDA. Issues in term use in PDA and in DDA are reviewed by Huberty and Olejnik (2006, Chap. 22).

More specifically, proper use of terms in reporting the use, and results, of a PDA and DDA is given by Huberty and Olejnik (2006, Chaps. 7 and 20). Also, reliance on terms found in computer output is *not* strongly recommended.

More Issues and Problems

Five issues in PDA are discussed by Huberty and Olejnik (2006, Chap. 22): (1) linear versus quadratic classification

rules; (2) nonnormal classification rules; (3) prior probabilities; (4) misclassification costs; and (5) hit-rate estimation. Also discussed in that chapter are three issues in DDA: (1) stepwise analyses (yuk!); (2) standardized LDF weights versus structure *r*'s; and (3) data-based structure. All of these issues involve the use of researcher judgment – of course, consultation with respected "experts" would help in making some judgment calls.

There are, also, some additional problems in the use of both PDA and DDA – see Huberty and Olejnik (2006, Chap. 23). Three of these problems involve missing data, outliers, and misclassification costs. A detailed discussion of these problems will not be presented herein. With regard to the first two problems, a recommendation is to "look at your data" (via graphs).

About the Author

Dr. C. J. Huberty is a Professor Emeritus, Department of Research, Measurement, and Statistics, University of Georgia, USA. He served UGA from 1969 to 2002, and was Department Head (1992–1999). He was given the College of Education Faculty Research Award in 1998, and the Career Center Award in 2002. He was also given recognition by being nominated for 27 national academic awards. In 2004, he was recognized as a Fulbright Senior Specialist. He was an ASA member for 32 years. He authored two books: *Applied Discriminant Analysis* (Wiley 1994) and *Applied MANOVA and Discriminant Analysis* (Wiley, 2006, with Stephen Olejnik). He was the sole/senior author of 15 invited book chapters, 58 journal articles, 13 book reviews, and 93 papers presented at professional meetings. He conducted 42 workshops on multivariate statistics at professional meetings in Australia, Belgium, Egypt, and USA. He served as a statistical consultant for 18 grants. Dr. Huberty served on 12 journal editorial boards, including the founding board of the *Journal of Educational Statistics*. Finally, he sponsored academic visitors from Canada (2), Egypt (16), Iran (1), Italy (1), Japan (1), South Korea (1), St. Lucia (1), and Sweden (1).

Cross References

►Canonical Correlation Analysis
►Discriminant Analysis: An Overview
►Multivariate Analysis of Variance (MANOVA)
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis

References and Further Reading

Cooley WW, Lohnes PR (1962) Multivariate procedures for the behavioral sciences. Wiley, New York
Huberty CJ (2002) Discriminant analysis. In: Meij J (ed) Dealing with the data flood. Study Centre for Technology Trends, The Netherlands, pp 585–600
Huberty CJ (2005) Discriminant analysis. In: Everitt BS, Howell DC (eds) Encyclopedia of statistics in behavioral science, vol 1. Wiley, Chichester, pp 499–505
Huberty CJ, Hussein MH (2003) Some problems in reporting use of discriminant analyses. J Exp Educ 71:177–191
Huberty CJ, Olejnik S (2006) Applied MANOVA and discriminant analysis. Wiley, Hoboken

Dispersion Models

Bent Jørgensen
Professor
University of Southern Denmark, Odense M, Denmark

Introduction

A *dispersion model,* denoted $Y \sim \mathrm{DM}(\mu, \sigma^2)$, is a two-parameter family of distributions with probability density functions on $\mathbb{R}$ of the form

$$f(y;\mu,\sigma^2) = a(y;\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\}. \qquad (1)$$

Here μ and σ^2 are real parameters with domain $(\mu,\sigma^2) \in \Omega \times \mathbb{R}_+$ (Ω being an interval), called the *position* and *dispersion* parameters, respectively. Also a and d are suitable functions such that (1) is a probability density function for all parameter values. In particular, d is assumed to be a *unit deviance*, satisfying $d(\mu;\mu) = 0$ for $\mu \in \Omega$ and $d(y;\mu) > 0$ for $y \neq \mu$. Dispersion models were introduced by Sweeting (1981) and Jørgensen (1983; 1987b) who extended the analysis of deviance for ►generalized linear models in the sense of Nelder and Wedderburn (1972) to non-linear regression models with error distribution $\mathrm{DM}(\mu,\sigma^2)$.

In many cases, the unit deviance d is *regular,* meaning that it is twice continuously differentiable and $\partial^2 d/\partial\mu^2(\mu;\mu) > 0$ for $\mu \in \Omega$. We then define the *unit variance function* for $\mu \in \Omega$ by

$$V(\mu) = \frac{2}{\frac{\partial^2 d}{\partial\mu^2}(\mu;\mu)}. \qquad (2)$$

For example, the normal distribution $\mathrm{N}(\mu,\sigma^2)$ corresponds to the unit deviance $d(y;\mu) = (y-\mu)^2$ on $\mathbb{R}^2$, with unit variance function $V(\mu) \equiv 1$ and $a(y;\sigma^2) = \left(2\pi\sigma^2\right)^{-1/2}$. Note that for $0 < \rho < 2$ the unit deviance $d(y;\mu) = |y-\mu|^\rho$, with $a(y;\sigma^2)$ constant (not depending on y) provides an example of a dispersion model where d is not regular.

The *renormalized saddlepoint approximation* for a dispersion model with regular unit deviance d is defined for $y \in \Omega$ by

$$f(y;\mu,\sigma^2) \sim a_0(\mu,\sigma^2)V^{-\frac{1}{2}}(y)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\}, \quad (3)$$

where $a_0(\mu,\sigma^2)$ is a normalizing constant, making the right-hand side of (3) a probability density function on Ω with respect to Lebesgue measure. The ordinary saddlepoint approximation is obtained by the asymptotic approximation $a_0(\mu,\sigma^2) \sim (2\pi\sigma^2)^{-1/2}$ as $\sigma^2 \downarrow 0$, and correspondingly, the distribution $\mathrm{DM}(\mu,\sigma^2)$ is asymptotically normal $\mathrm{N}\{\mu,\sigma^2V(\mu)\}$ for σ^2 small, so that μ and $\sigma^2V(\mu)$ are the asymptotic mean and variance of Y, respectively. In this way, dispersion models have properties that resemble those of the normal distribution. In particular the unit deviance $d(y;\mu)$ is often a measure of squared distance between y and μ.

There are two main classes of dispersion models, namely *proper dispersion models* (PD) and *reproductive exponential dispersion models* (ED), as defined below. The class of *additive exponential dispersion models* (ED*) is not of the dispersion model form (1), but is closely related to reproductive exponential dispersion models. These three types of models cover a comprehensive range of non-normal distributions, and as shown in Table 1 they include many standard statistical families as special cases. The exponential and Poisson families are examples of natural exponential families, which correspond to exponential dispersion models with $\sigma^2 = 1$.

Proper Dispersion Models

Jørgensen (1997a) proposed a special case of (1), called a *proper dispersion model*, which corresponds to a density of the form

$$f(y;\mu,\sigma^2) = a(\sigma^2)V^{-\frac{1}{2}}(y)\exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\}, \quad (4)$$

where d is a given regular unit deviance with unit variance function $V(\mu)$. We denote this model by $Y \sim \mathrm{PD}(\mu,\sigma^2)$. Note that the unit deviance d characterizes the proper dispersion model (4), because the unit variance function (2) is a function of d, while $a(\sigma^2)$, in turn, is a normalizing constant. It also follows that the normalizing constant of (3) is $a_0(\mu,\sigma^2) = a(\sigma^2)$, making the the two sides of (3) identical in this case. Conditions under which a given unit deviance gives rise to a proper dispersion model, and the connection with exactness of Barndorff-Nielsen's p^*-formula, are discussed by Jørgensen (1997b, Chap. 5).

Many proper dispersion models are transformation models when σ^2 is known. For example, unit deviances of the form $d(y;\mu) = h(y-\mu)$ on $\Omega = \mathbb{R}$, for a suitable function h, give rise to so-called *location-dispersion models*, where the unit variance function is constant. A specific example is given by the unit deviance $d(y;\mu) = \log\{1+(y-\mu)^2\}$, which corresponds to the Student t location family. Another important case corresponds to unit deviances of the form $d(y;\mu) = h(y/\mu)$ on $\Omega = \mathbb{R}_+$, which give rise to so-called *scale-dispersion models*. For example, Jørgensen (1997b, Chap. 5) showed that the generalized ▶inverse Gaussian distribution is a family of scale-dispersion models with unit deviances

$$d_\beta(y;\mu) = 2\beta\log\frac{\mu}{y} + (1+\beta)\frac{y}{\mu} + (1-\beta)\frac{\mu}{y} - 2, \quad (5)$$

for $y,\mu > 0$, where $\beta \in [-1,1]$ is an additional parameter. In particular, the values $\beta = \pm 1$ correspond to the gamma and reciprocal gamma families, respectively, and $\beta = 0$ to the one-dimensional hyperboloid distribution. The von Mises distribution, corresponding to the unit deviance $d(y;\mu) = 2\{1-\cos(y-\mu)\}$ on $\Omega = (0,2\pi)$ is also a transformation model when σ^2 is known, corresponding to the group of additions modulo 2π.

The class of *simplex distributions* of Barndorff-Nielsen (1991) contains several proper dispersion models as special cases. The simplest example is given by the probability density function

$$f(y;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\{y(1-y)\}^{-3/2} \times \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2y(1-y)\mu^2(1-\mu)^2}\right\}, \quad (6)$$

where $y,\mu \in (0,1)$ and $\sigma^2 > 0$. The *transformed Leipnik distribution*, defined by the unit deviance

$$d(y;\mu) = \log\left\{1+\frac{(y-\mu)^2}{y(1-y)}\right\} \text{ for } y,\mu \in (0,1)$$

provides a further examples of proper dispersion models. This and the simplex distribution are examples of proper dispersion models that are not transformation models when σ^2 is known.

Exponential Dispersion Models

Exponential dispersion models are two-parameter extensions of natural exponential families. A *natural exponential family* has densities of the form

$$f(x;\theta) = c(x)\exp\{\theta x - \kappa(\theta)\}, \quad (7)$$

with respect to a suitable measure ν, usually Lebesgue or counting measure. Here $c(x)$ is a given function and $\kappa(\theta) = \log\int c(x)e^{\theta x}\,\nu(dx)$ is the corresponding *cumulant function*. The *canonical parameter* θ has domain Θ; the

Dispersion Models. Table 1 Main examples of univariate dispersion models

Data type	Support	Examples	Type
Real data	$\mathbb{R}$	Normal, generalized hyperbolic secant	ED
Positive data	$\mathbb{R}_+$	Exponential, gamma, inverse Gaussian	ED
Positive data with zeros	$\mathbb{R}_0 = [0, \infty)$	Compound Poisson distribution	ED
Proportions	$(0,1)$	Simplex, Leipnik distributions	PD
Directional data	$[0, 2\pi)$	von Mises distribution	PD
Count data	$\mathbb{N}_0 = \{0,1,2,\dots\}$	Poisson, negative binomial distributions	ED*
Binomial count data	$\{0,1,\dots,m\}$	Binomial distribution	ED*

largest interval for which $\kappa(\theta)$ is finite. The mean of a random variable X with distribution (7) is $\mu = \tau(\theta)$, where $\tau(\theta) = \kappa'(\theta)$ (a one-to-one function of θ), and the domain for μ is $\Omega = \tau(\text{int}\Theta)$, where $\text{int}\Theta$ denotes the interior of the domain Θ. If necessary, we define the value of μ by continuity at boundary points of Θ. Examples of natural exponential families include the exponential, geometric, Poisson, and logarithmic families.

The variance of X is $V(\mu) = \tau'\{\tau^{-1}(\mu)\}$, where V is known as the *variance function* of the family (7). This function characterizes (7) among all natural exponential families, see for example Morris (1982), and V is an important convergence and characterization tool for natural exponential families, see, e.g., Jørgensen (1997b, Chap. 2).

An *additive exponential dispersion model* is a two-parameter extension of natural exponential families with densities of the form

$$f^*(x;\theta,\lambda) = c^*(x;\lambda)\exp\{\theta x - \lambda\kappa(\theta)\}, \qquad (8)$$

where the *index parameter* λ has domain Λ consisting of those values of $\lambda > 0$ for which $\lambda\kappa(\theta)$ is a cumulant function of some function $c^*(x;\lambda)$. The mean and variance of a random variable X with distribution (8) are $\lambda\mu$ and $\lambda V(\mu)$, respectively. Many discrete distributions are additive exponential dispersion models, for example the binomial and negative binomial families. An additive exponential dispersion model is denoted $X \sim \text{ED}^*(\mu,\lambda)$.

A *reproductive exponential dispersion model* is defined by applying the transformation $Y = X/\lambda$ to (8), known as the *duality transformation*. The reproductive form is denoted $Y \sim \text{ED}(\mu,\sigma^2)$, corresponding to densities of the form

$$f(y;\theta,\lambda) = c(y;\lambda)\exp[\lambda\{\theta y - \kappa(\theta)\}] \qquad (9)$$

for a suitable function $c(y;\lambda)$. Here μ (with domain Ω) is the mean of Y, $\sigma^2 = 1/\lambda$ is the dispersion parameter, and the variance of Y is $\sigma^2 V(\mu)$. Reproductive exponential dispersion models were proposed by Tweedie (1947), and again by Nelder and Wedderburn (1972) as the class of error distributions for generalized linear models. The inverse duality transformation $X = \lambda Y$ takes (9) back into the form (8) so that each exponential dispersion model has, in principle, an additive as well as a reproductive form. Both (9) and (8) are natural exponential families when the index parameter λ is known.

To see the connection with dispersion models as defined by (1), we introduce the unit deviance corresponding to (9),

$$d(y;\mu) = 2\left[\sup_\theta\{\theta y - \kappa(\theta)\} - y\tau^{-1}(\mu) + \kappa\{\tau^{-1}(\mu)\}\right], \qquad (10)$$

which is a Kullback-Leibler distance, see Hastie (1987). Defining

$$a(y;\sigma^2) = c(y;\sigma^{-2})\exp\left[\sup_\theta\{\theta y - \kappa(\theta)\}\right],$$

we may write the density (9) in the dispersion model form (1). Consequently, reproductive exponential dispersion models form a sub-class of dispersion models, whereas additive exponential dispersion models are not in general of the form (9).

The overlap between exponential and proper dispersion models is small; in fact the normal, gamma and inverse Gaussian families are the only examples of exponential dispersion models that are also proper dispersion models.

An additive exponential dispersion model satisfies a convolution formula, defined as follows. If $X_1,\dots,X_n$ are independent random variables, and $X_i \sim \text{ED}^*(\mu,\lambda_i)$ with

$\lambda_i \in \Lambda$ for $i = 1, \ldots, n$, then the distribution of $X_+ = X_1 + \ldots + X_n$ is

$$X_+ \sim \mathrm{ED}^*(\mu, \lambda_1 + \ldots + \lambda_n), \tag{11}$$

where in fact $\lambda_1 + \cdots + \lambda_n \in \Lambda$. This is called the *additive property* of an additive exponential dispersion model.

An additive exponential dispersion model $\mathrm{ED}^*(\mu, \lambda)$ gives rise to a stochastic process $\{X(t) : t \in \Lambda \cup \{0\}\}$ with stationary and independent increments, called an *additive process*. This process is defined by assuming that $X(0) = 0$, along with the following distribution of the increments:

$$X(t+s) - X(t) \sim \mathrm{ED}^*(\mu, s),$$

for $s, t \in \Lambda$. An additive process with time domain $\Lambda = \mathbb{N}$ (the positive integers) is a random walk, the simplest example being the Bernoulli process. An additive process with time domain $\Lambda = \mathbb{R}_+$ (which requires the distribution to be infinitely divisible), is a Lévy process (see ►Lévy Processes), including examples such as Brownian motion with drift and the Poisson process (see ►Poisson Processes). Note, in particular, that the average of the process over the interval $(0, t)$ has distribution $X(t)/t \sim \mathrm{ED}(\mu, t^{-1})$, which follows by an application of the duality transformation.

A reproductive exponential dispersion model $\mathrm{ED}(\mu, \sigma^2)$ satisfies the following *reproductive property*, which follows as a corollary to (11). Assume that $Y_1, \ldots, Y_n$ are independent and $Y_i \sim \mathrm{ED}\left(\mu, \sigma^2/w_i\right)$ for $i = 1, \ldots, n$, where $w_1, \ldots, w_n$ are positive weights such that $w_i/\sigma^2 \in \Lambda$ for all i. Then, with $w_+ = w_1 + \ldots + w_n$,

$$\frac{1}{w_+}\sum_{i=1}^{n} w_i Y_i \sim \mathrm{ED}\left(\mu, \frac{\sigma^2}{w_+}\right). \tag{12}$$

Tweedie Exponential Dispersion Models

The class of *Tweedie models,* denoted $\mathrm{Tw}_p(\mu, \sigma^2)$, consist of reproductive exponential dispersion models corresponding to the unit variance functions $V(\mu) = \mu^p$, where p is a parameter with domain $(-\infty, 0] \cup [1, \infty]$ and Ω is defined in Table 2. Here we let $p = \infty$ correspond to the variance function $V(\mu) = e^{\beta\mu}$ for some $\beta \in \mathbb{R}$. These models were introduced independently by Tweedie (1984), Morris (1981), Hougaard (1986) and Bar-Lev and Enis (1986). As shown in Table 2, the Tweedie class contains several well-known families of distributions. For $p < 0$ or $p > 2$, the Tweedie models are natural exponential families generated by extreme or positive stable distributions with index $\alpha = (p-2)/(p-1)$. For $p \neq 0, 1, 2$, the unit deviance of the family $\mathrm{Tw}_p(\mu, \sigma^2)$ is given by

$$d_p(y; \mu) = 2\left[\frac{\{\max(y, 0)\}^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right],$$

where y belongs to the support, see Table 2.

The Tweedie models are characterized by the following scaling property:

$$b\,\mathrm{Tw}_p(\mu, \sigma^2) \sim \mathrm{Tw}_p(b\mu, b^{2-p}\sigma^2). \tag{13}$$

for all $b > 0$, see Jørgensen et al. (1994) and Jørgensen (1997b, Chap. 4). The former authors showed that the scaling property implies a kind of central limit theorem (see ►Central Limit Theorems) for exponential dispersion models, where Tweedie models appear as limiting distributions.

For $1 < p < 2$, the Tweedie models are compound Poisson distributions, which are continuous non-negative distributions with an atom at zero. Such models are useful for measurements of for example precipitation, where wet periods show positive amounts, while dry periods are recorded as zeros. Similarly, the total claim on an insurance policy over a fixed time interval (Renshaw 1993; Jørgensen and Souza 1994), may be either positive if claims were made in the period or zero if no claims were made. Other values of p (except $p = 1$) correspond to continuous distributions with support either $\mathbb{R}_+$ or $\mathbb{R}$.

Dispersion Models. Table 2 Summary of Tweedie exponential dispersion models

Distribution family	p	Support	Ω	Θ
Extreme stable exponential family	$p < 0$	$\mathbb{R}$	$\mathbb{R}_+$	$\mathbb{R}_0$
Normal distribution	$p = 0$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
Poisson distribution	$p = 1$	$\mathbb{N}_0$	$\mathbb{R}_+$	$\mathbb{R}$
Compound Poisson distribution	$1 < p < 2$	$\mathbb{R}_0$	$\mathbb{R}_+$	$\mathbb{R}_-$
Gamma distribution	$p = 2$	$\mathbb{R}_+$	$\mathbb{R}_+$	$\mathbb{R}_-$
Positive stable exponential family	$p > 2$	$\mathbb{R}_+$	$\mathbb{R}_+$	$-\mathbb{R}_0$
Inverse Gaussian distribution	$p = 3$	$\mathbb{R}_+$	$\mathbb{R}_+$	$-\mathbb{R}_0$
Extreme stable exponential family	$p = \infty$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}_-$

Notation: $-\mathbb{R}_0 = (-\infty, 0]$

Let us apply the inverse duality transformation $X = \lambda Y$ to the Tweedie variable $Y \sim \mathrm{Tw}_p(\mu, 1/\lambda)$, in order to obtain the additive form of the Tweedie distribution. By the scaling property (13) this gives $X \sim \mathrm{Tw}_p(\lambda\mu, \lambda^{1-p})$, which shows that the Tweedie exponential dispersion models have an additive as well as a reproductive form. The additive form of the Tweedie model gives rise to a class of additive processes $\{X(t) : t \geq 0\}$, defined by the following distribution of the increments:

$$X(t+s) - X(t) \sim \mathrm{Tw}_p\left(s\mu, s^{1-p}\right),$$

for $s, t > 0$. This class of *Hougaard Lévy processes* was introduced by Jørgensen (1992) and Lee and Whitmore (1993). Brownian motion (see ►Brownian Motion and Diffusions) with drift ($p = 0$) and the Poisson process ($p = 1$) mentioned above are examples of Hougaard processes, and other familiar examples include the gamma process ($p = 2$) and the inverse Gaussian process ($p = 3$). The Hougaard processes corresponding to $1 < p < 2$ are compound Poisson processes with gamma distributed jumps.

Multivariate Dispersion Models

One way to generalize (1) to the multivariate case is to consider a probability density function of the form

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \sigma^2) = a(\boldsymbol{y}; \sigma^2) \exp\left\{-\frac{1}{2\sigma^2} d(\boldsymbol{y}; \boldsymbol{\mu})\right\} \text{ for } \boldsymbol{y} \in \mathbb{R}^k, \tag{14}$$

where $(\boldsymbol{\mu}, \sigma^2) \in \Omega \times \mathbb{R}_+$, and Ω is now a domain in $\mathbb{R}^k$. Here a is a suitable function such that (14) is a probability density function, and d is again a unit deviance satisfying $d(\boldsymbol{\mu}; \boldsymbol{\mu}) = 0$ for $\boldsymbol{\mu} \in \Omega$ and $d(\boldsymbol{y}; \boldsymbol{\mu}) > 0$ for $\boldsymbol{y} \neq \boldsymbol{\mu}$. The multivariate exponential dispersion models of Jørgensen (1986; 1987a) provide examples of (14). Other examples include the multivariate von Mises-Fisher distribution, the multivariate hyperboloid distribution (see Jensen 1981), and some special cases of the multivariate simplex distributions of Barndorff-Nielsen (1991).

A more flexible definition of multivariate dispersion models is obtained by considering models of the form

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = a(\boldsymbol{y}; \boldsymbol{\Sigma}) \exp\left[-\frac{1}{2} g\left\{r^\top(\boldsymbol{y}; \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} r(\boldsymbol{y}; \boldsymbol{\mu})\right\}\right], \tag{15}$$

where $\boldsymbol{\Sigma}$ is a symmetric positive-definite $k \times k$ matrix, g is an increasing function satisfying $g(0) = 0$ and $g'(0) > 0$, and $r(\boldsymbol{y}; \boldsymbol{\mu})$ is a suitably defined k-vector of residuals satisfying $r(\boldsymbol{\mu}; \boldsymbol{\mu}) = \mathbf{0}$ for $\boldsymbol{\mu} \in \Omega$, see Jørgensen and Lauritzen (2000). An example is the ►multivariate normal distributions $\mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, which is obtained for $r(\boldsymbol{y}; \boldsymbol{\mu}) = \boldsymbol{y} - \boldsymbol{\mu}$ and g the identity function. Further examples include the multivariate Laplace and t distributions, along with the class of elliptically contoured distributions of Fang (1997). Further generalizations and examples may be found in Jørgensen and Lauritzen (2000), see also Jørgensen and Rajeswaran (2005).

About the Author

Bent Jørgensen is Professor of Mathematical Statistics at the University of Southern Denmark, Odense. He has also held positions at the Institute of Pure and Applied Mathematics, Rio de Janeiro, and the University of British Columbia, Vancouver. He is an Elected member of the International Statistical Institute. Professor Jørgensen has authored or co-authored more than 45 papers and three books, including *The Theory of Dispersion Models* (Chapman & Hall, 1997).

Cross References

►Exponential Family Models
►Generalized Linear Models
►Saddlepoint Approximations

References and Further Reading

Bar-Lev SK, Enis P (1986) Reproducibility and natural exponential families with power variance functions. Ann Stat 14:1507–1522

Barndorff-Nielsen OE, Jørgensen B (1991) Some parametric models on the simplex. J Multivar Anal 39:106–116

Fang K-T (1997) Elliptically contoured distributions. In: Kotz S, Read CB, Banks DL (eds) Encyclopedia of statistical sciences, update vol 1. Wiley, New York, pp 212–218

Hastie T (1987) A closer look at the deviance. Am Stat 41:16–20

Hougaard P (1986) Survival models for heterogeneous populations derived from stable distributions. Biometrika 73:387–396

Jensen JL (1981) On the hyperboloid distribution. Scand J Stat 8: 193–206

Jørgensen B (1983) Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. Biometrika 70:19–28

Jørgensen B (1986) Some properties of exponential dispersion models. Scand J Stat 13:187–197

Jørgensen B (1987a) Exponential dispersion models (with discussion). J R Stat Soc Ser B 49:127–162

Jørgensen B (1987b) Small-dispersion asymptotics. Braz J Probab Stat 1:59–90

Jørgensen B (1992) Exponential dispersion models and extensions: a review. Int Stat Rev 60:5–20

Jørgensen B (1997a) Proper dispersion models (with discussion). Braz J Probab Stat 11:89–140

Jørgensen B (1997b) The theory of dispersion models. Chapman & Hall, London

Jørgensen B, Lauritzen SL (2000) Multivariate dispersion models. J Multivar Anal 74:267–281

Jørgensen B, Rajeswaran J (2005) A generalization of Hotelling's T2. Commun Stat – Theory and Methods 34:2179–2195

Jørgensen B, Souza MP (1994) Fitting Tweedie'scompound Poisson model to insurance claims data. Scand Actuarial J 69–93

Jørgensen B, Martínez JR, Tsao M (1994) Asymptotic behaviour of the variance function. Scand J Statist 21:223–243

Lee M-LT, Whitmore GA (1993) Stochastic processes directed by randomized time. J Appl Probab 30:302–314
Morris CN (1981) Models for positive data with good convolution properties. Memo no. 8949. Rand Corporation, California
Morris, CN (1982) Natural exponential families with quadratic variance functions. Ann Stat 10:65–80
Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A 135:370–384
Renshaw AE (1993) An application of exponential dispersion models in premium rating. Astin Bull 23:145–147
Sweeting TJ (1981) Scale parameters: a Bayesian treatment. J R Stat Soc Ser B 43:333–338
Tweedie MCK (1947) Functions of a statistical variate with given means, with special reference to Laplacian distributions. Proc Camb Phil Soc 49:41–49
Tweedie MCK (1984) An index which distinguishes between some important exponential families. In: Ghosh JK, Roy J (eds) Statistics: applications and new directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Indian Statistical Institute, Calcutta, pp 579–604

Distance Measures

Bojana Dalbelo Basic
Professor, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia

Distance is a key concept in many statistical and pattern recognition methods (e.g., clustering, ►multidimensional scaling, ►correspondence analysis, ►principal component analysis, k-nearest neighbors, etc.). From the mathematical point of view, any set X (whose elements are called points) is said to be a *metric space* (Rudin 1976) if for any two points $\boldsymbol{a}, \boldsymbol{b} \in X$ there is an associated real number $d(\boldsymbol{a}, \boldsymbol{b})$ called *distance* if the following properties hold true:

1. $d(\boldsymbol{a}, \boldsymbol{b}) \geq 0$, (non-negativity)
2. $d(\boldsymbol{a}, \boldsymbol{b}) = 0 \iff \boldsymbol{a} = \boldsymbol{b}$ (definiteness)
3. $d(\boldsymbol{a}, \boldsymbol{b}) = d(\boldsymbol{b}, \boldsymbol{a})$ (symmetry)
4. $d(\boldsymbol{a}, \boldsymbol{c}) \leq d(\boldsymbol{a}, \boldsymbol{b}) + d(\boldsymbol{b}, \boldsymbol{c})$, for any $\boldsymbol{b} \in X$ (triangle inequality)

Any function with these properties is called a *distance function*, a *distance measure*, or a *metric*.

An example of the distance measure in metric space R^n is the so-called *Minkowski distance* (L_n) defined by:

$$d_n(\boldsymbol{a}, \boldsymbol{b}) = \sqrt[n]{\sum_{i=1}^{n} (a_i - b_i)^n},$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are vectors, elements of R^n.

The Minkowski metric for $n = 1$ is called the *Manhattan* or the *Cityblock distance* (L_1) and is given by:

$$d_1(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} |a_i - b_i|.$$

For binary vectors (whose components are restricted to values 0 or 1) this metric is called the *Hamming distance*. A special case of the Minkowski metric for the $n = 2$ metric is called the Euclidian distance (L_2) and it is the most popular metric:

$$d_2(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}.$$

For $n \to \infty$, we have the *distance* $(\boldsymbol{L_\infty})$ defined by:

$$d_\infty(\boldsymbol{a}, \boldsymbol{b}) = \max_{1 \leq i \leq n} \{|a_i - b_i|\}.$$

As an example, consider a set of points $\boldsymbol{x}$ in R^2 that have a constant distance r from the origin, that is, $(0, 0)$ point in R^2. Then, a set of points $\boldsymbol{x} \in R^2$ having the property $d_2(0, \boldsymbol{x}) = r$ is a circle with radius r and origin $(0, 0)$; a set of points $\boldsymbol{x} \in R^2$ having $d_1(0, \boldsymbol{x}) = r$ is a "diamond" having vertices at $(r, 0), (0, r), (-r, 0)$ and $(0, -r)$; and a set of points $\boldsymbol{x} \in R^2$ having the property $d_\infty(0, \boldsymbol{x}) = r$ is a square having vertices at $(r, r), (-r, r), (-r, -r), (r, -r)$.

Although the Euclidean distance is the most widely used, it is not an appropriate measure when the statistical properties of variables (attributes of the items) are being explicitly considered because it assumes that the variables are uncorrelated.

Also, if we measure the distance between two items, the Euclidean distance could change with a change in the scale of different variables. For that purpose, the *statistical distance* is used, since it is not affected by a change in scale.

An example of such *scale-invariant distance* measures is the squared (weighted) Euclidean distance for standardized data and the Mahalanobis distance. The *squared Euclidean distance for standardized data* is weighted by $1/s_i^2$, where s_i^2 is the standard deviation of the i^{th} variable:

$$d^2(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{n} \frac{(a_i - b_i)^2}{s_i^2}.$$

This distance is also called the *Karl Pearson distance*.

Another example of a weighted Euclidean distance is the *chi-squared distance* used in the ►correspondence analysis. A very important statistical measure that is scale invariant is the *Mahalanobis distance* (Mardia and Kent 1982), defined by:

$$d^2_{Mahalanobis}(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{a} - \boldsymbol{b})' \Sigma^{-1} (\boldsymbol{a} - \boldsymbol{b}),$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are two multivariate observations, Σ^{-1} is the inverse of the variance-covariance matrix and $(\boldsymbol{a}-\boldsymbol{b})'$ is the transpose of vector $(\boldsymbol{a}-\boldsymbol{b})$.

The Mahalanobis distance is designed to take into account the correlation between all variables (attributes) of the observations under consideration. For uncorrelated variables, the Mahalanobis distance reduces to the Euclidean distance for standardized data.

As an example, consider a set of points $\boldsymbol{x}$ in R^2 that have the constant distance r from the origin, that is, $(0,0)$. Then, the set of points having the property $d^2_{Mahalanobis}(0,\boldsymbol{x}) = r$ is an ellipse. The Mahalanobis distance is a positive definite quadratic form $\boldsymbol{x}'\mathbf{A}\boldsymbol{x}$, where the matrix $\mathbf{A} = \Sigma^{-1}$.

Distance measures or metrics are members of a broader concept called *similarity measures* (or *dissimilarity measures*) (Theodoridis and Koutroumbas 2009) that measure likeness (or affinity) in the case of the similarity measure, or difference (or lack of affinity) in the case of dissimilarity between objects. Similarity measures can be converted to dissimilarity measures using a monotone decreasing transformation and vice versa.

The main difference between metrics and broader concepts of similarity/dissimilarity measures is that some of the properties (1)–(4) do not hold for similarity/dissimilarity measures. For example, definiteness, or the triangle inequality, usually do not hold for similarity/dissimilarity measures.

The cosine similarity and the Pearson's product moment coefficient are two similarity measures that are not metric. The cosine similarity is the cosine of an angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ from R^n and is given by:

$$s(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|\,\|\boldsymbol{y}\|},$$

where $\|\boldsymbol{x}\|$ and $\|\boldsymbol{y}\|$ are norms of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. This measure is very popular in information retrieval and text-mining applications.

In statistical analysis (especially when applied to ecology, natural language processing, social sciences, etc.) there are often cases in which similarity or the distance between two items (e.g., sets, binary vectors) is based on two-way contingency tables with elements a, b, c, and d, where a represents the number of elements (attribute values, variables values) present in both items, b is the number of elements present in the first but absent in the second item, c is the number of elements present in the second but absent in the first item, and d is number of elements absent simultaneously in both items. The numbers a, b, c, and d can be defined as properties of two sets or two binary vectors.

Similarity coefficients (Theodoridis and Koutroumbas 2009) or *associations measures* can be defined as a combination of numbers a, b, c, and d. Examples of associations measures are:

Simple matching coefficient	$(a+d)/n$,
Dice coefficient	$2a/(2a+b+c)$,
Jaccard (or *Tanimoto*) *coefficient*	$a/(a+b+c)$.

Although association measures, similarity measures, and correlation coefficients are not metric, they are applicable in the analysis where they are consistent with the objective of the study and where they have meaningful interpretation (Sharma 1996).

Cross References

►Cook's Distance
►Data Mining Time Series Data
►Entropy and Cross Entropy as Diversity and Distance Measures
►Multidimensional Scaling: An Introduction
►Multivariate Outliers
►Statistical Natural Language Processing
►Statistics on Ranked Lists

References and Further Reading

Mardia KV, Kent JT, Bibby JM (1982) Multivariate analysis. Academic, New York
Rudin W (1976) Principles of mathematical analysis, 3rd edn. McGraw Hill, New York
Sharma SC (1996) Applied multivariate techniques. Wiley, New York
Theodoridis S, Koutroumbas K (2009) Pattern recognition, 4th edn. Elsevier

Distance Sampling

Tiago A. Marques[1,2], Stephen T. Buckland[1], David L. Borchers[1], Eric A. Rexstad[1], Len Thomas[1]
[1]University of St Andrews, St Andrews, UK
[2]Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal

Distance sampling is a widely used methodology for estimating animal density or abundance. Its name derives from the fact that the information used for inference are the recorded distances to objects of interest, usually animals, obtained by surveying lines or points. The methods

are also particularly suited to plants or immotile objects, as the assumptions involved (see below for details) are more easily met. In the case of lines the perpendicular distances to detected animals are recorded, while in the case of points the radial distances from the point to detected animals are recorded. A key underlying concept is the detection function, usually denoted $g(y)$ (here y represents either a perpendicular distance from the line or a radial distance from the point). This represents the probability of detecting an animal of interest, given that it is at a distance y from the transect. This function is closely related to the probability density function (pdf) of the detected distances, $f(y)$, as

$$f(y) = \frac{g(y)\pi(y)}{\int_0^w g(y)\pi(y)dy}, \quad (1)$$

where $\pi(y)$ is the distribution of distances available for detection and w is a truncation distance, beyond which distances are not considered in the analysis. The above pdf provides the basis of a likelihood from which the parameters of the detection function can be estimated. An important and often overlooked consideration is that $\pi(y)$ is assumed known. This is enforced by design, as the random placement of transects, independently of the animal population, leads to a distribution which is uniform in the case of line transects and triangular in the case of point transects (Buckland et al. 2001).

Given the n distances to detected animals, density can be estimated by

$$\hat{D} = \frac{n\hat{f}(0)}{2L} \quad (2)$$

in the case of line transects with total transect length L, where $f^{(0)}$ is the estimated pdf evaluated at zero distance, and by

$$\hat{D} = \frac{n\hat{h}(0)}{2k\pi} \quad (3)$$

in the case of k point transects, where $h^{(0)}$ is the slope of the estimated pdf evaluated at zero distance (Buckland et al. 2001). This is a useful result because we can then use all the statistical tools that are available to estimate a pdf in order to obtain density estimates. So one can consider plausible candidate models for the detection function and then use standard maximum likelihood to obtain estimates for the corresponding parameters and therefore density estimates.

The most common software to analyze distance sampling data, Distance (Thomas et al. 2010), uses the semi-parametric key+series adjustment formulation from Buckland (1992), in which a number of parametric models are considered as a first approximation and then some expansion series terms are added to improve the fit to the data. Standard model selection tools and goodness-of-fit tests are available for assisting in ▶model selection.

Variance estimates can be obtained using a delta method approximation to combine the individual variances of the random components in the formulas above (i.e., n and either $f^{(0)}$ or $h^{(0)}$; for details on obtaining each component variance, see Buckland et al. 2001). In some of the more complex scenarios, one must use resampling methods based on the non-parametric bootstrap, which are also available in the software.

Given a sufficiently large number of transects randomly allocated independently of the population of interest, estimators are asymptotically unbiased if (1) all animals on the transect are detected, i.e., $g(0) = 1$, (2) sampling is an instantaneous process (typically it is enough if animal movement is slow relative to the observer movement), and (3) distances are measured without error. See Buckland et al. (2001) for discussion of assumptions. Other assumptions, for example that all detections are independent events, are strictly required as the methods are based on maximum likelihood, but the methods are extraordinarily robust to their failure (Buckland 2006). Failure of the $g(0) = 1$ assumption leads to underestimation of density. Violation of the movement and measurement error assumption have similar consequences. Underestimation of distances and undetected responsive movement toward the observers lead to overestimation of density, and overestimation of distances and undetected movement away from the observer lead to underestimation of density. Random movement and random measurement error usually leads to overestimation of density. Naturally the bias depends on the extent to which the assumptions are violated. Most of the current research in the field is aimed at relaxing or avoiding the need for such assumptions. As there are no free lunches in statistics, these come at the expense of more elaborate methods, additional data demands and additional assumptions.

Further details about conventional distance sampling, including dealing with clustered populations, cue counting methods and field methods aspects, can be found in Buckland et al. (2001), while advanced methods, including the use of multiple covariates in the detection function, double platform methods for when $g(0) < 1$, spatial models, automated survey design, and many other specialized topics, are covered in Buckland et al. (2004).

About the Authors

The authors have a wide experience in developing and applying methods for estimation of animal abundance. They have published collectively around one hundred papers in distance sampling methods and applications,

and are actively involved in dissemination of distance sampling methods through international workshops and consultancy. Three of them are also co-authors and editors of the following key books on distance sampling: *Introduction to Distance Sampling: Estimating Abundance of Biological Populations* (Oxford University Press, 2001) and *Advanced Distance Sampling: Estimating Abundance of Biological Populations* (Oxford University Press, 2004).

Cross References

►Statistical Ecology
►Statistical Inference in Ecology

References and Further Reading

Buckland ST (1992) Fitting density functions with polynomials. Appl Stat 41:63–76

Buckland ST (2006) Point transect surveys for songbirds: robust methodologies. The Auk 123:345–357

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2001) Introduction to distance sampling: estimating abundance of biological populations. Oxford University Press, Oxford

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2004) Advanced distance sampling. Oxford University Press, Oxford

Thomas L, Buckland ST, Rexstad R, Laake L, Strindberg S, Hedley S, Bishop J, Marques TA, Burnham KP (2010) Distace software: design and analysis of distance sampling surveys for estimating population size. J App Ecol 47:5–14

Distributions of Order *k*

Andreas N. Philippou[1], Demetrios L. Antzoulakos[2]
[1]Professor of Probability and Statistics
University of Patras, Patras, Greece
[2]Associate Professor
University of Piraeus, Piraeus, Greece

The distributions of order k are infinite families of probability distributions indexed by a positive integer k, which reduce to the respective classical probability distributions for $k = 1$, and they have many applications. We presently discuss briefly the geometric, negative binomial, Poisson, logarithmic series and binomial distributions of order k.

Geometric Distribution of Order *k*

Denote by T_k the number of independent Bernoulli trials with success (S) and failure (F) probabilities p and $q = 1 - p$ $(0 < p < 1)$, respectively, until the occurrence of the kth consecutive success. Philippou and Muwafi (1982) observed that a typical element of the event $\{T_k = x\}$ is an arrangement

$$a_1 a_2 \dots a_{x_1+x_2+\cdots+x_k} \underbrace{SS\dots S}_{k}, \tag{1}$$

such that x_1 of the a's are $E_1 = F$, x_2 of the a's are $E_2 = SF, \dots,$ x_k of the a's are $E_k = \underbrace{SS\dots S}_{k-1}F$, and proceeded to obtain the following exact formula for the probability mass function (pmf) of T_k, namely,

$$f(x) = P(T_k = x) = p^x \sum \binom{x_1 + x_2 + \cdots + x_k}{x_1, x_2, \dots, x_k} \left(\frac{q}{p}\right)^{x_1+x_2+\cdots+x_k}, \quad x \geq k, \tag{2}$$

where the summation is taken over all non-negative integers $x_1, x_2, \dots, x_k$ satisfying the condition $x_1 + 2x_2 + \cdots + kx_k = x - k$. Alternative simpler formulas have been derived. The following recurrence for example, due to Philippou and Makri (1985), is very efficient for computations

$$f(x) = f(x-1) - qp^k f(x-1-k), \quad x > 2k \tag{3}$$

with initial conditions $f(k) = p^k$ and $f(x) = qp^k$ for $k < x \leq 2k$. Furthermore, it shows that $f(x)$ attains its maximum p^k for $x = k$, followed by a plateau of height qp^k for $x = k+1, k+2, \dots, 2k$, and decreases monotonically to 0 for $x \geq 2k+1$.

Philippou et al. (1983) employed the transformation $x_i = m_i$ $(1 \leq m_i \leq k)$ and $x = m + \sum_{i=1}^{k}(i-1)m_i$ to show that $\sum_{x=k}^{\infty} f(x) = 1$ (and hence $f(x)$ is a proper pmf). They named the distribution of T_k *geometric distribution of order k with parameter p* and denoted it by $G_k(p)$, since for $k = 1$ it reduces to the classical geometric distribution with pmf $f(x) = q^{x-1}p$ $(x \geq 1)$. It follows from (2), by means of the above transformation and the multinomial theorem, that the probability generating function (pgf) of T_k is given by

$$\phi_k(w) = \sum_{x=k}^{\infty} x^w f(x) = \frac{p^k w^k (1 - pw)}{1 - w + qp^k w^{k+1}}, \quad |w| \leq 1. \tag{4}$$

The mean and variance of T_k readily follow from its pgf and they are given by

$$E(T_k) = \frac{1 - p^k}{qp^k}, \quad Var(T_k) = \frac{1 - (2k+1)qp^k - p^{2k+1}}{(qp^k)^2}. \tag{5}$$

A different derivation of (4) was first given by Feller (1968), who used the method of partial fractions on $\phi_k(w)$ to

derive the surprisingly good approximation

$$P(T_k > x) \simeq \frac{1 - pw_0}{(k + 1 - kw_0)q\, w_0^{x+1}}, \tag{6}$$

where w_0 is the unique positive root of $p^k w^k / \phi_k(w)$.

It is well known that the negative binomial distribution arises as the distribution of the sum of r independent rv's distributed identically as geometric. This fact motivated the genesis of the following distribution of order k.

Negative Binomial Distribution of Order k

Let $X_1, X_2, \ldots, X_r$ be independent rv's distributed identically as $G_k(p)$ and set $T_{r,k} = \sum_{i=1}^{r} X_i$. Then, relation (4) implies

$$\begin{aligned}\phi_{r,k}(w) &= \sum_{x=rk}^{\infty} x^w P(T_{r,k} = x) \\ &= \left(\frac{p^k w^k (1 - pw)}{1 - w + qp^k w^{k+1}}\right)^r, \quad |w| \le 1. \end{aligned} \tag{7}$$

Expanding $\phi_{r,k}(w)$ in a series around 0, Philippou et al. (1983) obtained the pmf of $T_{r,k}$ as

$$\begin{aligned}P(T_{r,k} = x) = p^x \sum \binom{x_1 + x_2 + \cdots + x_k + r - 1}{x_1, x_2, \ldots, x_k, r - 1} \\ \times \left(\frac{q}{p}\right)^{x_1 + x_2 + \cdots + x_k}, \quad x \ge rk, \end{aligned} \tag{8}$$

where the summation is taken over all non-negative integers $x_1, x_2, \ldots, x_k$ satisfying the condition $x_1 + 2x_2 + \cdots + kx_k = x - rk$. They called the distribution of $T_{r,k}$ *negative binomial distribution of order k with parameter vector* (r, p) and denoted it by $NB_k(r, p)$, since for $k = 1$ it reduces to the negative binomial distribution with pmf $f(x) = \binom{x-1}{r-1} p^r q^{x-r}$ $(x \ge r)$. For $r = 1$, it reduces to the geometric distribution of order k. Obviously, $T_{r,k}$ denotes the waiting time for the occurrence of the rth non-overlapping success run of length k. Alternative formulas, simpler than (8), have been derived by Godbole (1990) (see also Antzoulakos and Philippou (1999)) and Philippou and Georghiou (1989). The latter authors established also the following efficient recurrence

$$\begin{aligned}g(x) = P(T_{r,k} = x) = \frac{q/p}{x - rk} \sum_{i=1}^{k} [x - rk + i(r-1)] \\ \times p^i g(x - i), \quad x \ge rk + 1, \end{aligned} \tag{9}$$

with $g(x) = 0$ for $0 \le x \le rk - 1$ and $f(rk) = p^{rk}$, which recaptures (3) for $r = 1$.

Let $X_1, X_2, \ldots, X_n$ be independent rv's distributed identically as $NB_k(r, p)$, and set $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Philippou (1984) observed that the moment estimator $\hat{p}$ of p is the unique admissible root of the equation $\frac{r(1-p^k)}{(1-p)p^k} = \overline{X}$. In particular, for $k = 2$, $\hat{p} = \frac{r + (r^2 + 4r\overline{X})^{1/2}}{2\overline{X}}$ and it is consistent for p.

The fact that the Poisson and the logarithmic series distributions arise as appropriate limits of the negative binomial distribution, motivated the following.

Poisson and Logarithmic Series Distributions of Order k

Let X be a rv distributed as $NB_k(r, p)$ and set $Y_r = X - rk$.

(a) Assume that $q \to 0$ and $rq \to \lambda$ (> 0) as $r \to \infty$. Then, for $y = 0, 1, \ldots,$

$$\lim_{r \to \infty} P(Y_r = y) = e^{-k\lambda} \sum \frac{\lambda^{y_1 + y_2 + \cdots + y_k}}{y_1! y_2! \ldots y_k!} = f_Y(y), \tag{10}$$

where the summation is taken over all non-negative integers $y_1, y_2, \ldots, y_k$ satisfying the condition $y_1 + 2y_2 + \cdots + ky_k = y$. The distribution of Y has been named *Poisson distribution of order k with parameter* λ.

(b) Assume that r is positive and real. Then, for $z = 1, 2, \ldots,$

$$\begin{aligned}\lim_{r \to 0} P(Y_r = z \mid Y_r \ge 1) = \alpha p^z \sum \frac{(z_1 + z_2 + \cdots + z_k - 1)!}{z_1! z_2! \ldots z_k!} \\ \times \left(\frac{q}{p}\right)^{z_1 + z_2 + \cdots + z_k} = f_Z(z), \end{aligned} \tag{11}$$

where $\alpha = -1/k \log p$, and the summation is taken over all non-negative integers $z_1, z_2, \ldots, z_k$ satisfying the condition $z_1 + 2z_2 + \cdots + kz_k = z$. This result is due to Aki et al. (1984) who extended the definition of $NB_k(r, p)$ to positive real r. They called the distribution of Z *logarithmic series distribution of order k with parameter p*.

We end this note with a few words on the binomial distribution of order k.

Binomial Distribution of Order k

Let $N_{n,k}$ denote the number of non-overlapping success runs of length k in n (≥ 1) independent trials with success probability p $(0 < p < 1)$. The ▶asymptotic normality of a normalized version of $N_{n,k}$ was established by von Mises (see Feller 1968:324), where a simpler proof is presented). Its exact pmf was obtained by Hirano (1986) and Philippou and Makri (1986), who named the distribution *binomial distribution of order k with parameter vector* (n, p). They

found that, for $x = 0, 1, \ldots, [n/k]$,

$$P(N_{n,k} = x) = \sum_{i=0}^{k-1} \sum \binom{x_1 + x_2 + \cdots + x_k + x}{x_1, x_2, \ldots, x_k, x} p^n \times \left(\frac{q}{p}\right)^{x_1 + x_2 + \cdots + x_k}, \quad (12)$$

where $[u]$ denotes the greatest integer in u and the inner summation is taken over all non-negative integers $x_1, x_2, \ldots, x_k$ satisfying the condition $x_1 + 2x_2 + \cdots + kx_k + i = n - kx$. The following exact formula for the mth descending factorial moment μ'_m of $N_{n,k}$ was established by Antzoulakos and Chadjiconstantinidis (2001),

$$\mu'_m = m! p^{km} \sum_{j=0}^{\min(n-km,m)} (-1)^j \binom{m}{j} p^j \times \sum_{i=0}^{[\frac{n-km-j}{k}]} \binom{m+i-1}{i} \binom{m+n-k(m+i)}{m} p^{ki}. \quad (13)$$

The above results for the pmf's of T_k and $N_{n,k}$ give exact formulas for the reliability of a *consecutive-k-out-of-n:F system*, which is very important in Engineering. For more on distributions of order k and their applications, we refer to Balakrishnan and Koutras (2002).

About the Authors

Professor Andreas Philippou is a member of the Greek Statistical Authority and he has been teaching statistics and probability at the University of Patras since 1980. He has also taught earlier at the University of Texas at El Paso and the American University of Beirut. He wrote 75 papers, mainly in statistics and probability, and edited 7 books on Fibonacci numbers and their applications. He served as Vice-President of Hellenic Aerospace Industry (1981–1982), Vice-Rector of the University of Patras (1983–1986), and Minister of Education (1988–1990), Member of the House of Representatives (1991–2001) and Member of the Tax Tribunal (2004–2007) of the Republic of Cyprus. Professor Philippou is Grande Ufficiale of Italy and Honorary President of the Mathematical Association of Cyprus. He is a reviewer of more than 400 papers and books (including papers written by Paul Erdős, Peter Hall, C. R. Rao, and N. Balakrishnan).

Dr. Demetrios Antzoulakos is Associate Professor at the University of Piraeus where he has been teaching since 1997. He has also taught earlier at the University of Crete. He wrote 30 research papers mainly on the topics of statistical process control and distribution theory of runs and patterns.

Cross References

►Binomial Distribution
►Geometric and Negative Binomial Distributions

References and Further Reading

Aki S, Kuboki H, Hirano K (1984) On discrete distributions of order k. Ann Inst Stat Math 36:431–440
Antzoulakos DL, Chadjiconstantinidis S (2001) Distributions of numbers of success runs of fixed length in Markov dependent trials. Ann Inst Stat Math 53:599–619
Antzoulakos DL, Philippou AN (1999) Multivariate Pascal polynomials of order k with probability applications. In: Howard FT (ed) Applications of Fibonacci numbers, vol 8. Kluwer, Dordrecht, pp 27–41
Balakrishnan N, Koutras MV (2002) Runs and scans with applications. Wiley, New York
Feller W (1968) An introduction to probability theory and its applications, vol I, 3rd edn. Wiley, New York
Godbole AP (1990) Specific formulae for some success run distributions. Stat Probab Lett 10:119–124
Hirano K (1986) Some properties of the distributions of order k. In: Philippou AN et al (eds) Fibonacci numbers and their applications. Reidel, Dordrecht, pp 43–53
Philippou AN (1984) The negative binomial distribution of order k and some of its properties. Biom J 26:784–789
Philippou AN, Georghiou C (1989) Convolutions of Fibonacci-type polynomials of order k and the negative binomial distributions of order k. Fibonacci Q 27:209–216
Philippou AN, Makri FS (1985) Longest success runs and Fibonacci-type polynomials. Fibonacci Q 23:338–346
Philippou AN, Makri FS (1986) Successes, runs and longest runs. Stat Probab Lett 4:211–215
Philippou AN, Muwafi AA (1982) Waiting for the kth consecutive success and the Fibonacci sequence of order k. Fibonacci Q 20:28–32
Philippou AN, Georghiou C, Philippou GN (1983) A generalized geometric distribution and some of its properties. Stat Probab Lett 1:171–175

Diversity

Tarald O. Kvålseth
Professor Emeritus
University of Minnesota, Minneapolis, MN, USA

Introduction

Diversity is a concept that appears in various fields of study. In the most general situation involving a set of s exhaustive and mutually exclusive events with probabilities (or proportions) p_i $(i = 1, \ldots, s)$, diversity is an attribute that depends on s and all the individual p_i's. In particular, diversity is typically considered to increase with increasing s and with increasing evenness (uniformity) among the p_i's.

For any given s, diversity is considered to be minimum for the degenerate distribution P_s^0 and maximum for the completely even (uniform) distribution P_s^1 defined by

$$P_s^0 = (0,\ldots,0,1,0,\ldots,0),\ P_s^1 = (1/s,\ldots,1/s). \quad (1)$$

As a clarification of these concepts and their measurement, consider a generic measure of diversity D that takes on the value $D(P_s)$ for the probability distribution $P_s = (p_1,\ldots,p_s)$ with

$$D\left(P_s^0\right) \le D\left(P_s\right) \le D\left(P_s^1\right) \quad (2)$$

for the P_s^0 and P_s^1 in (1). Furthermore, define the normed form of $D(P_s)$ as

$$D^*(P_s) = \frac{D(P_s) - D\left(P_s^0\right)}{D\left(P_s^1\right) - D\left(P_s^0\right)} \in [0,1] \quad (3)$$

which effectively controls for s. From (3),

$$D(P_s) = D\left(P_s^0\right) + D^*(P_s)\left[D\left(P_s^1\right) - D\left(P_s^0\right)\right] \quad (4)$$

showing that $D(P_s)$ is an increasing function of both the evenness (uniformity) $D^*(P_s)$ and of $D(1/s,\ldots,1/s)$, which, for an appropriate measure D, is strictly increasing in s.

Numerous measures of diversity D and evenness D^* have been proposed. This is especially the case in biology (and ecology) where such measures are frequently used. In such applications, the concern is with a sample or a population of s different species, with p_i being the probability of the event that a randomly selected specimen belongs to the ith species for $i = 1,\ldots,s$. The $P_s = (p_1,\ldots,p_s)$ is then usually referred to as the *species abundance distribution* and s is often called the *species richness.*

Commonly Used Measures

The most widely used diversity measures appear to be the following:

$$D_1(P_s) = s, \quad (5)$$

$$D_2(P_s) = 1 - \sum_{i=1}^{s} p_i^2, \quad (6)$$

$$D_3(P_s) = -\sum_{i=1}^{s} p_i \log p_i. \quad (7)$$

where both base −2 and base −e (natural) logarithms are being used. The D_2 is most often referred to as the *Simpson index*, after Simpson (1949), although it is sometimes called the *Gini index* (e.g., Upton and Cook 2002). However, in some historical comments about D_2 and $1 - D_2$, Good (1982) suggested that "it is unjust to associate ρ with any one person" (where $\rho = 1 - D_2$). The measure of D_3 is the Shannon's entropy (Shannon 1948).

While D_1 has very limited information content, both D_2 and D_3 can be seen to have a number of desirable properties. In particular, for both $i = 2$ and $i = 3$, D_i is (a) zero-indifferent (expansible), i.e., $D_i(p_1,\ldots,p_s,0) = D_i(p_i,\ldots,p_s)$, (b) permutation symmetric in the p_i's, (c) continuous in the p_i's and (d) strictly Schur-concave in P_s (Marshall and Olkin 1979, ch. 3) and also strictly concave (in the usual sense). Also, with $D_i\left(P_s^0\right) = 0$ for both $i = 2$ and $i = 3$ in (2)–(4),

$$D_i(P_s) = D_i^*(P_s) D_i\left(P_s^1\right),\ i = 2,3 \quad (8)$$

with the evenness measure $D_i^*(P_s)$ for $i = 2,3$ (Magurran 2004, ch. 4) and showing that $D_i(P_s)$ increases with both evenness and s since $D_2\left(P_s^1\right) = 1 - 1/s$ and $D_3\left(P_s^1\right) = \log s$.

The D_2 in (6) has the advantage of having a reasonable probabilistic meaning, which helps with interpretation of its numerical values. In terms of a statistical experiment with s possible events whose probability distribution is $P_s = (p_1,\ldots,p_s)$, $D_2(P_s)$ is simply the probability that two independent repetitions of the experiment result in two different events occurring. In the case of a biological sample or population, $D_2(P_s)$ is the probability that two specimen, selected at random with replacement, belong to different species. By contrast, the D_3 in (7) has no such convenient probability interpretation. As a measure of diversity, this ▶entropy has no real or important, theoretical or practical, advantage over D_2. Nevertheless, and for no good reason, the D_3, often referred to as the Shannon index, remains a most popular diversity measure, causing some to lament that "Simpson's index remains inexplicably less popular than the Shannon index" (Magurran 2004: 115).

Cardinal Measures

Various other diversity measures and their corresponding evenness measures have been suggested (e.g., Magurran 2004). Some of these are cardinal measures or so-called *numbers equivalent* measures. Such a measure ranges in value from 1 for the distribution P_s^0 to s for the distribution P_s^1 in (1). For some diversity measure D, the numbers equivalent D_e is defined as being the number of elements in an equiprobable (completely even or uniform) distribution for which the value of D equals $D(P_s)$ for any given $P_s = (p_1,\ldots,p_s)$, i.e.,

$$D\left(1/D_e(P_s),\ldots,1/D_e(P_s)\right) = D(P_s) \quad (9)$$

where $D_e(P_s)$ is not necessarily an integer number, although this expression can only be strictly true in the integer case.

From (6), (7), and (9), the numbers equivalent of D_2 and D_3 (with base −e logarithms) are readily seen to be

$$D_4(P_s) = D_{2e}(P_s) = \frac{1}{1 - D_2(P_s)} = \frac{1}{\sum_{i=1}^{s} p_i^2},$$

$$D_5(P_s) = D_{3e}(P_s) = e^{D_3(P_s)}. \tag{10}$$

It is clear from (10) that, for the distributions in (1), $D_4\left(P_s^0\right) = D_5\left(P_s^0\right) = 1$ and $D_4\left(P_s^1\right) = D_5\left(P_s^1\right) = s$. Also, by subtracting 1 from $D_4(P_s)$, one gets the following measure (Kvålseth 1991):

$$D_6(P_s) = \frac{1}{\sum_{i=1}^{s} p_i^2} - 1 \tag{11}$$

which has a statistical odds interpretation equivalent to the probability interpretation of $D_2(P_s)$: it is the odds that different events will occur during two independent repetitions of a statistical experiment. Or, in the case when two specimen are randomly selected with replacement from a biological community (sample or population), $D_6(P_s)$ is the odds that the specimen will be of different species.

An interesting and rather intuitive diversity measure can be formulated by simply adjusting s for the absolute differences between the p_i's as follows:

$$D_7(P_s) = s - \sum_{1 \le i < j \le s} \sum |p_i - p_j| \tag{12}$$

where, for the distribution in (1), $D_7\left(P_s^0\right) = 1$ and $D_7\left(P_s^1\right) = s$. Since the summation term in (12) can be shown to be strictly Schur-convex (Marshall and Olkin 1979: 13, 411), $D_7(P_s)$ is strictly Schur-concave in P_s.

If the p_i's are ordered such that $p_{[1]} \ge p_{[2]} \ge \ldots \ge p_{[s]}$ and since

$$\sum_{1 \le i < j \le s} \sum |p_i - p_j| = \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} (p_{[i]} - p_{[j]}) = \sum_{i=1}^{s} (s + 1 - 2i) p_{[i]}$$

the $D_7(P_s)$ in (12) can also be expressed as

$$D_7(P_s) = 2 \sum_{i=1}^{s} i p_{[i]} - 1. \tag{13}$$

The D_7 in (12)–(13) has been discovered and re-discovered as a potential diversity measure (Carmargo 1993; Kvålseth 1998; Patil and Taillie 1982). While Carmargo (1993) proposed $D_7(P_s)/s \in [1/s, 1]$ as an evenness index for which the lower bound depends on s, a preferred measure of evenness would seem from (3) to be

$$D_7^*(P_s) = \frac{2 \sum_{i=1}^{s} i p_{[i]}}{s - 1} \in [0, 1]. \tag{14}$$

Families of Measures

Parameterized families of diversity measures have also been proposed and of which some of the above measures are special members. Hill (1973) proposed the following inverse self-weighted arithmetic mean of order β:

$$D_{(\beta)}(P_s) = \left(\sum_{i=1}^{s} p_i^{\beta+1} \right)^{-1/\beta}, \quad -\infty < \beta < \infty \tag{15}$$

($\beta = \alpha - 1$ in Hill's notation). Patil and Taillie (1982) proposed

$$D^{(\beta)}(P_s) = \left(1 - \sum_{i=1}^{s} p_i^{\beta+1} \right) \Big/ \beta, \; -1 \le \beta < \infty. \tag{16}$$

From (5) to (7) and (16), it is seen that $D_1 = D^{(-1)} + 1$, $D_2 = D^{(1)}$, and $D_3 = D^{(0)}$ (in the limiting sense of $\beta \to 0$ and with natural logarithm in (7)). Also from (10) and (15), $D_4 = D_{(1)}$ and $D_5 = D_{(0)}$ (in the limit as $\beta \to 0$). It can also be seen from (9), (15) and (16) that $D_{(\beta)}$ is simply the numbers equivalent of $D^{(\beta)}$.

Statistical Inferences

Consider now that the p_i's are multinomial sample estimates (and estimators) $p_i = n_i/N$ for $i = 1, \ldots, s$ and with sample size $N = \sum_{i=1}^{s} n_i$. One may then be interested in making statistical inferences about the population value $D(\Pi_s)$ of some diversity measure D for the corresponding population distribution $\Pi_s = (\pi_1, \ldots, \pi_s)$. In particular, besides the estimate $D(P_s)$, one may want to construct a confidence interval for $D(\Pi_s)$. When N is reasonably large and $D(\Pi_s)$ is differentiable, it follows from the *delta method* Agresti (2002) that $D(P_s)$ has approximately a normal distribution with mean $D(\Pi_s)$ and estimated variance

$$\hat{\sigma}_D^2 = \frac{1}{N} \left[\sum_{i=1}^{s} p_i \hat{\phi}_{Di}^2 - \left(\sum_{i=1}^{s} p_i \hat{\phi}_{Di} \right)^2 \right] \tag{17}$$

where

$$\hat{\phi}_{Di} = \left. \frac{\partial D(\Pi_s)}{\partial \pi_i} \right|_{\pi_i = p_i}, \; i = 1, \ldots, s \tag{18}$$

i.e., $\hat{\phi}_{Di}$ is the partial derivative of $D(\Pi_s)$ with respect to π_i, which is then replaced with p_i, for $i = 1, \ldots, s$.

From (17)–(18) and for the D_i in (6), (7), (10), (11), and (13), the following estimated variances are obtained:

$$\hat{\sigma}^2_{D2} = \frac{4}{N}\left[\sum_{i=1}^{s} p_i^3 - \left(\sum_{i=1}^{s} p_i^2\right)^2\right], \tag{19}$$

$$\hat{\sigma}^2_{D3} = \frac{1}{N}\left[\sum_{i=1}^{s} p_i(\log p_i)^2 - \left(\sum_{i=1}^{s} p_i \log p_i\right)^2\right], \tag{20}$$

$$\hat{\sigma}^2_{D4} = \frac{4}{N\left(\sum_{i=1}^{s} p_i^2\right)^4}\left[\sum_{i=1}^{s} p_i^3 - \left(\sum_{i=1}^{s} p_i^2\right)^2\right], \tag{21}$$

$$\hat{\sigma}^2_{D5} = \left[D_5(P_s)\right]^2 \hat{\sigma}^2_{D3},\ \hat{\sigma}^2_{D6} = \hat{\sigma}^2_{D4}, \tag{22}$$

$$\hat{\sigma}^2_{D7} = \frac{4}{N}\left[\sum_{i=1}^{s} i^2 p_{[i]} - \left(\sum_{i=1}^{s} i p_{[i]}\right)^2\right] \tag{23}$$

where D_3 in (7) and (10) and hence (20) and D_5 in (22) are assumed to be based on natural logarithms.

For example, consider the multinomial frequencies $n_i = 17, 33, 20, 30$ for which $D_2(P_4) = 0.7322$, and from (19), $\hat{\sigma}^2_{D2} = 0.00017$. Consequently, an approximate 95% confidence interval for $D_2(\Pi_4)$ becomes $0.7322 \pm 1.96\sqrt{0.00017}$, or $(0.71, 0.76)$. Similarly, from (13) and (23), $D_7(P_4) = 3.4200$ and $\hat{\sigma}^2_{D7} = 0.0466$ so that an approximate 95% confidence interval for $D_7(\Pi_4)$ is $3.4200 \pm 1.96\sqrt{0.0466}$ or $(3.00, 3.84)$.

About the Author

For biography *see* the entry ►Entropy.

Cross References

►Entropy
►Entropy and Cross Entropy as Diversity and Distance Measures
►Information Theory and Statistics
►Lorenz Curve
►Measurement of Uncertainty
►Statistical View of Information Theory

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken
Carmargo JA (1993) Most dominance increase with the number of subordinate species in competitive interactions? J Theor Biol 161:537–542
Good IJ (1982) Comment to paper by Patil and Taillie. J Am Stat Assoc 77:561–563
Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54:427–432
Kvålseth TO (1991) Note on biological diversity, evenness, and homogeneity measures. Oikos 62:123–127
Kvålseth TO (1998) On difference – based summary measures. Percept Mot Skills 87:1379–1384
Magurran AE (2004) Measuring biological diversity. Blackwell, Oxford
Marshall AW, Olkin I (1979) Inequalities: theory of majorization and its applications. Academic, San Diego
Patil GP, Taillie C (1982) Diversity as a concept and its measurement. J Am Stat Assoc 77:548–561
Shannon CE (1948) The mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656
Simpson EH (1949) Measure of diversity. Nature 163:688
Upton G, Cook I (2002) Oxford dictionary of statistics. Oxford University Press, Oxford

Divisible Statistics

Estate V. Khmaladze
Professor
Victoria University of Wellington, Wellington, New Zealand

Suppose $\nu_{in}, i = 1, \ldots, M$, are frequencies of M disjoint events in a sample of size n. Suppose the probabilities of these events are $p_i, i = 1, \ldots, M$. The statistics of the form

$$\sum_{i=1}^{M} g(\nu_{in}, np_i) \quad \text{or} \quad \sum_{i=1}^{M} h\left(\frac{\nu_{in} - np_i}{\sqrt{np_i}}, np_i\right)$$

are called divisible or additively divisible statistics. The total space was "divided" into disjoint events, then frequencies of these events were treated, so to say, "individually" and only then an aggregated sum was formed – hence the term. It will be more justified when we start speaking about finer and finer divisions of the space.

Examples of divisible statistics are easy to find. To some extent, to attribute to them the name of divisible statistics is similar to the discovery that we speak prose, as statisticians could very well have used and studied them without paying attention or knowing that these were divisible statistics. In the following examples,

$$X^2_{nM} = \sum_{i=1}^{M} \frac{(\nu_{in} - np_i)^2}{np_i}, \quad L_{nM} = \sum_{i=1}^{M} \nu_{in} \ln \frac{\nu_{in}}{np_i},$$

$$\mu_n = \sum_{i=1}^{M} \mathbb{I}\{\nu_{in} > 0\},$$

where $\mathbb{I}\{A\}$ denotes the indicator function of the event A, the reader will easily recognize the classical chi-square goodness of fit statistic, the maximum likelihood statistic for the ►multinomial distribution and what is often called an "empirical vocabulary" – the number of different events

in a sample. Versions of this latter statistic are often used in problems of statistical diversity, cryptology and analysis of texts:

$$\mu_n(0) = \sum_{i=1}^{M} \mathbb{I}\{\nu_{in} = 0\},$$
$$\mu_n(k) = \sum_{i=1}^{M} \mathbb{I}\{\nu_{in} = k\},\ k = 1, 2, \ldots.$$

The classical theory studies the asymptotic behavior of divisible statistics when the sample size $n \to \infty$, but the number of different events M stays fixed.

In many problems, however, a different setting is necessary: not only $n \to \infty$, but also $M \to \infty$ at the same time, so that $p_i, i = 1, \ldots, M$, form a triangular array of an increasing number of diminishing probabilities. Indeed, in statistical analysis of random number generators often no less than $M = 10^7$ different events are identified and samples of $n = 10^7$–10^8 are generated. In analysis of a corpus of a language, the number of different words is often of the order $M = 10^5$–10^6, while the size of texts analyzed is about $n = 10^6$–10^7 word-usages (see, e.g., Baayen 2001; Simpson et al. 2002).

As a rule, however, one cannot access this asymptotic range by using the classical results first and then letting M tend to ∞. For example, although it is true that

$$X^2_{nM} \to_d \chi^2_{M-1},\ n \to \infty, \quad \text{and}$$
$$\frac{\chi^2_{M-1} - (M-1)}{\sqrt{2(M-1)}} \to_d N(0,1),\ M \to \infty,$$

where χ^2_{M-1} is a chi-square random variable with $M-1$ degrees of freedom and $N(0,1)$ is a standard normal random variable, for most cases it is not true that

$$\frac{X^2_{nM} - (M-1)}{\sqrt{2(M-1)}} \to_d N(0,1)$$

if $n, M \to \infty$ simultaneously and $n/M \to const$.

One of the frequently used tools to study the asymptotic behavior of divisible statistics is the so-called Poissonization (see, e.g., Ivchenko and Medvedev 1980; Morris 1975): if sample size N is Poisson(n) random variable then frequencies $\nu_{iN}, i = 1, \ldots, M$, become independent Poisson(np_i) random variables and for $B_n \subset \mathbb{R}$

$$P\left\{\sum_{i=1}^{M} g(\nu_{in}, np_i) \in B_n\right\} = \frac{P\left\{\sum_{i=1}^{M} g(\nu_{iN}, np_i) \in B_n, N = n\right\}}{P\{N = n\}}.$$

In the numerator of the right side there is now a sum of independent random variables, which is convenient. Some trouble is, however, that the probability of $N = n$ tends to 0 and some sort of a local limit theorem must be used, while the probability on the left side is obviously of a global nature. In certain problems limit theorems for the process of partial sums

$$Z_n(m) = \sum_{i=1}^{m} g(\nu_{in}, np_i),\ m = 1, 2, \ldots, M,$$

are needed and the use of Poissonization above, again, becomes technically unpleasant.

Alternative approach, demonstrated in Khmaladze (1983), suggests the use of an "additional" object – the filtration $\{\mathcal{H}_{mn}\}_{m=1}^{M}$, where each σ-algebra $\mathcal{H}_{mn}$ is generated by the first m frequencies $\nu_{1n}, \ldots, \nu_{mn}$. The gain here lies in the following: in functional limit theorems for the resulting semi-martingale $\{Z_n(m), \mathcal{H}_{mn}\}_{m=1}^{M}$, conditional distribution of $\nu_{m+1,n}$ given $\mathcal{H}_{mn}$ plays a central role (see, e.g., Liptzer and Shiryaev 1980; Rebolledo 1975); however, this distribution is extremely simple – just binomial distribution $b(\cdot, \tilde{n}_{m+1}, \tilde{p}_{m+1})$ with parameters

$$\tilde{n}_{m+1} = n - \sum_{i=1}^{m} \nu_{in} \quad \text{and} \quad \tilde{p}_{m+1} = p_{m+1} \Big/ \left(1 - \sum_{i=1}^{m} p_i\right).$$

This allows to prove the functional limit theorem, that is, convergence to a Brownian motion (see ▶Brownian Motion and Diffusions), for the martingale part of $\{Z_n(m), \mathcal{H}_{mn}\}_{m=1}^{M}$ in a simple and natural way.

True that if $\max p_i \to 0$, while $n, M \to \infty$, the mutual dependence of multinomial frequencies $\nu_{in}, i = 1, \ldots, M$, decreases, but there is an increasing number of them. Therefore the process $\{Z_n(m), \mathcal{H}_{mn}\}_{m=1}^{M}$ should not be expected to converge to a process with independent increments. Centered and normalized in a usual way and in time $t = m/M$ this process converges to a Gaussian semi-martingale,

$$W(t) + K(t), \quad t \in [0,1],$$

where W is a Brownian motion and the compensator K, which accounts for the above mentioned dependence, is a process with differentiable trajectories.

This result, and its extensions, was established in Khmaladze (1983) under an assumption that all frequencies ν_{in} are asymptotically Poisson random variables. A very interesting case arises when some of the frequencies can be asymptotically Gaussian, that is, some probabilities in our triangular array do not decrease sufficiently quickly. Indeed, in a typical corpus of a contemporary language, with word-count of order 10^6, some words will occur tens of thousands of times, while many others only a few hundred times, while still many more will occur only a few times.

In this mixed case even an asymptotic expression for the expected values

$$E\sum_{i=1}^{M} h\left(\frac{v_{in}-np_i}{\sqrt{np_i}}, np_i\right)$$

becomes an object of research. In particular, asymptotic expressions for $E\mu_n(k)$ may reveal the presence or absence of well known limiting expressions like Zipf's law or Karlin–Rouault law (see, e.g., Baayen 2001; Karlin 1968; Rouault 1978). Also the central limit theorem (see ►Central Limit Theorems) for the mixed case is yet to be established. At the moment of writing this article some work on this was in progress (e.g., Khmaladze 2009; Kvizhinadze and Wu 2010).

About the Author

For biography *see* the entry ►Testing Exponentiality of Distribution.

Cross References

►Brownian Motion and Diffusions
►Chi-Square Tests
►Multinomial Distribution
►Poisson Distribution and Its Application in Statistics
►Random Walk
►Uniform Random Number Generators

References and Further Reading

Baayen RH (2001) Word frequency distributions. Kluwer, Dordrecht
Ivchenko GI, Medvedev YuI (1980) Decomposable statistics and hypothesis testing for grouped data. Theory Probab Appl 25:540–551
Karlin S (1968) Central limit theorems for certain infinite urn schemes. Indiana Univ Math J 17:373–401
Khmaladze EV (1983) Martingale limit theorems for divisible statistics. Theory Probab Appl 28:530–549
Khmaladze EV (2009) Diversity of responses in questionnaires and similar objects. MSOR Research Report 09–3, VUW
Kvizhinadze G, Wu H (2010) Diversity of responses in general questionnaires, Stat Prob Lett 80:1103–1110
Liptzer RSh, Shiryaev AN (1980) Functional central limit theorem for semimartingales. Theory Probab Appl 25:680–703
Morris C (1975) Central limit theorem for multinomial sums. Ann Stat 3:165–188
Rebolledo R (1980) Central limiit theorem for local martingales. Z Warsch Verw Geb 51:269–286
Rouault A (1978) Loi de Zipf et sources markoviennes. Ann Inst H Poincaré 14:169–188
Simpson RC, Briggs SL, Ovens J, Swales JM (2002) The Michigan Corpus of academic spoken English. The Regents of the University of Michigan, Ann Arbor

Dummy Variables

Nataša Erjavec
Professor, Faculty of Economics
University of Zagreb, Zagreb, Croatia

D

A dummy variable is a binary variable that can take only two values, 0 and 1. It is often used in the regression model to incorporate qualitative (categorical) explanatory variables, such as gender (male or female), marital status (single or married), union membership (yes or no), etc. In the above examples, the dummy variable can be defined to take the value one if the observation relates to a male and zero for a female (or vice versa), one for single and zero if married, one for a union member and zero if a nonmember. The dummy variable can then be used in a standard regression model like any other regressor (Green 2002). As an example, if we are interested whether earnings are subject to gender discrimination, the simple regression model can be defined as

$$Y_i = \alpha + \beta X_i + \gamma D_i + \varepsilon_i \quad (1)$$

in which the regressand (Y) are earnings and the regressor (X) is the working hours of an individual. In the model, one dummy variable is introduced that takes the value one ($D_i = 1$) if the ith individual is male and zero ($D_i = 0$) if the individual is female. The parameter γ in Eq. (1) represents the difference between the expected earnings of men and women who work the same hours.

In general, the qualitative (categorical) variable can have k classifications (categories). The examples are ethnicity (i.e., Hispanic, black, or white), region (north, south, east, and west) in a study of salary levels in a sample of highly educated employees or the effects of company size (small, medium, or large) on wages. In such cases, in the standard regression model $(k-1)$ dummy variables are required, one less than the number of classifications. In this way, the dummy trap (perfect ►multicollinearity when the model cannot be estimated) is avoided. For example, when analyzing the effects of company size on wages in which there are three types of companies (small, medium, and large), $k = 3$, two dummy variables are required in the regression model. Thus, the regression model (1) can be extended to

$$Y_i = \alpha + \beta X_i + \gamma D_{Mi} + \delta D_{Li} + \varepsilon_i \quad (2)$$

where the dummy variable D_{Mi} takes the value of one if the ith individual works in a medium size company and zero otherwise, while D_{Li} takes the value of one if the ith individual works in the large company and zero otherwise. The small company is taken as a reference category to

which Eq. (2) applies. The dummy variable for the reference category is not included. In Eq. (2), the parameter γ now represents the expected extra earnings of an individual who works in medium size company relative to the earnings of an individual in the reference category, i.e., in a small company.

Dummy variables have many applications. For example, with dummy variables one can treat seasonality in the data or capture the effect of ►outliers (e.g., abrupt shifts in data). Furthermore, in the regression model, a dummy variable can be used as a regressor variable for coefficient stability tests, for obtaining predictions or for imposing cross-equation constraints (Maddala 2002). On the other hand, a dummy variable as a regressand variable is used in the linear probability model, logit model, and probit model.

For more details on the topic, readers may refer to the list of references.

Cross References

►Linear Regression Models

References and Further Reading

Green WH (2002) Econometric analysis, 5th edn. Prentice Hall, Upper Saddle River

Maddala GS (2002) Introduction to econometrics, 3rd edn. Wiley, Chichester

Durbin–Watson Test

Walter Krämer
Professor and Chairman
Technische Universität Dortmund, Dortmund, Germany

The Durbin–Watson test is arguably, next to the method of ►least squares, the most widely applied procedure in all of statistics; it is routinely provided by most software packages and almost automatically applied in the analysis of economic time series when a researcher is fitting a linear regression model (see ►Linear Regression Models)

$$y = X\beta + u, \tag{1}$$

where y $(T \times 1)$ and X $(T \times K$, nonstochastic, rank $K)$ denote the vector of observations of the dependent and the matrix of observations of K independent (regressor-) variables, respectively, β is a $K \times 1$ vector of unknown regression coefficients to be estimated, and u $(T \times 1)$ is an unobservable vector of stochastic errors (disturbances, latent variables) with mean zero and equal variances. In the case of uncorrelated disturbances it is known from the ►Gauss-Markov-Theorem that the ordinary least squares estimator $\hat{\beta}$ for β, where $\hat{\beta} = (X'X)^{-1}X'y$ is best linear unbiased (BLUE) for β. In a time-series context, however, one often suspects that the components u_t of u might be correlated with each other, so it is of much interest to test whether the latter is indeed the case.

Due to the vastness of the alternative, there is no hope that a uniformly most powerful test exists (Anderson 1948) and one has to focus on more restricted alternatives. The one investigated by Durbin and Watson (1950, 1951, 1971) stipulates that the u_t's follow a stationary first order autoregressive process

$$u_t = \rho u_{t-1} + \epsilon_t \quad (|\rho| < 1, \epsilon_i \sim iid), \quad t = 2, \ldots, T \tag{2}$$

so the correlation matrix V of u is given by

$$V = \begin{bmatrix} 1 & \rho & & \rho^{T-1} \\ \rho & 1 & & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{t-2} & & 1 \end{bmatrix} \tag{3}$$

and the null hypothesis boils down to $H_0 : \rho = 0$. The DW test statistic is

$$d = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{T} \hat{u}_t^2} = \frac{\hat{u}' A \hat{u}}{\hat{u}'\hat{u}}, \tag{4}$$

where $\hat{u} = y - X\hat{\beta}$ and

$$A = \begin{bmatrix} 1 & -1 & & & 0 \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ 0 & & & -1 & 1 \end{bmatrix}. \tag{5}$$

It is easily checked that $0 \le d \le 4$. With positive serial correlation among the u_t's, neighboring u_t's and thus $\hat{u}_t$'s will tend to be close to each other, i.e., d will tend to be small. On the other hand, when ρ is negative, d will be large. A plausible two-sided test therefore rejects H_0 whenever d moves too far from 2, the mid-point of its range. One-sided tests will reject H_0 whenever d is too small $(H_1 : \rho > 0)$ or too large $(H_1 : \rho < 0)$.

Before the era of modern computers, the usefulness of this rule was compromised by the dependence of the null distribution on the regressor matrix X. Given T and K and selected significance levels, Durbin and Watson provided upper and lower bounds for the critical values; no conclusion was possible for a test statistic in between (the "inconclusive range"). Today, exact prob-values are easily available via some variant of the Imhof-algorithm (1961). Also, there exist competitors like the Breusch (1978)–Godfrey (1978) test which do not require fixed regressors or AR(1)-alternatives.

The theoretical foundation for the DW test dates back to Anderson (1948). He showed that whenever the disturbance vector u has density

$$f(u) = K\, exp\left[-\frac{1}{2\sigma^2}((1+\rho^2)u'u - 2\rho u'\phi u)\right], \qquad (6)$$

where K is some constant and ϕ is a symmetric $T \times T$ matrix, a uniformly most powerful test of $H_0 : \rho = 0$ against $H_1 : \rho > 0$ is given by

$$\frac{\hat{u}'\phi\hat{u}}{\hat{u}'\hat{u}} > k, \qquad (7)$$

provided the columns of X are linear combinations of eigenvectors of ϕ. This result is linked to the DW test as follows: The inverse of $\mathsf{Cov}(u)$ is

$$\frac{1}{\sigma_u^2}V^{-1} = \frac{1}{\sigma_\epsilon^2}\begin{bmatrix} 1 & -\rho & & & 0 \\ -\rho & 1+\rho^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 1+\rho^2 & -\rho \\ 0 & & & -\rho & 1 \end{bmatrix} \qquad (8)$$

$$= \tfrac{1}{2\sigma_\epsilon^2}[(1+\rho^2)I - 2\rho\phi - \rho(\rho-1)C],$$

where

$$\phi = \frac{1}{2}\begin{bmatrix} 1 & 1 & & & 0 \\ 1 & 0 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 0 & 1 \\ 0 & & & 1 & 1 \end{bmatrix}$$

and C= diag(1,0,…,0,1). Thus when we neglect the term $\rho(1-\rho)C$, the density of u is of the form (6), and from Anderson's result, an UMP rejection region against $H_1 : \rho > 0$ is obtained from large values of $\hat{u}'\phi\hat{u}/\hat{u}'\hat{u}$. Since $A = 2I - 2\phi$, this is equivalent to rejecting whenever $d = \frac{\hat{u}'A\hat{u}}{\hat{u}'\hat{u}}$ is too small.

A crucial condition for the approximate optimality of the DW test is that the column space of X be spanned by eigenvectors of ϕ. Much less is known about the power of the DW test when this condition fails. In fact, the power of the DW test can even drop to zero for certain regressors (Krämer 1985). This is so because for regressions without an intercept, d tends to some constant $\bar{d}$ as $\rho \to 1$. If $\bar{d}$ is less than the critical level $d^\star$ corresponding to the given X matrix and significance level, the limiting power of the DW test is 1, otherwise it is zero (neglecting the possibility that $\bar{d} = d^\star$).

About the Author

For biography *see* the entry ►Statistical Fallacies.

Cross References

►Approximations to Distributions
►Autocorrelation in Regression
►Best Linear Unbiased Estimation in Linear Models
►Gauss-Markov Theorem
►Least Squares
►Time Series Regression

References and Further Reading

Anderson TW (1948) On the theory of testing serial correlation. Skand Aktuarietidskr 31:88–116
Breusch TS (1978) Testing for autocorrelation in dynamic linear models. Aust Econ Pap 17:334–355
Durbin J, Watson GS (1950) Testing for serial correlation in least sqares regression I. Biometrika 37:409–428
Durbin J, Watson GS (1951) Testing for serial correlation in least sqares regression II. Biometrika 38:159–178
Durbin J, Watson GS (1971) Testing for serial correlation in least sqares regression III. Biometrika 58:1–19
Godfrey LG (1978) Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. Econometrica 46:1303–1310
Imhof JP (1961) Computing the distribution of quadratic forms in normal variables. Biometrika 48:419–426
Krämer W (1985) The power of the Durbin-Watson test for regression without an intercept. J Econom 28:363–370

E

Econometrics

Peter Kennedy
Professor Emeritus
Simon Fraser University, Burnaby, B.C., Canada

Several definitions of econometrics exist, a popular example being "Econometrics is the study of the application of statistical methods to the analysis of economic phenomena." The variety of definitions is due to econometricians wearing many different hats. First, and foremost, they are *economists*, capable of utilizing economic theory to improve their empirical analyses of the problems they address. At times they are *mathematicians*, formulating economic theory in ways that make it appropriate for statistical testing. At times they are *accountants*, concerned with the problem of finding and collecting economic data and relating theoretical economic variables to observable ones. At times they are *applied statisticians*, spending hours with the computer trying to estimate economic relationships or predict economic events. And at times they are *theoretical statisticians*, applying their skills to the development of statistical techniques appropriate to the empirical problems characterizing the science of economics. It is to the last of these roles that the term "econometric theory" applies, and it is on this aspect of econometrics that most textbooks on the subject focus.

The workhorse of econometrics is the linear regression model (see ▶Linear Regression Models), extended in a wide variety of nonlinear ways. So, for example, the log of wages might be regressed on explanatory variables such as years of education and a dummy for gender, with interest focusing on the slope of years of education, reflecting the return to investing in education, and the slope of the gender dummy, measuring discrimination in the labor market. Because this kind of empirical analysis is an important element of traditional statistics, one might naturally ask "How does econometrics differ from statistics?" There are two main differences between econometrics and statistics. The first stems from the fact that most economic data come from the real world rather than from controlled experiments, forcing econometricians to develop special techniques to deal with the unique statistical problems that accompany such data. The second difference is that econometricians believe that economic data reflect strategic behavior by the individuals and firms being observed, and so they employ models of human behavior to structure their data analyses. Statisticians are less willing to impose this kind of structure, mainly because doing so usually is not fully consistent with the data. Econometricians ignore such inconsistencies, so long as they are not gross, to enable them to address issues of interest. Some examples can illustrate these differences, the first four below resulting in Nobel prizes for econometricians.

1. Other things equal, females with young children earn higher wages than females without young children. Why? Females with children only appear in the labor market if their wage is large enough to entice them away from being a homemaker; this means that a sample of female wage earners is not a random sample of potential female workers – other things equal, low-wage earners are under-represented. This *self-selection* phenomenon causes bias in estimates of the wage equation, a bias that is of consequence because it narrows the estimated discrimination gap between male and female wages.
2. Economics is all about people making choices in a world of scarcity. People make choices (and thereby produce data for the econometrician) by maximizing an objective function. How do people choose which transportation mode to use to commute to work? Economic theory suggested addressing this question using the *random utility* model in which people choose on the basis of which option provides them with the greatest utility. Viewing the problem this way greatly enhanced the development of *multinomial logit/probit* analyses, resulting in markedly superior predictions of the ridership of the San Francisco bay area rapid transit system, then under construction.
3. For many years, in time series data it was thought that it did not make sense to employ levels and first differenced data in the same specification, and statisticians were careful to avoid doing this. But economic theory suggested that a certain combination of levels

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

data, representing an economic equilibrium position, could be compatible with differenced data. This concept of *cointegration* has changed forever the way in which time series analysis is undertaken.

4. Why would we ever be interested in how volatile some variable will be in the future? When pricing financial instruments such as options such information is crucial. An econometrician developed a way of addressing this problem, creating a major statistical industry, referred to as ARCH, *autoregressive conditional heteroskedasticity*.
5. Supply and demand curves are at the heart of economic analysis. When we regress quantity on price how do we know if the result is an estimated supply curve, an estimated demand curve, or garbage (a linear combination of the two)? To some the solution to this *identification* problem in the late 1940s marked the beginning of econometrics as a separate discipline. (But to others the founding of the Econometric Society and its journal *Econometrica* in the early 1930s marks the beginning of econometrics.)
6. The interaction of supply and demand illustrates another special feature of econometric work. Suppose we are estimating a demand curve and suppose amount demanded changes due to a bump in the error term in the demand equation. This shifts the demand curve, changes the intersection of the supply and demand curves, and so changes price. A regression of quantity on price produces biased estimates because the explanatory variable price is correlated with the error term in the demand equation. This dilemma of *simultaneous equation bias* is widespread, occurring whenever variables feed back to one another. *Instrumental variable* estimation, developed by econometricians, is a way of avoiding bias in this context.
7. When a sporting event is sold out, we don't know the actual demand for tickets; all we know is that the demand was at least as big as the venue's seating capacity. When estimating the determinants of demand what should we do with the observations on sold-out games? Estimation in the context of such *limited dependent variables* is another example of special estimation procedures developed by econometricians.

For a much more detailed overview of econometrics, including discussion of its historical development, see Geweke et al. (2008). For both students and practitioners, Kennedy (2008) is a very popular source of information on econometrics, offering intuition, skepticism, insights, humor, and practical advice. Section 1.1, beginning on p. 6, offers general commentary on "What is Econometrics?" and provides further references.

About the Author

Peter Kennedy received his BA from Queen's in 1965 and Ph.D. from Wisconsin in 1968. He is Professor Emeritus, Simon Fraser University. He held visiting positions at Cornell, Wisconsin, the London School of Economics, Singapore, Deakin, Cape Town, Canterbury, Curtin, Adelaide, Otago, and EERC (Ukraine). Peter Kennedy is best known as the author of *A Guide to Econometrics* (sixth edition, 2008, Wiley-Blackwell). He is the recipient of four awards for excellence in teaching, and one award for excellence in research. Upon retiring from Simon Fraser University in 2008, he stepped down as an associate editor of the *International Journal of Forecasting* and of *Economics Bulletin*, but continues to edit the research section of the *Journal of Economic Education*.

"Kennedy's Guide is a handbook of important concepts, techniques, and mathematical results that are developed at length in the standard textbooks but for which insights, interpretations, and intuition are missing. The popularity of his earlier three editions and their presence on students' shelves around the world documents the usefulness of Kennedy's ability to communicate the essence of econometrics without losing the reader in technical details" (Becker W. E., *The Journal of Economic Education*, 30, 1999, p. 89.).

Cross References

►Components of Statistics
►Econometrics: A Failed Science?
►Economic Statistics
►Linear Regression Models
►Principles Underlying Econometric Estimators for Identifying Causal Effects

References and Further Reading

Geweke JK, Horowitz JL, Pesaran MH (2008) Econometrics: a bird's eye view. In: Blume L, Durlauf S (eds) The new palgrave dictionary of economics, 2nd edn. Macmillan, London

Kennedy PE (2008) A guide to econometrics, 6th edn. Blackwell, Oxford

Econometrics: A Failed Science?

Jan Kmenta
Professor Emeritus
University of Michigan, Ann Arbor, MI, USA
Center for Economic Research and Graduate Education, Prague, Czech Republic

The official beginning of ►econometrics can be traced to the first issue of *Econometrica* (1933), the journal

dedicated to the "advancement of economic theory in its relation to statistics and mathematics." This makes it clear that the development of economic theory as a discursive process is to be supplemented by quantitative analysis and that econometrics is a part of economics. The earliest major attempts at specifying and estimating economic relationships include the famous – and to this day used – Cobb-Douglas production function (1928), the first macro-econometric models of Jan Tinbergen (1936 and 1939), and various unsuccessful attempts at estimating supply and demand functions in the 1930s.

The first systematic development of the mathematical formulation of economic theory, without which rigorous research in econometrics could not exist, is due to Paul Samuelson's *Foundations of Economic Analysis* (1947). The next step was the development of econometric methods largely due to the work of the researchers at the Cowles Commission at the University of Chicago (1939–1955). The resulting series of monographs included a solution to the identification problem that allowed separate estimations of supply and demand functions. An irony of fate is that the solution of the identification problem was provided already by Tinbergen in an overlooked article written in German and published in Vienna in 1930. The Cowles Commission monographs laid foundation to the formulation and estimation of simultaneous equation models that swamped the econometric literature in the 1960s and 1970s. All of this led W.C. Mitchell, one of the founders of the National Bureau for Economic Research (NBER), to declare in 1950 that "we must expect a recasting of old problems into new forms amenable to statistical attack."

The monographs of Cowles Commission would not have made an impact on the profession of economics and would not have led to the ensuing "scientific revolution" in economics – had they not given rise to a number of early popularizers who translated the sophisticated and technical language of the monographs into a language understandable to economists. Prominent among these were Gerhard Tintner, Henri Theil, Lawrence Klein, and Denis Sargan who carried the torch of econometric enlightenment in the 1950s and 1960s.

The subsequent two decades were marked by the introduction of the computer technology that enabled implementation of Monte Carlo experiments, formulation and estimation of large macro-econometric models, handling of large micro-data sets, and ►numerical integration of Bayesian inference. Indeed, at a conference in Princeton in 1972 computers were reported to have been of only moderate influence in various disciplines except economics. Of course, the use of computers required an accessory development of software. In this context a breakthrough was made by Harry Eisenpress who in 1962 compiled a program for computing maximum likelihood estimates of the coefficients of simultaneous equation models.

The introduction of computer technology really helped econometrics to take off. The macro-econometric modeling efforts were crowned by the Nobel Prize awarded to Lawrence Klein in 1980, the upsurge of micro-econometrics culminated with the award of the Nobel Prize to James Heckman and Daniel McFadden in 2000, and the modern "time series revolution" was blessed by the award of the Nobel Prize to Robert Engle and Clive Granger in 2003.

However, the golden years of econometrics were not without criticism. Articles and books with ominous sounding titles made an appearance that did not go unnoticed in the profession. Following are some of the titles appearing in the literature:

"Econometrics: Alchemy or Science?" (Hendry 1980)
"Lets Take the 'Con' out of Econometrics" (Leamer 1983)
"Data Mining" (Lovell 1983)
"What Will Take the Con out of Econometrics" (McAleer et al. 1985)
"The Foundations of Econometrics: Are There Any?" (Swamy et al. 1985)
Economics in Disarray (wiles and Routh 1985)
The Death of Economics (Ormerod 1994)

There were several bases for these criticisms, one of which was the poor specification and forecasting performance of econometric models. The sharpest criticism came from Sims (1980) who declared that "among academic macro-economists the conventional methods (of macro-econometric modeling), have not just been attacked, they have been discredited." Sims' response to the challenge was the introduction of the vector-autoregressive (VAR) model formulation (see ►Vector Autoregressive Models).

Another line of criticism, coming from David Hendry, was about estimation being the focus of econometric practice. According to Hendry (1980), "what should have been a relatively minor aspect of the subject, namely estimation, has been accorded the centre of the stage…the three golden rules of econometrics are test, test, test !" The common practice of "data mining," (see ►Data Mining) namely trying different model formulations until the estimated parameters conform to the pre-judgment of the investigator, also did not escape criticism. A sarcastic but fitting description of this practice is due to Leamer (1983) who noted that "if you torture data long enough, Nature will confess."

Hendry (1980) provided further elaboration of this point by a cute example, "Econometricians have found their Philosopher's stone. It is called regression analysis and it is used for transforming data into 'significant results'.

Deception is easily practiced." As an example, Hendry came up with his own 'theory of inflation' in which the price level is a function of a variable $C(t)$ and an auto regressive disturbance, where $C(t)$ is cumulative rainfall in the UK. The fit was spectacular, even better than if $C(t)$ was replaced by the quantity of money.

Perhaps the most serious criticism was leveled against the alleged meager contribution of econometrics to economic knowledge. The earliest criticism on this count came from Leontief (1971) who in his presidential address declared that, "in no other field of empirical inquiry has so massive and sophisticated statistical machinery been used with such indifferent results." Blaug (1980) elaborated on this theme by stating that, "empirical work that fails utterly to discriminate between competing explanations quickly degenerates into a sort of mindless instrumentalism and it is not too much to say that the bulk of empirical work in modern economics is guilty on that score."

The "time series revolution" in econometric practice, initiated by Engle and Granger in the 1970s, does not help in the direction of expanding economic knowledge either. On the contrary the mechanistic modeling of time series analysis pushes econometrics away from economics. This "revolution" is based on the contention – not rejected by statistical tests – that many economic variables have infinite variances, that is, that they grow without limit in time, and thus observed relationships maybe purely fictitious. The consoling fact is that while the variables themselves have infinite variances, typically the linear combinations of their first differences may well have variances that are finite. However, there are some serious problems with the way time series analysis is presented in the text books and applied in research. With respect to the former, the standard approach of teaching time series analysis ignores completely the classical econometrics, whose basic foundation is the regression model. A logical way would be to explain which of the assumptions of the classical regression model are likely to be violated when dealing with economic time series data. This would then be followed by discussion of the undesirable consequences of this on the properties of the least squares estimators and what to do about it. As it is, there is a complete disassociation between classical econometrics and time series econometrics. A lack of connection between time series analysis and economics as such is even more remarkable. The standard explanation of the long-run path of any economic variables as being purely determined by the passage of time plus a stochastic disturbance, with no reference to economic factors, is "primitive beyond words" (Kocenda and Cerny 2007). As for the so-called "co-integration analysis," uncovering that certain economic variables move together in stable relationships and looking for their economic interpretation is in contrast to the traditional deductive approach in which theory comes first and is followed by testing.

A final bone of contention regarding econometrics are the common results of empirical research that conflict with each other. This was pointedly illustrated by Leamer (1985) in his comments on the study of the deterrent of capital punishment …"there was a great outpouring of papers that showed that the results depend on (i) which variables are included; (ii) which observations are included; (iii) how simultaneity is treated, etc."

The source of failure of econometrics to satisfy its critics is undoubtedly due to, "a huge credibility gap between economic theory, empirical evidence and policy prescriptions" (Phillips 1988). The fault is on both sides; on the side of economic theorists as well as on the side of applied econometricians. Most economic theorists do not allow for the stochastic nature of the world, and most applied researchers tend to be very casual about the stochastic specifications of their models. Most of the work of economic theorists deals with deterministic models, while most of the applied workers treat the stochastic disturbance in their models as an afterthought barely worth mentioning. There are exceptions, of course, and they should be appreciated. In the field of micro-economics, the shining example of stochastic consciousness are the labor economists, while in the field of macro-economics the orchid goes to the dynamic stochastic general equilibrium models (DSGE), the modern successors of the simultaneous equation models of yesterday. On the whole, though, the gap between economic theory and empirical research is daunting. To rescue the reputation of econometrics and to "advance economic theory in relation to statistics and mathematics" it is, in my opinion, absolutely necessary for economic theory and econometrics to merge. There should be no separation between economic theory and econometrics as it by and large exists today.

About the Author

Jan Kmenta received his B.Ec with First Class Honors from the University of Sydney, Australia in 1955 and graduated from Stanford University with a Ph.D. in Economics with a minor in Statistics in 1964. He is a Fellow of the American Statistical Association since 1970 and a Fellow of the Econometric Society since 1980. Professor Kmenta has received the Alexander von Humboldt Foundation Award

for Senior U.S. Scientists (1979–1980). He was awarded an Honorary Doctorate from the University of Saarland, Germany (1989), and in 1998 he was the first recipient of the Karel Engliš Medal from the Academy of Sciences of the Czech Republic. In 2007 he received an annual award from the Czech Society of Economic for long-term contribution to the development of teaching of economics in the Czech Republic. Professor Kmenta has published a number of articles in prominent economic and statistical journals. He was ranked as 40th among all economists ranked by the total number of citations (1971–1985). Professor Kmenta is well known as the author of the internationally respected text *The Elements of Econometrics* (University of Michigan Press, 2nd edition, 1997), that was translated into several foreign languages.

Cross References

►Econometrics
►Linear Regression Models
►Seasonal Integration and Cointegration in Economic Time Series
►Time Series
►Vector Autoregressive Models

References and Further Reading

Blaug M (1980) The methodology of economics. Cambridge University Press, Cambridge
Hendry DF (1980) Econometrics: alchemy or science? Economica 47:387–406
Kocenda G, Cerny A (2007) Elements of time series econometrics. Charles University Press, Prague
Leamer EE (1983) Lets take the 'Con' Out of econometrics. Am Econ Rev 73:31–43
Leamer EE (1985) Sensitivity analysis would help. Am Econ Rev 75:308–313
Leontief W (1971) Theoretical assumptions and nonobserved facts. Am Econ Rev 61:1–7
Lovell MC (1983) Data mining. Rev Econ Stats 65:1–12
McAleer M et al (1985) What will take the con out of econometrics. Am Econ Rev 75:293–307
Mitchell WC (1950) The backward art of spending money. Augustus M. Kelley, New York
Ormerod P (1994) The death of economics. Wiley, New York
Phillips PCB (1988) Reflections on econometric methodology. Econ Rec 64:344–359
Sims C (1980) Macroeconomics and reality. Econometrica 48: 1–48
Swamy PAVB et al (1985) The foundations of econometrics: are there any? Econometric Reviews 4:1–62
Tinbergen J (1930) Bestimmung und Deutung von Angebotskurven: Ein Beispiel. Zeitschrift für Nat, Vienna 1(5)
Wiles P, Routh G (1985) Economics in disarray. Basil Blackwell, New York

Economic Growth and Well-Being: Statistical Perspective

Zarylbek I. Kudabaev
Chairman of the Statistical Society of Kyrgyz Republic, Professor, Head of Economic Department
American University of Central Asia, Bishkek, Kyrgyz Republic

► There is a clear case for complementing GDP with statistics covering the other economic, social and environmental issues, on which people's well-being critically depends.
European Commission (2009)

Access to well-developed, cross-cutting, high-quality shared information is a basic need of an effective democracy. The more information people have about economic and social conditions and trends in their countries, the better able they are to demand from politicians policies that improve economic growth and development. Another issue is the public's trust in its government. If people don't believe the figures that their official statistical bodies produce, they are likely to lose trust in government as a whole, which can have a negative impact on the democratic process.

So, the challenge is avoiding a gap between the official statistics on economic performance and the public's perception of their own living conditions. Such a gap has been evident over a long period in many countries around the world, but it is especially the case for the developing countries. Possible reasons for the gap between the official statistics and people's perceptions are misuse of certain statistics or the poor quality of official statistics.

Misuse of Certain Statistics

Statistical data produced for one purpose may not be appropriate for other purposes. A good example is gross domestic product (GDP), which tells us a nation's total income and the total expenditure on its output of goods and services and mainly is a measure of market production. GDP is not well suited as a metric of well-being, but it has been increasingly used for this purpose. GDP fails to capture such factors as unpaid work of households, distribution of income among the population groups, and depletion of resources, which creates a problem in using GDP (or GDP per capita) as an indicator of well-being. For example, if inequality in income distribution among population groups increases enough relative to the increase in average GDP per capita, then most people will be worse off even though average income is increasing. Or, if the number of cars drastically increases, then GDP goes up, but it

also leads to traffic jams and an increase in air pollution. In these examples, from the public's perspective, an increase in GDP does not improve their quality of life at all. For these reasons the original architects of the United Nations System of National Accounts (SNA) knew that well-being could not be measured by national income alone.

Low Quality of Official Statistics

Developing countries are more likely to produce official statistics of lower quality. Due to economic difficulties, developing countries struggle with making significant investments to develop statistical systems that are able to produce, using internationally recognized methodologies, a whole set of necessary indicators of economic and social progress.

The quality of official statistics might also be low because existing methodologies are imperfect. For example, consider how well statisticians can measure a non-observed, or shadow, economy. The economy of any country consist of two parts, observed and non-observed. The share of non-observed (shadow) economy in the GDP can be significant, especially in developing countries (up to 50% and higher). Also, it is well known that, in developing countries, many city-dwellers and internal migrants from rural areas are able to survive only because of their employment in the non-observed economy. Therefore, it is not possible to ignore a measure of non-observed economy in terms of both economic growth and development (well-being). It has to be separately estimated (measured) and then included in the country's GDP. There are several methodologies, based on different approaches, for measurement of the non-observed economy: conducting surveys of households on production of goods and services in the non-observed economy, labor force surveys, analysis of cash operations, analysis of electricity supply, and so on. But the problem is that all of the existing methodologies for measurement of the non-observed economy are not as good as those for the observed economy, and they should be improved significantly. The same criticism can applied to the statistical data on services, which should be improved as well.

What Should Be Done?

Statistical indicators for economic performance, although they are well developed, need to be improved in order to provide accurate monitoring of the evolution of modern economies (Stiglitz et al. 2009). Evolution toward a more complex economy means the growth of the share of services and production of a more complex quality of goods.

The "Going beyond GDP" strategy is very promising as a better measure of societal well-being, quality of life, and progress. "Going beyond GDP" does not mean dismissing GDP and production measures. It means complementing GDP with statistics on which people's well-being critically depends.

Well-being is multidimensional and includes not just economic but also social, environmental, and governance, and, as mentioned in (Stiglitz et al. 2009), the following dimensions should be considered simultaneously:

1. Material living standards (income, consumption, and wealth)
2. Health
3. Education
4. Personal activities including work
5. Political voice and governance
6. Social connections and relationships
7. Environment (present and future conditions)
8. Insecurity, of an economic as well as a physical nature

Real implementation of the "Going beyond GDP" strategy is a huge challenge for statisticians, and it likely requires a shift in emphasis from measuring economic production to measuring people's well-being.

Ideally, national accounts of well-being should be developed for the successful monitoring of social progress, as is already done for monitoring of economic progress (Michaelson et al. 2009). Forty-eight indicators of the Millennium Development Goals (MDG), in combination with the indicators of SNA, could be a good basis for construction of the national accounts of well-being.

How can developing countries improve their statistical products? International organizations and well-developed countries could play a key role. Strengthening the statistical capacities should become an important part of their programs of support for developing countries.

About the Author

Professor Zarylbek I. Kudabaev, Doctor of Economics and Ph.D in Physics. He is a Professor and Head of the Economics Department, American University of Central Asia (AUCA) in Bishkek (Kyrgyz Republic). He is Past Chairman (1997–2005) of the National Statistical Committee of the Kyrgyz Republic. He is a Chairman of the Statistical Society of Kyrgyz Republic. He was a member of Steering Committee of the International Consortium "Partnership in Statistics for Development in the 21st Century" (PARIS21, 2000–2005). In 2002 he was a Chairman of the Council of the Heads of Statistical Services of CIS countries. He is National Coordinator for subscribing the Kyrgyz Republic to the International Monetary Fund's (IMF) General Data Dissemination Standard (2001) and IMF's Special Data Dissemination Standard (2004). Professor

Kudabaev is author of 5 monographs and more than 100 scientific articles on statistics and economic development.

Cross References

►Economic Statistics
►Measurement of Economic Progress

References and Further Reading

European Commission (2009) GDP and beyond: measuring progress in a changing world, communication from the Commission to the Council and the European Parliament, Brussels, COM (2009) 433 final

Michaelson J, Abdallah S, Steuer N, Thompson S, Marks N, Aked J, Cordon C, Potts R (2009) National accounts of well-being: bringing real wealth onto the balance sheet. The New Economics Foundation, London (available at www.neweconomics.org)

Stiglitz JE, Sen A, Fitoussi J-P (2009) Report by the commission on the measurement of economic performance and social progress. (available at www.stiglitz-sen-fitoussi.fr)

Economic Statistics

Yasuto Yoshizoe
Professor, President of the Japan Statistical Society
Aoyama Gakuin University, Tokyo, Japan

Introduction

In this article, we regard economic statistics as a branch of applied statistics which deals with the following topics: (1) collection of statistics on socioeconomic conditions, (2) compilation of surveys and registered records to produce various economic indicators, such as consumer price index or GDP, (3) evaluation of economic indicators from the viewpoint of reliability.

Economic statistics is closely related to official statistics, since most of the statistics of the society and economy are provided by official organizations like national statistical offices or intergovernmental organizations such as United Nations, OECD, and World Bank. On the other hand, economic statistics is different from econometrics in a narrow sense. Typically, objective of econometrics lies in developing the theory and its applications to various economic data. In contrast, economic statistics places more emphasis on quality of data before applying sophisticated methods to analyze them. In other words, we are more interested in appropriate interpretation of the data paying attention to their detailed characteristics.

Here, we describe some typical issues from economic statistics: censuses, sample surveys, index numbers, system of national accounts, followed by an illustrative example.

Censuses – Complete Enumeration

Consumers and producers consist of two major components in economics. Correspondingly, two fundamental information of economic statistics are (1) population and households, and (2) firms and establishments. U.S. Census Bureau, http://www.census.gov/econ/, provides definitions of firms and establishments among other related concepts.

To construct reliable statistics on these subjects, the complete enumeration is required. A ►*census* is the procedure of collecting information about all members of a population. In many countries, population censuses are carried out periodically. In contrast, information of firms and establishments are obtained either by statistical surveys (so-called economic censuses) or by registered information collected through some legal requirement.

The complete enumeration is necessary to provide accurate information for small areas. In planning the location of elementary schools, hospitals or care agencies for elderly people, local governments need information for small areas. If the censuses or registration records provide accurate information, local governments can depend on them in making important policies.

Another role of the census is that it provides a list of households or firms/establishments. Such a list is used as the *sampling frame*, and it plays an essential role in many sampling surveys described in the next section. In this sense, inaccurate censuses, or administrative records that are out of date, reduces the reliability of sample surveys, and eventually increase social costs.

For household surveys, efficient sampling designs, such as stratified sampling, become difficult if accurate sampling frame is not available. The inaccurate sampling frame can sometimes cause extremely biased estimates for business surveys due to frequent entry/exit behavior of firms/establishments. An example is given later.

Sample Surveys

Since economic conditions vary dynamically, we need monthly or quarterly indicators, such as consumer price index or industrial production index. These statistics have to be compiled from sample surveys on households and/or establishments. Therefore, the accuracy of sampling survey is crucial in maintaining reliability of economic indicators. Major factors that affect performance of the estimators are: sampling design (e.g., simple random sampling, stratified sampling, cluster sampling, etc.) and estimation methods (e.g., linear estimator, ratio estimator, regression estimator, etc.). See Cochran (1977), Groves et al. (2004), Lesser and Kalsbeek (1992), and Särndal et al. (1992) for detail.

The auxiliary information available from censuses are used to set up appropriate strata, or to apply a ratio estimation, resulting in more efficient and inexpensive surveys. An example of ratio estimator is to use the size of population who are older than 15 years (Y) to estimate the number of employees (X). Let the sample means are $\bar{x}$ and $\bar{y}$ in a sample of size n from a population of size N. In case of simple random sampling, the linear estimator is $\hat{X} = N\bar{x}$. On the other hand, if population Y is obtained from a census and additional registration records, then the ratio estimator defined by $\hat{X}_r = Y(\bar{x}/\bar{y})$ can be applied. Since X and Y are correlated, $\hat{X}_r$ has a smaller variance than $\hat{X}$.

More important and difficult task is that we need to evaluate the *non-sampling errors*, such as nonresponse caused by many reasons, possible bias from a deteriorated sampling frame, etc. In the example above, the ratio estimator may become worse if Y is incorrectly measured.

As for business surveys, about 5% of establishments start business and another 5% shut down annually in case of Japan. It is clear that maintaining accurate list of establishments is vital for high quality of economic statistics.

Index Numbers

►Index numbers are widely used to compare levels and evaluate changes of economic variables over time or among regions (see Schultze and Mackie 2002; ILO 2004). Typically, the base value equals 100, but it can be other values. For example, the real GDP can be regarded as a quantity index, if we choose its base value as the nominal GDP of the base year. Most often, index numbers are used to compare economic variables, such as unemployment or prices, over time. In some cases, however, index numbers may compare geographic areas at a specific time. Such an example is the *purchasing power parity*, which is an international comparison of prices among countries.

As for price index and quantity index, the most well-known formulas are Laspeyres, Paasche, and Fisher. If we deal with $i = 1, \ldots, n$ commodities (goods or services), we need to collect data of prices p_{it} and quantities q_{it} at time t. Then the three price indexes (base period is $t = 0$) are defined as Laspeyres: $P_L = \sum p_{it}q_{i0}/\sum p_{i0}q_{i0}$, Paasche: $P_P = \sum p_{it}q_{it}/\sum p_{i0}q_{it}$, and Fisher: $P_F = \sqrt{P_L P_P}$. Similarly, quantity indexes are defined by interchanging p and q in the definitions.

These formulas are, however, too superficial for practical purposes. In calculating consumer price index (CPI), national statistical offices of developed countries adopt much more complicated methods using vast amount of related data. The price of a specific merchandize (a pack of milk, for example) varies among stores and areas. Moreover, there are several kinds of milk that consumers purchase in a given period. The observed prices are averaged by some method before the price is used as p_{it} in an index formula like Laspeyres. Hence, we should review and evaluate the appropriateness of the specific method employed. The judgment also depends on the data collection procedure of prices and weights, which varies among national statistical offices.

Since CPI has a rigorous background in economic theory, there is a good amount of reference, including the definition of the *true index number*. Helped by the theory, some statistical offices provide, although approximately, the *superlative price index*, which has enhanced the reliability of CPI.

In other applications of index numbers, however, the interpretation cannot be so exact, and the quality of an index is judged by coverage of commodities and reliability of data sources used in calculating the index. For example, the Bank of Japan publishes a monthly price index called Corporate Goods Price Index (CGPI). It is a Laspeyres price index, but the prices are mixtures of producers prices and wholesale prices. Thus, despite the title, experienced economists do not regard summary CGPI (for all commodities) as a price index but an indicator of business activities. On the other hand, since the data collection process guarantees high reliability of CGPI data, collected prices are quite important for many official statisticians to calculate other statistics involving prices.

System of National Accounts

The purpose of the United Nations System of National Accounts (SNA) is to provide an integrated accounts of significant economic activities to make international or temporal comparisons possible (Lequiller and Blades 2006). The most important implication of SNA to economic statistics is that the system provides the conceptual and actual basis to judge and achieve coherence among all statistical sources available in a country. In this sense, national accounts are the core of modern economic statistics.

Since the data from national accounts are so rich and full, it is not easy for economists to understand the whole system. To make things more complicated, some important changes have been made in the new version called SNA 2003 followed by a revision (SNA 2008). The difference from the earlier version becomes quite important for drawing meaningful analytical conclusions when we are interested in making inflation forecasts or in assessing economic growth capabilities.

Rigorously speaking, the national accounts system adopted by national statistical offices differ in some extent from country to country. Therefore, we need to make sure whether the differential in economic growth observed in the 1990s between the United States and the European Union is real or not. The answer is that it just reflected the different treatment of prices of the products for information and communication technology (ICT), where rapid technical changes have been taking place.

The quality of national accounts depends heavily on the statistical system of the country. To be more specific, GDP and other national accounts are not the result of a single survey, but the result of combining a mixture of various data from many sources. Since it is practically impossible for national accounts to cover all units in a country, a significant number of adjustments has to be made. In the process of adjusting those data to achieve coherence with SNA, quite often, some obscure methods are applied. Hence, it is quite difficult to obtain a formal assessment on the accuracy of the GDP, and an attempt to construct an interval estimation is almost meaningless. National accounts should be regarded at best as approximations.

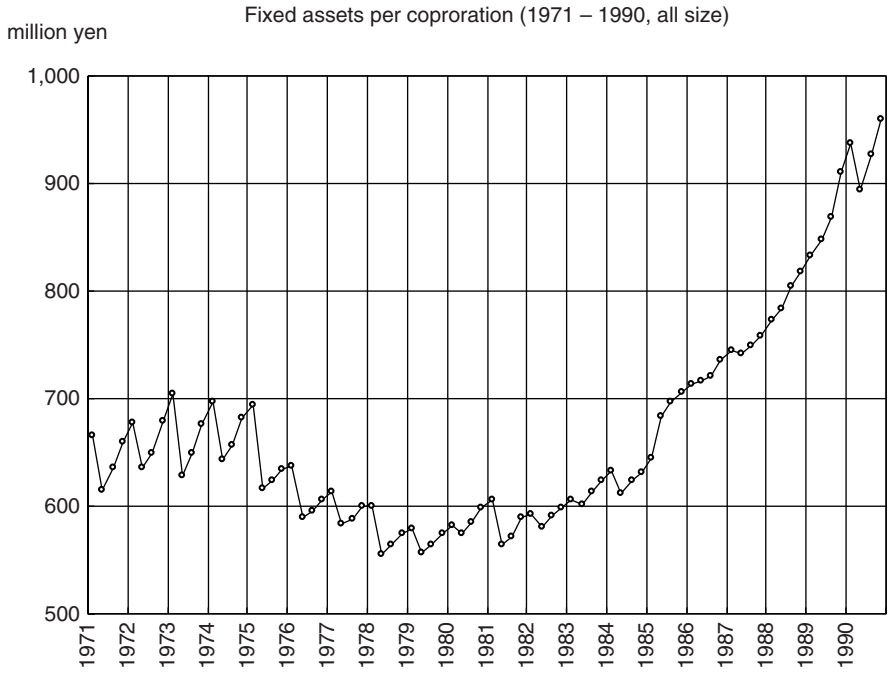
Another current issue is increasing publication of quarterly accounts in many countries. One of the essential objectives of economic statistics is to provide appropriate information to policy-makers at the right moment. In particular, it is requested to provide the newest information regarding the business cycle. In this context, traditional annual national accounts are not useful, and the demand for quarterly accounts has been increasing. The availability and the reliability of the statistics or administrative records is the key to reliable quarterly GDP, but in some countries, even monthly GDP are compiled by national statistical offices or by private research institutions.

Today, it is well-known that GDP is not a single measure of well-being, and it has many drawbacks. Yet, GDP remains to be the most important economic statistics as it defines a standard. To conclude, we should be aware of the usefulness and limitations of GDP and other economic statistics.

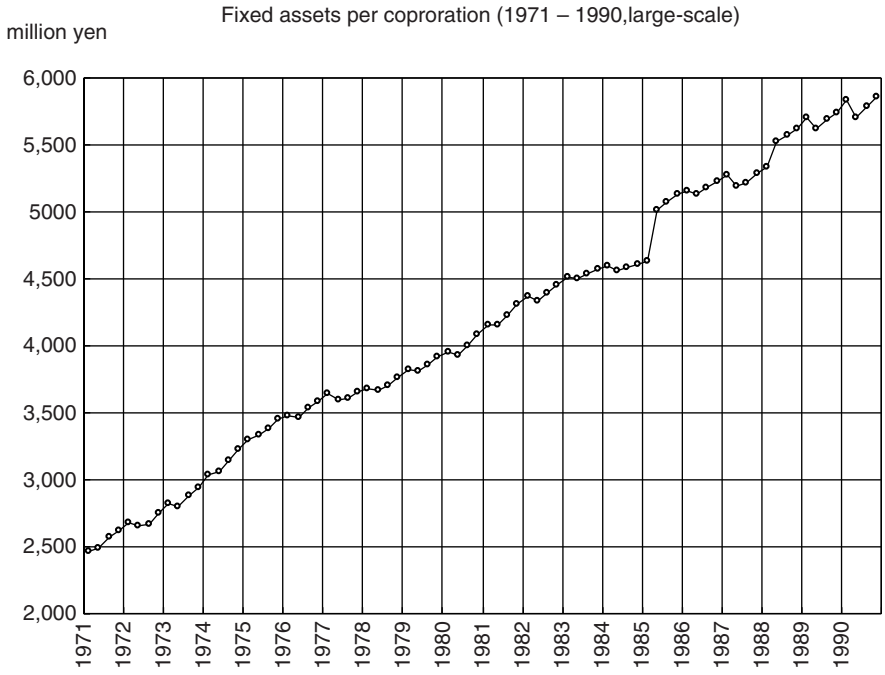
An Example

In this section, we describe some issues in the statistics of corporation. Refer to Yoshizoe et al. (2007) for a more detailed description.

The Japanese Ministry of Finance publishes *Financial Statement Statistics of Corporations by Industry* consisting of *Annual Survey* and *Quarterly Survey*, whose accuracy is vital for the government to judge the current economic situation. The large-scale corporations are quite few in number (the population ratio is 0.5%), but their influence is predominant (37.5% in sales, 56.9% in profit, and 52.0%



Economic Statistics. Fig. 1 All corporations



Economic Statistics. Fig. 2 Large-scale corporations

in fixed assets, in 2005). Nevertheless, trends of small-scale corporations are also important, especially in economic forecasts.

One of the problems is that some statistics related to small-scale corporations occasionally show unnatural patterns. Figures 1 and 2 show "fixed assets per corporation" of all corporations and those of large-scale corporations, respectively. Figure 1 indicates considerable drops each year from January–March period to April–June period. On the other hand, Fig. 2 for large-scale corporations (enumerated completely) does not show noticeable drops. We found out that the gaps in Fig. 1 was caused by the sampling design.

Since the response burden is rather heavy for small corporations, sample corporations are replaced after a year. Thus, newly sampled corporations start to respond to the questionnaires in April–June period. During the survey period (one year), some corporations go out of business and drop out of the survey. On the other hand, newly established corporations which start operating in that year cannot be included in the survey until the new list of corporations becomes available. The fact that these new entrants are usually smaller in size means that in *Quarterly Survey*, the corporations that survive and respond throughout the survey period tend to have better business performance than average. This in turn implies that, in the sample, indicators such as fixed capital per corporation will gradually have upward bias from April–June to January–March in the next year. When corporations are selected from the new list which becomes available in April, the new sample represent the population more appropriately. Thus, the sampling design explains a major portion of the gaps in the statistics.

We also examined the way corporations provide financial statements as another possible source of the gap. While large-scale corporations prepare quarterly statements, most small-scale corporations only prepare annual statements reflecting the difference of legal requirement. Moreover, it is often said that interim financial results are systematically different from the final financial statement even for large corporations. If we look at Fig. 2 carefully, we can find occasional gaps from January–March to April–June. The difference between provisional settlement of account and the final statement explains the gaps for larger corporations, hence a similar systematic behavior may partially explain the gaps of small-scale corporations. Further analyses and possible solutions for the gap problem can be found in the reference given above.

About the Author

Professor Yoshizoe is President, Japan Statistical Society (2009-present).

Cross References

- ►Business Statistics
- ►Composite Indicators
- ►Econometrics
- ►Index Numbers
- ►National Account Statistics
- ►Social Statistics

References and Further Reading

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Groves RM et al (2004) Survey methodology. Wiley, New York

International Labour Organization (2004) Consumer price index manual: theory and practice. International Labour Office, Geneva

Lequiller F, Blades D (2006) Understanding national accounts. OECD, Paris

Lesser JT, Kalsbeek WD (1992) Nonsampling error in surveys. Wiley, New York

Särndal C-E, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York

Schultze CL, Mackie C (eds) (2002) At what price? Conceptualizing and measuring cost-of-living indexes. National Academy Press, Washington, DC

Yoshizoe Y et al (2007) Correcting non-sampling errors in Financial statement statistics of Japanese Ministry of Finance. Proceedings of the 56th meeting of the International Statistical Institute in Lisbon, Portugal. http://www.yoshizoe-stat.jp/english/isi2007_yoshizoe.pdf

Edgeworth Expansion

Zhidong Bai
Professor
Northeast Normal University, Changchun, China

Introduction

In many statistical problems, for example, hypothesis testing and confidence region estimation, it is necessary to know the distribution functions of statistics of interest. When they are not exactly known, one might like to have an approximation of the unknown distribution as accurately as possible. One of the possibilities is the normal approximation. However, it is proved that normal approximation has an unimprovable rate of $O(n^{-1/2})$. To obtain a more accurate approximation for the distribution function, F_n, of a random variable X_n, Chebyshev (1890) expanded the function $\psi(t)/\phi(t)$ (definition given below) into a power series of (it) and obtained a formal asymptotic expansion of F_n. Edgeworth (1905, 1907) expanded $\chi(t,u)$ into a power series of u and then, by setting $u = 1$, obtained another formal expansion that is equivalent to a rearrangement of the Chebyshev expansion. Although the two are equivalent in the sense of formal expansion, the Edgeworth expansion (EE) provides an optimal approximation when only a finite number of moments of X_n exist.

Formal Expansion

Suppose that all moments of X_n exist and additionally all *cumulants* (also called *semi-invariants*) ν_{nk} of X_n exist. In this case, the characteristic function of X_n can be written as

$$\psi_n(t) = \exp\left(\sum_{r=1}^{\infty} \frac{\nu_{nr}(it)^r}{r!}\right) = \phi(t)\exp\left(\sum_{r=3}^{\infty} \frac{\nu_{nr}(it)^r}{r!}\right),$$

where $\phi(t) = \exp\left(-i\mu t - \frac{t^2\sigma^2}{2}\right)$, $\mu = \nu_{n1}$ and $\sigma^2 = \nu_{n2}$. In applications, the parameters μ and σ^2 are independent of n and $\nu_{nr}, r \geq 3$, has the order of $n^{-(r-2)/2}$.

Expand $\exp\left(\sum_{r=3}^{\infty} \frac{\nu_{nr}(it)^r u^{r-2}}{r!}\right)$ into a power series in u:

$$\chi(t,u) = \exp\left(\sum_{r=3}^{\infty} \frac{\nu_{nr}(it)^r u^{r-2}}{r!}\right) = 1 + \sum_{j=1}^{\infty} p_{nj}(it)u^j, \quad (1)$$

where $p_{nj}(x)$ is a polynomial with powers of x ranging between $j+2$ and $3j$ and coefficients depending on the cumulants ν_{nr}, $r = 3, 4, \ldots, j+2$. Some examples:

$$p_{n1}(it) = \frac{(it)^3\nu_{n3}}{6},\ p_{n2}(it) = \frac{(it)^4\nu_{n4}}{24} + \frac{(it)^6\nu_{n3}^2}{72},\ p_{n3}(it) = \frac{(it)^5\nu_{n5}}{120} + \frac{(it)^7\nu_{n3}\nu_{n4}}{144}.$$

Differentiating r times with respect to x both sides of the formula $\frac{1}{2\pi}\int e^{-itx}e^{-t^2/2}dt = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \equiv \varphi(x)$, the density of the standard normal variable, we have

$$\frac{1}{2\pi}\int (it)^r e^{-itx-t^2/2}dt = (-1)^r\frac{d^r}{dx^r}\varphi(x) = H_r(x)\varphi(x),$$

where H_r is the r-th Chebyshev–Hermitian polynomial. From this, one can easily derive that the inverse Fourier transform of $(it)^r\exp(it\mu - t^2\sigma^2/2)$ is $\sigma^{-r}H_r\left((x-\mu)/\sigma\right)\varphi\left((x-\mu)/\sigma\right)$.

Write $p_{nj}(it) = \sum_{r=j+2}^{3j} b_{nr}(it)^r$. By noting that

$$\psi_n(t) = \phi(t)\left(1 + \sum_{j=1}^{\infty} p_{nj}(it)\right),$$

the distribution function of X_n can be written as

$$F_n(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) - \sum_{j=1}^{\infty} Q_{nj}\left(\frac{x-\mu}{\sigma}\right)\varphi\left(\frac{x-\mu}{\sigma}\right),$$

where $Q_{nj}(x) = \sum_{r=j+2}^{3j} b_{nr}\sigma^{-r}H_{r-1}(x)$. This is called the *formal Edgeworth expansion* (FEE) of $F_n(x)$. The first success of rigorous theory was done by Cramér (1928).

Validity of Edgeworth Expansion

The FEE may not be valid in three senses: (1) in real applications, not all orders of moments exist and thus the FEE is not well-defined; (2) even if all terms of FEE are well-defined, the series may not converge; and (3) the equality may not be true even if the series converges. To this end, one needs to establish a valid approximation when only the first k ($k \geq 3$) moments are finite. In this case, all terms Q_j, $j \leq k-2$ in the FEE are well-defined and thus, our task reduces to establishing error bounds for

$$R_n(x) = F_n(x) - \left[\Phi\left(\frac{x-\mu}{\sigma}\right) - \sum_{j=1}^{k-2} Q_j\left(\frac{x-\mu}{\sigma}\right)\varphi\left(\frac{x-\mu}{\sigma}\right)\right]. \tag{2}$$

If the error bound for $R_n(x)$ is given by the form $Kg(n,k)$, the bound is then called a *uniform estimation*; if the bound is of the form $Kg(n,k)(1+|x|)^{-k}$, the bound is then called a *nonuniform estimation*, where K is an absolute constant and $g(n,k) \to 0$ describing the order of the convergence. The quantity in the brackets is called the $(k-2)$-term EE of $F_n(x)$ or X_n.

Sum of Independent Random Variables

Suppose that for each n, $\{X_{nj}, j = 1,2,\ldots,n\}$ is a sequence of random variables with means 0, variances σ_{nj}^2, finite k-th moments and ▶characteristic functions $v_{nj}(t)$ satisfying

$$\limsup_{|t|\to\infty} \sup_n \frac{1}{n}\sum_{j=1}^{n} |v_{nj}(t)| < 1 \tag{3}$$

(called the *Cramér condition*), then for the distribution function $F_n(x)$ of $S_n = B_n^{-1}\sum_{j=1}^n X_{nj}$, we have

$$|R_n(x)| \leq B\left(\frac{\sum_{k=1}^n \mathrm{E}|Z_{njx}|^k}{B_n^k(1+|x|)^k} + \frac{\sum_{j=1}^n \mathrm{E}|W_{njx}|^{k+1}}{B_n^{k+1}(1+|x|)^{k+1}}\right),$$

where $B_n^2 = \sum_{j=1}^n \sigma_{nj}^2$, $Z_{njx} = X_{nj}I(|X_{nj}| > B_n(1+|x|)$ and $W_{njx} = X_{nj}I(|X_{nj}| \leq B_n(1+|x|)$. If the r-cumulants of X_{nj} is denoted by v_{jr}, then the cumulants of S_n is given by $\mu = 0$, $\sigma^2 = 1$ and for $r \geq 3$, $v_{nr} = B_n^{-r}\sum_{j=1}^n v_{jr}$.

When the random variables are independent and identically distributed (IID) with mean 0 and variance σ^2 and r-th cumulant v_r, we have $v_{nr} = n^{-(r-2)/2}v_r/\sigma^r$ and thus $Q_{nj}(x) = n^{-j/2}q_j(x)$ the EE of F_n.

Function of Sample Means

Suppose that $\{\mathbf{Y}_1,\ldots,\mathbf{Y}_n\}$ is a sequence of iid random m-vectors with mean vector $\boldsymbol{\mu}$, covariance matrix Σ and finite k-th moment. Assume the function H defined on R^m is $k-1$ times continuously differentiable in a neighborhood of $\boldsymbol{\mu}$. By the Taylor expansion of $H(\bar{\mathbf{Y}})$, one has

$$\sqrt{n}\left(H(\bar{\mathbf{Y}}) - H(\boldsymbol{\mu})\right) = \sum_{r=1}^{k-1}\frac{\sqrt{n}}{r!}\sum_{i_1+\cdots+i_m=r} l_{i_1,\ldots,i_m} \prod_{j=1}^{m} i_j!(\bar{Y}_{(j)} - \mu_{(j)})^{i_j} + R_k, \tag{4}$$

where $l_{i_1,\ldots,i_m} = \frac{\partial^r H(\boldsymbol{\mu})}{\partial\mu_{(1)}^{i_1}\cdots\partial\mu_{(m)}^{i_m}}$, $\bar{Y}_{(j)}$ and $\mu_{(j)}$ are the j-th components of $\bar{\mathbf{Y}}$ and $\boldsymbol{\mu}$, respectively. If the cumulants are defined by those of the first term on the right-hand side of Eq. (1), then the $(k-2)$-term EE for $\sqrt{n}(H(\bar{\mathbf{Y}}) - H(\boldsymbol{\mu}))$ is valid with a uniform bound $o(n^{-(k-2)/2})$ if $\sigma^2 > 0$ and the Cramélr condition

$$\limsup_{\|\mathbf{t}\|\to\infty} |\mathrm{E}e^{i\mathbf{t}'\mathbf{Y}_1}| < 1, \tag{5}$$

holds, where $\sigma^2 = \boldsymbol{\ell}'\Sigma\boldsymbol{\ell}$ and the j-th component of $\boldsymbol{\ell}$ is $\frac{\partial H(\boldsymbol{\mu})}{\partial\mu_j}$.

Partial Cramér Condition

Condition (3) may not hold when $\mathbf{Y}_1$ contains a discrete component and the moment condition may not be minimal when some component of $\boldsymbol{\ell}$ is 0. To this end, we have the following theorem.

Theorem 1 *The $(k-2)$-term EE of $\sqrt{n}\left(H(\bar{\mathbf{Y}}) - H(\boldsymbol{\mu})\right)$ is valid with a uniform bound $o(n^{-(k-2)/2})$ if the following conditions hold:*

(1) *The following partial Cramér condition holds*

$$\limsup_{|t|\to\infty} \mathrm{E}|\mathrm{E}(e^{itY_{(1)}}|Y_{(2)},\ldots,Y_{(m)})| < 1.$$

(2) *There is an integer $p < m$ such that $\ell_1 \neq 0$ and $l_{0,\ldots,0,i_{p+1},\ldots,i_m} = 0$, provided $i_{p+1}+\cdots+i_m \geq 1$, where ℓ_1 is the first component of $\boldsymbol{\ell}$.*

(3) $\mathrm{E}|Y_{(j)}|^k < \infty$ *for* $j = 1,\ldots,p$ *and* $\mathrm{E}|Y_{(j)}|^{k/2} < \infty$.

Cornish–Fisher EE

For a constant $\alpha \in (0,1)$, denote by $\xi = \xi_{n,\alpha}$ and $z = z_\alpha$ the quantiles of F_n and $\Phi\left(\frac{z-\mu}{\sigma}\right)$, respectively. In view of the EE of F_n, one may expect ξ can be approximated by z of the form

$$\xi = z + \sum_{j=1}^{k-2}\eta_{n,j}n^{-j/2} + o(n^{-(k-2)/2}), \tag{6}$$

which is called the Cornish–Fisher EE of quantiles. Indeed, the coefficients $\eta_{n,j}$ can be determined by submitting (6)

into the valid EE of F_n, that is,

$$F_n(\xi) = \Phi\left(\tfrac{\xi-\mu}{\sigma}\right) - \sum_{j=1}^{k-2} Q_{n,j}\left(\tfrac{\xi-\mu}{\sigma}\right)\varphi\left(\tfrac{\xi-\mu}{\sigma}\right) + R_n.$$

As an example, for the iid case with $\mu = 0$ and $\sigma = 1$, we give expressions of the first three η_j as follows

$$\eta_1 = \frac{1}{6}\nu_3(z^2-1),\ \eta_2 = \frac{\nu_4}{24}(z^3-3z) - \frac{\nu_3^2}{36}(2z^3-5z)$$
$$\eta_3 = \frac{\nu_5}{120}(z^4-6z^2+3) - \frac{\nu_3\nu_4}{24}(4z^5-2z^3+5z) + \frac{\nu_3^3}{324}(12z^4-53z^2+17).$$

About the Author

"Zhidong has a legendary life. He was adopted into a poor peasant family at birth. He spent his childhood during the Chinese resistance war against Japan." (*Advances in statistics: proceedings of the conference in honor of Professor Zhidong Bai on His 65th Birthday*, Eds. Zehua Chen, Jin-ting Zhang, Feifang Hu, World Scientific Publishing Company, 2008, Preface). Zhidong Bai has started his illustrious career as a truck driver's team leader during the Cultural Revolution in China (1968–1978), but he managed to become a Professor of statistics, North East Normal University, China, and Department of Statistics and Applied Probability, National University of Singapore. He is a Fellow of the Institute of Mathematical Statistics (1995), and Fellow of the Third World Academy of Science (1990). Dr. Bai was the Editor of the *Journal of Multivariate Analysis* (1991–2001), and Associate editor of the *Journal of Statistical Planning and Inference* (2001–2003) and *Statistica Sinica* (1991–2000). Currently, he is the Associate editor of the *Journal of Probability and Statistics* and *Sankya*. He has published seven papers related to Edgeworth expansion (he was the first who proposed the concept of Partial Cramér condition, and published a series of papers in *Ann. Statist.*, *JMVA* and *Sankya* to solve a series of problems), and over 120 refereed papers in other areas (including 10 papers with Professor C.R. Rao). He has also (co)authored three books, including *Ranked set sampling. Theory and applications* (with Chen, Z.H. and Sinha, B.K., Springer-Verlag, New York, 2004). In 2008, the Conference in Honor of Professor Zhidong Bai on his 65th Birthday was held at the National University of Singapore.

"Professor Bai represents a special generation who experienced the most difficult period in China but never gave up their quest for science and eventually excelled in the world arena. This special issue is not only to honor Professor Bai for his great achievements but also to inspire the young generation to devote themselves to the promotion of Statistics in China as China enters into a great new era." (Loc. cit.)

Zhidong Bai is a professor of the School of Mathematics and Statistics at Northeast Normal University and Department of Statistics and Applied Probability at National University of Singapore. He is a Fellow of the Third World Academy of Sciences and a Fellow of the Institute of Mathematical Statistics.

Cross References

►Approximations for Densities of Sufficient Estimators
►Approximations to Distributions
►Cornish-Fisher Expansions
►Multivariate Statistical Distributions

References and Further Reading

Babu GJ, Bai ZD (1993) Edgeworth expansions of a function of sample means under minimal moment conditions and partial Cramèr's condition. Sankhya Ser A 55(2):244–258
Bai ZD, Rao CR (1991) Edgeworth expansion of a function of sample means. Ann Stat 19(3):1295–1315
Bai ZD, Zhao LC (1986) Asymptotic expansions of sums of independent random variables. Sci Sin 29(1):1–22
Battacharya RN, Ghosh JK (1978) On the validaty of the formal Edgeworth expansion. Ann Stat 6(2):434–451
Chebyshev PL (1890) Sur deux theoremes relatifs aux probabilites. Acta Math 14:305–315
Cramér H (1928) On the composition of elementary errors. Skand Aktuarietidskr 11:13–74, 141–180
Edgeworth FY (1905) The law of error. Trans Camb Philos Soc 20:36–65, 113–141
Edgeworth FY (1907) On the representatin of a statistical frequency by a series. J R Stat Soc 70:102–106
Fisher SRA, Cornish EA (1960) The percentile points of distributions having known cumulants. Technometrics 2:209–225
Hall P (1987) Edgeworth expansion for Student's t statistic under minimal moment conditions. Ann Probab 15(3):920–931
Petrov VV (1975) Sums of independent random variables (trans from Russian: Brown AA). Springer, New York/Heidelberg

Effect Modification and Biological Interaction

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

The term *effect modification* has been applied to two distinct phenomena. For the first phenomenon, effect modification simply means that some chosen measure of effect

varies across levels of background variables. This phenomenon is thus more precisely termed effect-measure modification, and in the statistics literature is more often termed heterogeneity or "interaction" (Greenland et al. 2008), although "interaction" is more often used as a synonym for a product term in a regression model (*see* **Interaction**). For the second phenomenon, effect modification means that the mechanism of effect differs with background variables, which is known in the biomedical literature as *dependent action* or (again) "interaction." The two phenomena are often confused, as reflected by the use of the same terms ("effect modification" and "interaction") for both. In fact they have only limited points of contact.

Effect-Measure Modification (Heterogeneity of Effect)

To make the concepts and distinctions precise, suppose we are studying the effects that changes in a variable X will have on a subsequent variable Y, in the presence of a background variable Z that precedes X and Y. For example, X might be treatment level such as dose or treatment arm, Y might be a health outcome variable such as life expectancy following treatment, and Z might be sex (1 = female, 0 = male). To measure effects, write Y_x for the outcome one would have if administered treatment level x of X; for example, if $X = 1$ for active treatment, $X = 0$ for placebo, then Y_1 is the outcome a subject will have if $X = 1$ is administered, and Y_0 is the outcome a subject will have if $X = 0$ is administered. The Y_x are often called *potential outcomes* (*see* **Causation and causal inference**).

One measure of the effect of changing X from 0 to 1 on the outcome is the difference $Y_1 - Y_0$; for example, if Y were life expectancy, $Y_1 - Y_0$ would be the change in life expectancy. If this difference varied with sex in a systematic fashion, one could say that the difference was modified by sex, or that there was heterogeneity of the difference across sex. Another common measure of effect is the ratio Y_1/Y_0; if this ratio varied with sex in a systematic fashion, one could say that the ratio was modified by sex.

For purely algebraic reasons, two measures may be modified in very different ways by the same variable. Furthermore, if both X and Z affect Y, absence of modification of the difference implies modification of the ratio, and vice-versa. For example, suppose for the subjects under study $Y_1 = 20$ and $Y_0 = 10$ for all the males, but $Y_1 = 30$ and $Y_0 = 15$ for all the females. Then $Y_1 - Y_0 = 10$ for males but $Y_1 - Y_0 = 15$ for females, so there is 5-year modification of the difference measure by sex. But suppose we measured the effects by expectancy ratios Y_1/Y_0, instead of differences. Then $Y_1/Y_0 = 20/10 = 2$ for males and $Y_1/Y_0 = 30/15 = 2$ for females as well, so there is no modification of the ratio measure by sex.

Consider next an example in which $Y_1 = 20$ and $Y_0 = 10$ for males, and $Y_1 = 30$ and $Y_0 = 20$ for females. Then $Y_1 - Y_0 = 10$ for both males and females, so there is no modification of the difference by sex. But $Y_1/Y_0 = 20/10 = 2$ for males and $Y_1/Y_0 = 30/20 = 1.5$ for females, so there is modification of the ratio by sex.

Finally, suppose $Y_1 = 20$ and $Y_0 = 10$ for males, and $Y_1 = 60$ and $Y_0 = 40$ for females. Then $Y_1 - Y_0 = 10$ for males and $Y_1 - Y_0 = 20$ for females, so the Y-difference is smaller among males than among females. But $Y_1/Y_0 = 20/10 = 2$ for males and $Y_1/Y_0 = 30/20 = 1.5$ for females, so the Y-ratio is larger among males than among females. Thus, modification can be in the opposite direction for different measures of effect.

Biological Interaction

The preceding examples show that one should not in general equate the presence or absence of effect-measure modification to the presence or absence of interactions in the biological (mechanistic) sense, because effect-measure modification depends entirely on what measure one chooses to examine, whereas the mechanism is the same regardless of that choice. Nonetheless, it is possible to formulate mechanisms of action that imply homogeneity (no modification) of a particular measure. For such a mechanism, the observation of heterogeneity in that measure can be taken as evidence against the mechanism (assuming of course that the observations are valid). It would be fallacious however to infer the mechanism is correct if homogeneity was observed, because many other mechanisms (some unimagined) would imply the observation.

A classic example is the simple "independent action" model for the effect of X and Z on Y, in which subjects affected by changes in X are disjoint from subjects affected by changes in Z (Greenland and Poole 1988; Weinberg 1986). This model implies homogeneity (*absence* of modification by Z) of the average X effect on Y when that effect is measured by the difference in the average Y. In particular, suppose Y is a disease indicator (1 if disease occurs, 0 if not). Then the average of Y is the proportion getting disease (the incidence proportion, often called the risk) and the average Y difference is the risk difference. Thus, in this context, the independent action model implies that the risk difference for the effect of X on Y will be constant across levels of Z; in other words, the risk difference will be homogeneous across Z, or unmodified by Z.

If both X and Z have effects, this homogeneity of the difference forces ratio measures of the effect of X on Y

to be heterogeneous across Z. When additional factors are present in the model (such as confounders) homogeneity of the risk differences can also lead to heterogeneity of the excess risk ratios (Greenland 1993a). Thus, under the simple independent action model, the independence of the X and Z effects will cause the measures other than the risk difference to be heterogeneous, or modified, across Z.

Biological models for the mechanism of X and Z interactions can lead to other patterns. For example, certain multistage models in which X and Z act at completely separate stages of a multistage mechanism can lead to homogeneity of ratios rather than differences, as well as particular dose–response patterns. Special caution is needed when interpreting observed patterns, however, because converse relations do not hold: Different plausible biological models may imply identical patterns of the effect measures (Moolgavkar 1986).

Synergism and Antagonism

Taking the simple independent-action model as a baseline, one may offer the following *dependent-action* definitions for an outcome indicator Y as a function of the causal antecedents X and Z. *Synergism* of $X = 1$ and $Z = 1$ in causing $Y = 1$ is defined as necessity and sufficiency of $X = 1$ and $Z = 1$ for causing $Y = 1$, i.e., $Y = 1$ if and only if $X = 1$ and $Z = 1$. We also may say that $Y = 1$ in a given individual would be a synergistic response to $X = 1$ and $Z = 1$ if $Y = 0$ would have occurred instead if either $X = 0$ or $Z = 0$ or both. In potential-outcome notation where Y_{xz} is an individual's outcome when $X = x$ and $Z = z$, this definition says synergistic responders have $Y_{11} = 1$ and $Y_{10} = Y_{01} = Y_{00} = 0$. *Antagonism* of $X = 1$ by $Z = 1$ in causing $Y = 1$ is defined as necessity and sufficiency of $Z = 0$ in order for $X = 1$ to cause $Y = 1$. This definition says antagonistic responders to X or Z have $Y_{10} = 1$ or $Y_{01} = 1$ or both, and $Y_{11} = Y_{00} = 0$.

With these definitions, synergism and antagonism are not logically distinct concepts, but depend on the coding of X and Z. For example, switching the labels of "exposed" and "unexposed" for one factor can change apparent synergy to apparent antagonism, and vice-versa (Greenland et al. 2008; Greenland and Poole 1988). The only label-invariant property is whether the effect of X on a given person is altered by the level of Z, i.e., the action of X is dependent on Z. If so, by definition we have biological interaction.

Absence of any synergistic or antagonistic interaction among levels of X and Z implies homogeneity (*absence* of modification by Z) of the average X effect across levels of Z when the X effect is measured by the differences in Y across levels of X (Greenland and Poole 1988; Greenland 1993b). The converse is false, however: Homogeneity of the difference measures (e.g., lack of modification of the risk difference) does not imply absence of synergy or antagonism, because such homogeneity can arise through other means (e.g., averaging out of the synergistic and antagonistic effects across the population being examined) (Greenland et al. 2008; Greenland and Poole 1988).

A more restrictive set of definitions is based on the sufficient-component cause model of causation (Greenland et al. 2008; Rothman 1976; VanderWeele and Robins 2007). Here, for two binary indicators X and Z, *synergism* of the effects of $X = 1$ and $Z = 1$ is defined as the presence of $X = 1$ and $Z = 1$ in the same sufficient cause of $Y = 1$, i.e., the sufficient cause cannot act without both $X = 1$ and $Z = 1$. Similarly, *antagonism* of the $X = 1$ effect by $Z = 1$ is defined as the presence of $X = 1$ and $Z = 0$ in the same sufficient cause of $Y = 1$. These definitions are also coding dependent.

Extensions to Continuous Outcomes

The use of indicators in the above definitions may appear restrictive but is not. For example, to subsume a continuous outcome T such as death time, we may define Y_t as the indicator for $T \leq t$ and apply the above definitions to each Y_t. Similar devices can be applied to incorporate continuous exposure variables (Greenland 1993b). The resulting set of indicators is of course unwieldy, and in application has to be simplified by modeling constraints (e.g., proportional hazards for T).

About the Author

For biography *see* the entry ►Confounding and Confounder Control.

Cross References

►Interaction

References and Further Reading

Greenland S (1993a) Additive-risk versus additive relative-risk models. Epidemiology 4:32–36

Greenland S (1993b) Basic problems in interaction assessment. Environ Health Perspect 101(Suppl 4):59–66

Greenland S, Poole C (1988) Invariants and noninvariants in the concept of interdependent effects. Scand J Work Environ Health 14:125–129

Greenland S, Lash TL, Rothman KJ (2008) Concepts of interaction. In: Modern epidemiology (Chapter 5), 3rd edn. Lippincott, Philadelphia, pp 71–83

Moolgavkar S (1986) Carcinogenesis modeling: from molecular biology to epidemiology. Annu Rev Public Health 7:151–169

Rothman KJ (1976) Causes. Am J Epidemiol 104:587–592

VanderWeele TJ, Robins JM (2007) The identification of synergism in the sufficient-component cause framework. Epidemiology 18:329–339

Weinberg CR (1986) Applicability of simple independent-action model to epidemiologic studies involving two factors and a dichotomous outcome. Am J Epidemiol 123:162–173

Effect Size

Shlomo Sawilowsky[1], Jack Sawilowsky[1], Robert J. Grissom[2]
[1]Wayne State University, Detroit, MI, USA
[2]San Francisco State University, San Francisco, CA, USA

In general, an effect size (ES) measures the extent parameters differ or variables are related. Effect-size methodology is barely out of its infancy and yet the effect size has already been proclaimed as the statistical coin of the realm for the 21st century. It was primarily due to the diligence of Cohen (1962, 1969, 1977, 1988) that the role and importance of the ES has attained great prominence in ▶power analysis, statistical analysis, and product and program evaluation.

The raw score difference may be a naïve ES measure in the context of a two independent samples layout consisting of a treatment and control group when the intervention is modeled as shift in location. It cannot always be prudently used to assess the magnitude of an intervention, or for that matter even a naturally occurring phenomenon, if arbitrary measures of the dependent variable (e.g., *Minnesota Multiphasic Personality Inventory subscale*, *Tennessee Self Concept Scale*) are used as a stand-in for latent variables, as opposed to inherently meaningful dependent variables (e.g., weight in pounds or kilograms, length of hospital stay).

The two primary parametric ES families are (1) d, a standardized difference between group's means, and (2) r^2, a measure of the proportion of variance in a variable that is attributable to the variability of another variable. (The Pearson correlation, r, is an ES measure of bivariate association.) Kirk (1996) listed 40 effect size indices among the d and r families, and the list has grown since then.

Cohen's (1969) d estimator of population d is the difference between two sample means divided by the pooled estimate of the assumed common population standard deviation:

$$d = \frac{\overline{X}_T - \overline{X}_C}{s_{Pooled}}$$

where $\overline{X}_T$ = mean of the treatment group, $\overline{X}_C$ = mean of the control group, and s_{Pooled} is the pooled standard deviation.

In the case of two samples of equal size the point-biserial correlation (r_{PB}) can be approximately obtained from d:

$$r_{PB_{X_1X_2}} = \frac{d}{\sqrt{d^2 + 4}},$$

where X_1 is a truly dichotomous independent variable (e.g., treatment group membership or gender) and X_2 is a continuous dependent variable.

Although his warning to be flexible about such very general designations in various areas of research is widely ignored, Cohen provided rule of thumb descriptors of the magnitude of d: $d \leq 0.2$ = small, $d = 0.5$ = medium, and $d \geq 0.8$ = large. Sawilowsky (1985) defined $d \geq 1.2$ = very large.

Glass' Δ (Glass et al. 1981) and Hedges' g (Hedges and Olkin, 1985) are indices that are related to d. They are obtained by substituting one group's standard deviation for the estimate based on the pooled variance to counter heterogeneity of variance (the ΔES), or substituting each $(n_i - 1)$ for each n_i, in the denominator when pooling variances for a better estimate of the assumed common population variance (the g ES). When sample sizes reach $n_1 = n_2 \approx 20$ these three indices tend to converge at the second decimal. A benefit of standardizing the mean difference via any of these three methods is it permits a ▶meta-analysis of ESs from different studies that used the same two levels of an independent variable, but different measures of the same latent dependent variable.

In terms of association, r^2 is called the sample coefficient of determination. (R^2, preferred by some textbook authors, refers to a hypothetical population parameter; see King 1986.) A related statistic, r^2_{adj}, is used in multiple regression to adjust for the number of variables in the model. Some caution is necessary in its use because squaring r, which can be positive or negative, yields a directionless r^2, and limitations inherent in r apply to r^2 (e.g., no causality is implied). To ensure correct interpretation of r^2, r (with its sign) should also be reported.

An approximate value of r^2 can be obtained from t:

$$r^2 = \frac{t^2}{t^2 + df}, \quad (1)$$

where t = obtained value of the Student t statistic and df = degrees of freedom. In terms of the two variable general linear model,

$$r^2 = \hat{\eta}^2 = \frac{SS_{Between}}{SS_{Total}},$$

where $SS_{Between}$ = sum of squares between groups and SS_{Total} = total sum of squares. ($\hat{\eta}^2$ is a sample estimate of the population η^2, the proportion of explained variance.) The $\hat{\eta}^2$ ES and its bias-adjusted counterpart, ω^2, can be extended to one-way and factorial ANOVA and MANOVA. According to Cohen, a general rule of thumb is $r^2 \leq 0.09$ = small, $0.09 < r^2 \leq 0.25$ = medium, and $r^2 > 0.25$ = large. Alternately, for r^2_{PB} or $\hat{\eta}^2$, 0.01 = small, 0.06 = medium, and 0.14 = large.

Attacks against hypothesis testing in the last quarter of the 20th century were accelerated with the emergence of the ES. For a response, see Sawilowsky (2003a). However, considerable controversy did arise regarding emerging ES reporting requirements in scholarly journals. For example, Thompson (1996:29, 1999:67) recommended the ES "can and should be reported and interpreted in all studies, regardless of whether or not statistical tests are reported" (1996), and "even [for] nonstatistically significant effects" (1999). This advice was initially, albeit lukewarmly, endorsed by the American Psychological Association (Wilkinson and APA 1999; APA 2001).

An invited debate on this topic was published in the *Journal of Modern Applied Statistical Methods* in 2003 (see, e.g., Sawilowsky 2003b, c) where it was argued that the ES should only be reported in the presence of a statistically significant hypothesis test. That is, the evidence against a null hypothesis must first be shown to be statistically significant before it makes sense to quantify the ES. Indeed, the typical null hypothesis implies that an ES = 0.

Subsequently, the APA (2010) recanted its support for publishing effect sizes in the absence of a statistically significant test result by at least removing active language supporting Thompson's recommendation. Thus, the words of Cohen prevailed: "The null hypothesis always means the effect size is zero...[but] when the null hypothesis is false, it is false to some specific degree, i.e., the effect size (ES) is some specific nonzero value in the population" (Cohen 1988:10). (In more current usage "nil hypothesis" would substitute for "null hypothesis".)

Nevertheless, this controversy persists. A primary reason is due to a misunderstanding attributable to r being commonly interpreted via r^2, even though no statistical test has been conducted. For example, suppose $r_{X_1X_2} = 0.2$. Then, $r^2 = 0.04$, indicating 4% of the variability in the X_1 scores are attributable to the X_2 variable. Similarly, if $r_{X_1X_3} = 0.8$, then $r^2 = 0.64$, or 64%. Using r^2 as the ES measure of the sample data, the X_3 variable is shown to be sixteen $\left(\frac{0.64}{0.04} = 16\right)$ times more effective in attributing variance in the X_1 scores than the X_2 variable, even though no statistical test has been conducted.

Rearranging (1), for $r \neq \pm 1$, Student's t is expressed in terms of r as:

$$t = \frac{r_{X_1X_2}}{\sqrt{(1 - r_{X_1X_2})(1 + r_{X_1X_2})/df}}.$$

Student's t can also be expressed in terms of r_{PB} as:

$$t = r_{PB_{X_1X_2}} \sqrt{\frac{N-2}{(1 - r_{PB_{X_1X_2}})(1 + r_{PB_{X_1X_2}})}},$$

where N = number of scores, and $r_{PB_{X_1X_2}} \neq \pm 1$. In either case it is erroneously believed that inherent in calculating a Pearson correlation for descriptive purposes a hypothesis test result, Student's t, has likewise been obtained. This is not correct, because r (or r_{PB}) is only a measure of co-relationship, whereas t determines the statistical significance of that correlation. The explained amount of variance in the dependent variable, r^2, is only a function of the size of r, whereas t is obtained based on the size of r as well as the size of the df. "Although t and r are related, they are measuring different things" (Gravetter and Wailnau 2005:431). Hence, knowing only the magnitude of r is insufficient to inform statistical significance.

The converse is also true. Conducting a Student's t test informs statistical significance, but does not convey the magnitude of an ES. Also, the result of an experiment might be statistically significant, but the ES might be too small to attain practical significance (e.g., no or insufficient clinical benefit).

Additional caution is in order in interpreting the magnitude of an ES. In terms of the explained proportion of variance ES, consider the Abelson (1985) paradox. The r^2 associated with batting average and the outcome of American baseball games was found to be 0.003, not quite $\frac{1}{3}$ of 1%, although it is universally accepted that batting average is the single most important characteristic of a successful baseball player. Therefore, heed Cohen's (1988:535) warning, "The next time you read that 'only X% of the variance is accounted for' remember Abelson's Paradox."

Similarly, in terms of the strength of an effect in an ANOVA layout, consider the Sawilowsky (2003a) paradox. The 1887 Michelson-Morley interferometer experiment was designed to detect the luminiferous ether. An effect of 30 kilometers/second (km/s) was hypothesized, but obtained results were only 5–7.5 km/s, with a paltry ES of $\hat{\eta}^2 = 0.005$. What is commonly referred to as the most famous *null* result in physics was not 0 km/s as might be imagined. Instead, it was actually more than 16,750

miles/h, a speed that exceeds Earth's satellite orbital velocity! The next time you read *Eta* squared was only "$\frac{1}{X}$ of 1%" remember Sawilowsky's Paradox.

Effect sizes developed to accompany parametric hypothesis tests have parametric assumptions (e.g., homoscedasticity, normality), which must not be overlooked. There have been advances in developing nonparametric ESs to accompany nonparametric tests. For example, see Hedges and Olkin (1984, 1985), Kraemer and Andrews (1982), and Newcombe (2006). Emphasizing confidence intervals for ESs, Grissom and Kim (2005) discussed traditional and robust parametric ESs for contingency tables with nominal or ordinal outcome categories (phi, risk difference, relative risk, odds ratio, cumulative odds ratio, and generalized odds ratio), and the nonparametric stochastic superiority ES for continuous or ordinal data: $\Pr(Y_1 > Y_2)$, where Y_1 and Y_2 are randomly sampled scores from two different populations.

For further reading see Algina et al. (2006), Baugh (2002), Bird (2002), Cohen (1973, 1990, 1992, 1994), Cooper and Findley (1982), Cortina and Nouri (2000), Dwyer (1974), Fern and Monroe (1996), Fleiss (1994), Fowler (1987), Grissom and Kim (2001), Hedges (1982), Hedges and Olkin (1985), Huberty (2002), Hunter and Schmidt (1990), Keselman (1975), Levin and Robinson (2003), Murray and Myors (2003), O'Grady (1982), Olejnik and Algina (2000, 2003), Ozer (1985), Parker (1995), Preece (1983), Prentice and Miller (1992), Richardson (1996), Ronis (1981), Rosenthal and Rubin (1982, 1994), Rosenthal et al. (2000), Rosnow and Rosenthal (2008), Sink and Stroh (2006), Vaughan and Corballis (1969), Volker (2006), and Wilcox and Muska (1999).

About the Authors

Biography of Shlomo Sawilowsky is in ►Frequentist hypothesis testing: A Defence.

Jack Sawilowsky, M. Ed., is a doctoral student in Evaluation and Research at Wayne State University. He is an Editorial Assistant for the *Journal of Modern Applied Statistical Methods*.

Dr. Robert J. Grissom is Professor Emeritus and Adjunct Professor, Department of Psychology, San Francisco State University, San Francisco, California, retiring early from teaching to focus on research and writing on effect sizes. He was a cofounder and Coordinator of the Graduate Program in Psychological Research. He received his doctorate in Experimental Psychology from Princeton University. He has authored and co-authored articles in edited books, Science magazine, and in methodological journals, where he is a regular reviewer on effect sizes. He is the senior author, with Dr. John J. Kim, of *Effect Sizes for Research: A Broad Practical Approach* (2005), published by Erlbaum; second edition scheduled for publication by Routledge in 2012.

Cross References

►Power Analysis
►Psychology, Statistics in
►Significance Tests: A Critique
►Statistical Fallacies: Misconceptions, and Myths
►Statistics: Controversies in Practice

References and Further Reading

Abelson RP (1985) A variance explanation paradox: When a little is a lot. Psychol Bull 97:129–133

Algina J, Keselman HJ, Penfield RD (2006) Confidence interval coverage for Cohen's effect size statistic. Educ Psychol Meas 66(6):945–960

American Psychological Association (2001) Publication manual, 5th edn. American Psychological Association, Washington, DC

American Psychological Association (2010) Publication manual, 6th edn. American Psychological Association, Washington, DC

Baugh F (2002) Correcting effect sizes for score reliability: a reminder that measurement and substantive issues are linked inextricably. Educ Psychol Meas 62:254–263

Bird KD (2002) Confidence intervals for effect sizes in analysis of variance. Educ Psychol Meas 62:197–226

Cohen J (1962) The statistical power of abnormal-social psychological research: a review. J Abnorm Soc Psychol 65:145–153

Cohen J (1969) Statistical power analysis for the behavioral sciences. Academic, San Diego

Cohen J (1973) Eta-squared and partial eta-squared in fixed factor ANOVA designs. Educ Psychol Meas 33:107–112

Cohen J (1977) Statistical power analysis for the behavioral sciences, rev. edn. Academic, San Diego

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Erlbaum, Hillsdale

Cohen J (1990) Things I have learned (so far). Am Psychol 45:1304–1312

Cohen J (1992) A power primer. Psychol Bull 112:155–159

Cohen J (1994) The earth is round ($p < .05$). Am Psychol 49:997–1003

Cooper HM, Findley M (1982) Expect effect sizes: estimates for statistical power analysis in social psychology. Personal Soc Psychol Bull 8:158–173

Cortina JM, Nouri H (2002) Effect sizes for ANOVA designs. Sage, Thousand Oaks

Dwyer JH (1974) Analysis of variance and the magnitude of effects: a general approach. Psychol Bull 81:731–737

Fern FE, Monroe KB (1996) Effect-size estimates: issues and problems in interpretation. J Consum Res 23:89–105

Fleiss JL (1994) Measures of effect size for categorical data. In: Cooper H, Hedges LV (eds) The handbook of research synthesis. Russell Sage, New York, pp 245–260

Fowler RL (1987) A general method for comparing effect magnitudes in ANOVA designs. Educ Psychol Meas 47:361–367

Glass GV, McGaw B, Smith ML (1981) Meta-analysis in social research. Sage, Thousand Oaks

Gravetter FJ, Wallnau LB (2005) Essentials of statistics for the behavioral sciences. 5th edn. Wadsworth/Thompson Learning, Belmont
Grissom RJ, Kim JJ (2001) Review of assumptions and problems in the appropriate conceptualization of effect size. Psychol Methods 6:135–146
Grissom RJ, Kim JJ (2005) Effect sizes for research: A broad practical approach. Erlbaum, Mahwah
Hedges L (1982) Estimation of effect size from a series of independent experiments. Psychol Bull 92:490–499
Hedges L, Olkin I (1984) Nonparametric estimators of effect size in meta-analysis. Psychol Bull 96:573–580
Hedges L, Olkin I (1985) Statistical methods for meta-analysis. Academic, New York
Huberty CJ (2002) A history of effect size indices. Educ Psychol Meas 62:227–240
Hunter JE, Schmidt FL (1990) Methods of meta-analysis: correcting error and bias in research findings. Sage, Newbury Park
Keselman H (1975) A Monte Carlo investigation of three estimates of treatment magnitude: epsilon squared, eta squared, and omega squared. Can Psychol Rev 16:44–48
King G (1986) How not to lie with statistics: avoiding common mistakes in quantitative political science. Am J Polit Sci 30(3): 666–687
Kirk RE (1996) Practical significance: a concept whose time has come. Educ Psychol Meas 56:746–759
Kraemer HC, Andrews G (1982) A non-parametric technique for meta-analysis effect size calculation. Psychol Bull 91: 404–412
Levin JR, Robinson DH (2003) The trouble with interpreting statistically nonsignificant effect sizes in single-study investigations. J Mod Appl Stat Methods 2:231–236
Murray LW, Myors B (2003) How significant is a significant difference? Problems with the measurement of magnitude of effect. J Couns Psychol 34:68–72
Newcombe RG (2006) Confidence intervals for an effect size measure based on the Mann–Whitney statistic. Stat Med 25(4): 559–573
O'Grady KE (1982) Measures of explained variance: cautions and limitations. Psychol Bull 92:766–777
Olejnik S, Algina J (2000) Measures of effect size for comparative studies: application, interpretations, and limitations. Contemp Educ Psychol 25:241–286
Olejnik S, Algina J (2003) Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol Methods 8:434–447
Ozer DJ (1985) Correlation and the coefficient of determination. Psychol Bull 97:307–315
Parker S (1995) The "difference of means" may not be the "effect size". Am Psychol 50:1101–1102
Preece PFW (1983) A measure of experimental effect size based on success rates. Educ Psychol Meas 43:763–766
Prentice DA, Miller DT (1992) When small effects are impressive. Psychol Bull 112:160–164
Richardson JTE (1996) Measures of effect size. Behav Res Methods: Instrum Comput 28:12–22
Ronis DL (1981) Comparing the magnitude of effects in ANOVA designs. Educ Psychol Meas 41:993–1000
Rosenthal R, Rubin DB (1982) A simple, general purpose display of magnitude of experimental effect. J Educ Psychol 76: 166–169
Rosenthal R, Rubin DB (1994) The counternull value of an effect size: a new statistic. Psychol Sci 5:329–334
Rosnow RL, Rosenthal R (2008) Assessing the effect size of outcome research. In: Nezu AM, Nezu CM (eds) Evidence-based outcome research. Oxford University Press, Oxford, pp 379–404
Rosenthal R, Rosnow RL, Rubin DB (2000) Contrasts and effect sizes in behavioral research: a correlational approach. Cambridge University Press, United Kingdom
Sawilowsky S (1985) Robust and power analysis of the 2 × 2 × 2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests. Unpublished doctoral dissertation, University of South Florida
Sawilowsky S (2003a) Deconstructing arguments from the case against hypothesis testing. J Mod Appl Stat Methods 2(2): 467–474
Sawilowsky S (2003b) You think you've got trivials? J Mod Appl Stat Methods 2(1):218–225
Sawilowsky S (2003c) Trivials: the birth, sale, and final production of meta-analysis. J Mod Appl Stat Methods 2(1):242–246
Sink CA, Stroh HR (2006) Practical significance: the use of effect sizes in school counseling research. Prof Sch Couns 9(5):401–411
Thompson B (1996) AERA editorial policies regarding statistical significance testing: three suggested reforms. Educ Res 25:26–30
Thompson B (1999) Five methodology errors in educational research: a pantheon of statistical significance and other faux pas. In: Thompson B (ed) Advances in social science methodology, vol 5. Sage, Thousand Oaks, CA, pp 23–86
Vaughan GM, Corballis MC (1969) Beyond tests of significance: estimating strength of effects in selected ANOVA designs. Psychol Bull 72:204–213
Volker MA (2006) Reporting effect size estimates in school psychology research. Psychol Sch 43(6):653–672
Wilcox RR, Muska J (1999) Measuring effect size: a non-parametric analog of ω^2. Br J Math Stat Psychol 52:93–110
Wilkinson L, the American Psychological Association (1999) Task force on statistical inference. Statistical methods in psychology journals: guidelines and explanations. Am Psychol 54: 594–604

Eigenvalue, Eigenvector and Eigenspace

Pui Lam Leung
Associate Professor, Director
The Chinese University of Hong Kong, Hong Kong, China

Let A be a $n \times n$ matrix and λ be a scalar. The *characteristic equation* of A is defined as

$$det(A - \lambda I_n) = 0, \tag{1}$$

where I_n is an identity matrix of order n and *det* denotes the determinant. The values of λ that satisfy Eq. (1) are called the *characteristic roots* or *eigenvalues* of A. In general, there are n eigenvalues of A. If λ_i is an eigenvalue, then the matrix

$(A - \lambda_i I_n)$ is singular and there exist a non-zero vector v_i such that $(A - \lambda_i I_n)v_i = 0$. Equivalently,

$$Av_i = \lambda_i v_i. \tag{2}$$

v_i is called the eigenvector corresponds to the eigenvalue λ_i. Geometrically, A represents a linear transformation in n-dimensional space. v_i is the direction remains unchanged or *invariant* under this transformation. In general, there are exactly n eigenvalues of A. These eigenvalues are not necessarily distinct and may be real or complex. Note that if v_i is the eigenvector corresponds to λ_i, then any scalar multiple of v_i is also an eigenvector of A. There may be more than one eigenvector correspond to the same eigenvalue λ_i, any linear combination of these eigenvectors will also be an eigenvector. The linear subspace spanned by the set of eigenvectors corresponds to the same eigenvalue together with the zero vector is called an *eigenspace*.

In statistics, we often deal with real symmetric matrices. Many test statistics in *Multivariate linear model* are functions of eigenvalues and eigenvectors of certain matrices. Some basic results concerning eigenvalues and eigenvectors are listed as follow:

1. If A is a real symmetric matrix then its eigenvalues are all real.
2. If A is a $n \times n$ matrix, then the eigenvalues of A and its transpose A' are the same.
3. If A and B are $n \times n$ matrices and A is non-singular, then the eigenvalues of AB and BA are equal.
4. If λ is an eigenvalue of A, then λ^k is an eigenvalue of A^k, where k is any positive integer.
5. If λ is an eigenvalue of a non-singular matrix A, then λ^{-1} is an eigenvalue of A^{-1}.
6. If A is an orthogonal matrix, (i.e., $AA' = I$), then all its eigenvalues have absolute value 1.
7. If A is symmetric, then A is idempotent (i.e., $A^2 = A$) if and only if its eigenvalues take values 0s and 1s only.
8. If A is a positive (non-negative) definite matrix, then all its eigenvalues are positive (non-negative).
9. If A is a real symmetric matrix and λ_i and λ_j are two distinct eigenvalues of A, then the two corresponding vectors v_i and v_j are orthogonal.
10. Let A be a $n \times n$ real symmetric matrix and $H = (v_1, \ldots, v_n)$ be a $n \times n$ matrix whose ith column is the eigenvectors v_i of A corresponds to the eigenvalues λ_i, then

$$H'AH = D = diag(\lambda_1, \ldots, \lambda_n), \tag{3}$$

where D is a diagonal matrix whose diagonal elements are λ_i. Furthermore, Eq. (3) can be rewritten as

$$A = HDH' = \lambda_1 v_1 v_1' + \cdots + \lambda_n v_n v_n'. \tag{4}$$

Equation (3) is known as the *diagonalization* of A while Eq. (4) is known as the *spectral decomposition* of A.

Extrema of Quadratic Form

There is an important theorem which is very useful in *Multivariate analysis* concerning the minimum and maximum of quadratic form.

Theorem Let A be a $n \times n$ positive definite matrix has the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n > 0$ and the corresponding eigenvectors are $v_1, \ldots, v_n$ and c is a $n \times 1$ vector. Then

1. $\max_{c \neq 0} \frac{c'Ac}{c'c} = \lambda_1$ and the maximum is attained at $c = v_1$.
2. $\min_{c \neq 0} \frac{c'Ac}{c'c} = \lambda_n$ and the minimum is attained at $c = v_n$.
3. $\max_{c \perp v_1, \ldots, v_k} \frac{c'Ac}{c'c} = \lambda_{k+1}$ and the maximum is attained at $c = v_{k+1}$ for $k = 1, 2, \ldots, n-1$. ($c \perp v_1, \ldots, v_k$ means c is orthogonal to $v_1, \ldots, v_k$).

This theorem is important since it provides the theory behind ►*principal component analysis* and ►*canonical correlation analysis*. In principal component analysis, we are looking for a linear transformation $Y = \alpha' X$ of the original variables X such that the variance of Y is maximum. In canonical correlation analysis, we are looking for a pair of linear transformation $U = \alpha' X$ and $V = \beta' Y$ of the original variables X and Y such that the correlation between U and V is maximum.

Software

Standard mathematical and statistical packages, such as MATLAB and R, have built-in function to compute eigenvalues and eigenvectors. The eigenvalues are usually output in descending order and the corresponding eigenvectors are normalized to unit length.

Cross References

►Correspondence Analysis
►Multivariate Technique: Robustness
►Principal Component Analysis
►Random Matrix Theory

References and Further Reading

Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley, New York
Graybill FA (1983) Matrices with applications in statistics, 2nd edn. Wadsworth, Belmont
Muirhead RJ (1982) Aspects of multivariate statistical theory. Wiley, New York
Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, New York

Empirical Likelihood Approach to Inference from Sample Survey Data

J. N. K. Rao
Professor Emeritus and Distinguished Research Professor
Carleton University, Ottawa, ON, Canada

Introduction

Traditional sample survey theory largely focused on descriptive parameters, such as finite population means, totals and quantiles, from samples selected according to efficient probability sampling designs (subject to cost constraints). Associated inferences consist of point estimation, standard errors of estimators and large-sample confidence intervals based on normal approximations. Inferences, based on the known probability distribution induced by the sampling design with the population item values held fixed but unknown, are essentially non-parametric. Standard text books on sampling (e.g., Cochran 1977) provide excellent accounts of the traditional design-based approach. Attempts were also made to integrate sampling theory with mainstream statistical inference based on likelihood functions, but retaining the non-parametric set up. Here we provide a brief account of some methods for sample survey data, based on non-parametric likelihoods. In particular, we focus on the "empirical likelihood" approach which was first introduced in the context of survey sampling by Hartley and Rao (1968) under the name "scale-load approach." Twenty years later, Owen (1988) introduced it in the main stream statistical inference, under the name "empirical likelihood." He developed a unified theory for the standard case of a random sample of independent and identically distributed (IID) variables, and demonstrated several advantages of the approach. In particular, the shape and orientation of empirical likelihood (EL) confidence intervals are determined entirely by the data, and the intervals are range preserving and transformation respecting, unlike the intervals based on normal approximations. An extensive account of EL, including various extensions and applications, is given in the excellent book by Owen (2001).

Scale-Load Approach

Suppose that the finite population U consists of N units labeled $i = 1,\ldots,N$ with associated item values y_i. A sample of units, s, is selected from U with specified probability $p(s)$. Godambe (1966) obtained the non-parametric likelihood function based on the full sample data $\{(i,y_i), i \in s\}$ and showed that it is non-informative in the sense that all possible non-observed values $y_i, i \notin s$ lead to the same likelihood. This difficulty arises because of the distinctness of labels i associated with the sample data that makes the sample unique. To get around this difficulty, Hartley and Rao (1968) suggested data reduction by ignoring some aspects of the data to make the sample non-unique and in turn make the associated likelihood informative. The reduction of sample data is not unique and it depends on the situation at hand. A basic feature of the Hartley–Rao (HR) approach is a specific representation of the finite population, assuming that the possible values of the variable y is a finite set of scale points $y_1^*,\ldots,y_T^*$ for some finite T. The associated population frequencies or "scale loads" are denoted by $N_1^*,\ldots,N_T^*$ and the population parameters can be expressed in terms of the scale loads; for example, the population mean $\overline{Y}$ can be expressed as $\bar{Y} = \sum_t p_t^* y_t^*$ where $p_t^* = N_t^*/N$ and $\sum_t p_t^* = 1$. Under simple random sampling without replacement, the sample labels i may be suppressed from the sample data in the absence of information relating the label to the corresponding item value, and the resulting nonparametric likelihood based on the reduced data $\{y_i, i \in s\}$ is given by the multivariate hyper-geometric distribution that depends on the scale loads N_t^* and the associated sample scale loads n_t^*. Thus the likelihood function based on the reduced sample data is informative, unlike the likelihood based on the full sample data. If the sampling fraction n/N is negligible, then the hyper-geometric likelihood may be approximated by the multinomial likelihood with probabilities p_t^* and associated sample frequencies n_t^* so that the log-likelihood function is given by $l(p^*) = \sum n_t^* \log\left(p_t^*\right)$ where the sum is over the observed non-zero sample scale loads. The resulting maximum likelihood estimator (MLE) of the mean $\overline{Y}$ is the sample mean $\bar{y} = \sum \hat{p}_t^* y_t^* = n^{-1}\sum_{i\in s} y_i$, where $\hat{p}_t^* = n_t^*/n$. Hartley and Rao (1968) also considered ML estimation when the population mean $\overline{X}$ of an auxiliary variable x is known and showed that the MLE is asymptotically equal to the customary linear regression estimator of $\overline{Y}$. The Hartley–Rao (HR) approach readily extends to stratified simple random sampling by retaining the strata labels h to reflect known strata differences and then regarding each stratum as a separate population and applying the scale-load approach to each stratum separately. The traditional stratified mean is the MLE under this approach.

Empirical Likelihood: IID Case

In the case of IID observations $y_1,\ldots,y_n$ from some distribution $F(.)$, Owen (1988) obtained a non-parametric

(or empirical) likelihood function which puts masses $p_i = \Pr(y = y_i)$ at the sample points y_i and the log-EL function is given by $l(p) = \sum_i \log(p_i)$ which is equivalent to the HR scale-load log-likelihood function. Maximizing $l(p)$ subject to $p_i > 0$ and $\sum p_i = 1$ gives maximum empirical likelihood (MEL) estimators of p_i and the mean $\mu = E(y)$ as $\hat{p}_i = 1/n$ and $\hat{\mu} = \sum \hat{p}_i y_i$. Chen and Qin (1993) studied EL in the sampling context, assuming simple random sampling, and considered parameters of the form $\theta = N^{-1}\sum_{i\in U} g(y_i)$ for specified functions $g(.)$ and known auxiliary information of the form $N^{-1}\sum_{i\in U} w(x_i) = 0$. In the special case of $g(y) = y$ and $w(x) = x - \overline{X}$, their results are equivalent to those of HR for estimating the mean $\overline{Y}$. By letting $g(y_i)$ be the indicator function $I(y_i \le t)$, the MEL estimator of the population distribution function is obtained as $\hat{F}(t) = \sum_{i\in s} \hat{p}_i I(y_i \le t)$, where $\hat{p}_i$ is the maximum EL estimator of p_i obtained by maximizing the log-EL function subject to the above restrictions on the p_i and the additional restriction $\sum_{i\in s} p_i w(x_i) = 0$. The estimator $\hat{F}(t)$ is non-decreasing and hence can be used to obtain estimators of population quantiles and other functionals, in particular the population median.

The EL approach can provide non-parametric confidence intervals on parameters of interest, similar to the classical parametric likelihood ratio intervals. For example, the EL intervals on the mean μ for the IID case are obtained by first deriving the profile EL ratio function $R(\mu)$ and then treating $r(\mu) = -2\log R(\mu)$ as χ^2_1, a chi-squared variable with one degree of freedom, where $R(\mu)$ is the maximum of $\prod(np_i)$ subject to the previous restrictions on the p_i and the additional constraint $\sum p_i y_i = \mu$. The $(1-\alpha)$-level EL interval is given by $\{\mu | r(\mu) \le \chi^2_1(\alpha)\}$, where $\chi^2_1(\alpha)$ is the upper α-point of the distribution of χ^2_1. This result is based on the fact that $r(\mu)$ is approximately distributed as χ^2_1 in large samples. As noted earlier, the shape and orientation of EL intervals are determined entirely by the data, and the intervals are range preserving and transformation respecting. Moreover, unlike the traditional normal approximation intervals the EL intervals do not require the evaluation of standard errors of estimators in the IID case and are particularly useful if balanced tail error rates are considered desirable. An important application of EL intervals in audit sampling is reported by Chen et al. (2003). Populations containing many zero item values are encountered in audit sampling, where y_i denotes the amount of money owed to a government agency and $\overline{Y}$ is the average amount of excessive claims. In some previous work on audit sampling, parametric likelihood ratio intervals based on parametric mixture distributions for the variable y were used because they performed better than the normal approximation-based intervals in terms of coverage. On the other hand, the EL intervals compared favorably to the parametric intervals when the assumed parametric distribution holds, and performed better than the parametric intervals under deviations from the assumed parametric distribution, by providing non-coverage rate below the lower bound on $\overline{Y}$ closer to the nominal value and also larger lower bound. Typically, government agencies use the lower bound for the collection of excess amounts claimed.

Pseudo-EL: Complex Surveys

The HR scale load approach or the EL approach does not readily extend to general sampling design involving unequal probability sampling. To circumvent this difficulty, a pseudo-EL approach has been proposed (Chen and Sitter 1999) by regarding the finite population as a random sample from an infinite super-population. This in turn leads to the "census" log-EL function $l_N(p) = \sum_{i\in U} \log(p_i)$ which is simply the finite population total of the $\log(p_i)$. From standard sampling theory, an unbiased estimator of the finite population total $l_N(p)$ is given by $\hat{l}(p) = \sum_{i\in s} d_i \log(p_i)$, where $d_i = 1/\pi_i$ is the design weight associated with unit i and π_i is the associated probability of inclusion in the sample. The pseudo empirical log-likelihood function is defined as $\hat{l}(p)$ and then one can proceed as in section "►Scale-Load Approach" to get point estimators of finite population parameters. For example, in the absence of auxiliary information, the pseudo-MEL estimator of the finite population mean $\overline{Y}$ is given by the well-known ratio estimator $\left(\sum_{i\in s} d_i\right)^{-1}\left(\sum_{i\in s} d_i y_i\right)$. Wu (2004) provided algorithms and R codes for computing the pseudo-MEL estimators. Interval estimation, however, runs into difficulties under this approach because the profile pseudo-EL ratio function obtained from the pseudo-EL does not lead to asymptotic χ^2_1 distribution in large samples. To get around this difficulty, Wu and Rao (2006) proposed an alternative pseudo empirical log-likelihood function. It is given by $\hat{l}_{\mathrm{mod}}(p) = n^* \sum_{i\in s} \tilde{d}_i \log(p_i)$, where $\tilde{d}_i = d_i / \sum_{k\in s} d_k$ are the normalized design weights and n^* is the "effective sample size" defined as the ratio of the sample size n and the estimated design effect; the latter is taken as the ratio of the estimated variance of the point estimator under the specified design and under simple random sampling. This design effect depends on the parameter of interest. For simple random sampling with replacement, $\hat{l}_{\mathrm{mod}}(p)$ reduces to the usual empirical log-likelihood function.

The pseudo-MEL estimator under the modified function remains the same as the estimator obtained from the original pseudo-EL function. But the modification leads to a different pseudo-EL ratio function that leads to asymptotic χ^2_1 distribution and hence the resulting intervals can be used. Wu and Rao (2010) proposed a bootstrap calibration method that bypasses the need of effective sample size and the resulting pseudo-EL intervals are asymptotically valid and performed well in simulation studies.

In a recent application, the modified pseudo-EL method, based on the χ^2_1 approximation, was applied to adaptive cluster sampling which is often used to sample populations that are sparse but clustered (Salehi et al. 2010). Simulation results showed that the resulting confidence intervals perform well in terms of coverage whereas the traditional normal approximation-based intervals perform poorly in finite samples because the distribution of the estimator under ►adaptive sampling is highly skewed.

Various extensions of the EL method for survey data have been reported, including multiple frame surveys and imputation for missing data (see ►Imputation). We refer the reader to a recent review paper (Rao and Wu 2009) on empirical likelihood methods for survey data.

About the Author

Dr. J. N. K. Rao is Distinguished Research Professor at Carleton University, Ottawa, Canada and a long term Consultant to Statistics Canada and a Member of Statistics Canada's Advisory Committee on Methodology. "He published his first research work in sampling at the age of twenty, four years before he received his PhD in 1961. Very quickly Jon became one of the leading researchers in survey sampling and has remained one of the field's leaders to this day" (David Bellhouse, *J.N.K. Rao: An Appreciation of his work*, SSC Annual Meeting, June 2001, Proceedings of the Survey Methods Section, p. 3). Professor Rao received the 2004 prestigious Waksberg Award for survey methodology and the 1993 Gold Medal of Statistical Society of Canada in recognition of "fundamental research achievements, especially in the theory and practice of surveys." Professor Rao is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics. He was elected as a Fellow of the Royal Society of Canada for "fundamental contributions to survey sampling, from its foundational principles to complex survey design and analysis. These contributions have influenced the way surveys are conducted and analyzed." His 1981 paper with Alastair Scott, published in the *Journal of American Statistical Association*, was selected as one of 19 landmark papers in the history of survey sampling for the 2001 centenary volume of the International Association of Survey Statisticians. Currently, he is Associate editor, *Survey Methodology* (1985–). He received Honorary Doctor of Mathematics degree in 2008 from the University of Waterloo. Professor Rao is the author of the well known book *Small Area Estimation* (Wiley-Interscience, 2003).

Cross References

►Likelihood
►Sample Survey Methods

References and Further Reading

Chen J, Qin J (1993) Empirical likelihood estimation for finite population and the effective usage of auxiliary information. Biometrika 80:107–116
Chen J, Sitter RR (1999) A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. Stat Sin 9:385–406
Chen J, Chen SY, Rao JNK (2003) Empirical likelihood confidence intervals for the mean of a population containing many zero values. Can J Stat 31:53–68
Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York
Godambe VP (1966) A unified theory of sampling from finite populations. J R Stat Soc Ser B 28:310–328
Hartley HO, Rao JNK (1968) A new estimation theory for sample surveys. Biometrika 55:547–557
Owen AB (1988) Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75:237–249
Owen AB (2001) Empirical likelihood. Chapman and Hall, New York
Rao JNK, Wu C (2009) Empirical likelihood methods. In: Pfeffermann D, Rao CR (eds) Handbook of statistics – sample surveys: inference and analysis, vol 29B. North-Holland, The Netherlands, pp 189–207
Salehi MM, Mohammadi M, Rao JNK, Berger YG (2010) Empirical likelihood confidence intervals for adaptive cluster sampling. Environ Ecol Stat 17:111–123
Wu C (2004) Algorithms and R codes for the pseudo empirical likelihood methods in survey sampling. Surv Methodol 31:239–243
Wu C, Rao JNK (2006) Pseudo-empirical likelihood ratio confidence intervals. Can J Stat 34:359–375
Wu C, Rao JNK (2010) Bootstrap procedures for the pseudo empirical likelihood method in sample surveys. Stat Prob Letters 80:1472–1478

Empirical Processes

Evarist Giné
Professor
University of Connecticut, Storrs, CT, USA

A basic question in Statistics is how well does the frequency of an event approach its probability when the number of repetitions of an experiment with "statistical regularity"

increases indefinitely, or how well does the average of the values of a function at the observed outcomes approach its expected value. "How well" of course may have different meanings. Empirical process theory has its origin on this question. The counterpart in the "probability model" of averaging the values of a function f at the outcomes of n repetitions of an experiment, is just $P_n f = n^{-1}\sum_{i=1}^{n} f(X_i)$, where X_i, the data, are independent, identically distributed random variables with common probability law P. The variables X_i need not be real but may take values in any measurable space $(S, \mathcal{S})$. The random measure $P_n = n^{-1}\sum_{i=1}^{n}\delta_{X_i}$, where δ_x denotes unit mass at x (the Dirac delta) is the "empirical measure," and $P_n f = \int f dP_n$, assigns to each function its average over the values $X_i(\omega)$, and in particular to each set A the frequency of the occurrences $X_i(\omega) \in A$. This random measure is often understood as a stochastic process indexed by $f \in \mathcal{F}$, $\mathcal{F}$ being a collection of measurable functions (which can in particular be indicators). The object of empirical process theory is to study the properties of the approximation of Pf by $P_n f$, uniformly in $\mathcal{F}$, mainly, probabilistic limit theorems for the processes $\{(P_n - P)f : f \in \mathcal{F}\}$ and probability estimates of the random quantity

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |P_n f - Pf|.$$

This program has a long history. It starts with Bernoulli and de Moivre in the early 1700s: they studied the approximation of the probability of an event, PA, by its frequency, $P_n A$, thus obtaining respectively the Bernoulli law of large numbers and the de Moivre(-Laplace) normal approximation of the binomial law.

Next, Glivenko and Cantelli, in the 1920s showed, for X_i real valued, that $F_n(x) = P_n(-\infty, x]$ converges uniformly almost surely to the "true" distribution function $F(x) = P(-\infty, x]$, and this was the first "uniform" law of large numbers for the empirical measure. Kolmogorov in the 1930s gave the limiting distribution of $\sqrt{n}\sup_{x\in\mathbf{R}} |F_n(x) - F(x)|$, and Dvoretzky–Kiefer–Wolfowitz in the 1950s gave an inequality of the right order, that applies for all values of n, for $\Pr\left\{\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| > \varepsilon\right\}$, namely, this is less than or equal to $2e^{-2n\varepsilon^2}$ (where the constant 2 in front of the exponent was determined by Pascal Pascal Massart (1990), cannot be improved, and neither can the exponent).

The theorem of Kolmogorov is a corollary of Donsker's theorem, who viewed the empirical process as a random function and proved for it a central limit theorem (see ►Central Limit Theorems), that is, a limit theorem for its probability law defined on function space (in this case, the space of right continuous functions with left limits, $D(-\infty, \infty)$, with a separable metric on it, the Skorokhod metric, with the law of a Gaussian process (the P-Brownian bridge) as limit). Then, by the "continuous mapping theorem" $H(\sqrt{n}(F_n - F))$ converges in law to $H(G_F)$ for all functionals H continuous on D, and Kolmogorov's theorem follows by taking $H(f) = \sup_{x\in\mathbf{R}} |f(x)|$. [Actually, Donsker's theorem is a result of efforts by at least Kolmogorov, Skorokhod, Doob, Donsker and, in several dimensions, Dudley]. This was mostly done in the mid 1950s.

Another important area on what we may call the classical empirical process theory is that of the "strong approximation" in a suitable probability space of the empirical process for the uniform distribution (see ►Uniform Distribution in Statistics) on $[0, 1]$ by Brownian bridges, in the supremum norm ("Hungarian constructions"), very useful e.g., in the study of limit theorem of complicated functions of the empirical process (Komlos et al. 1975). Other results include laws of the iterated logarithm and large deviation principles (Sanov's theorem). The book to consult on the classical theory of empirical processes is Shorack and Wellner's (1986).

The Glivenko–Cantelli, Donsker, and Dvoretzky–Kiefer–Wolfowitz results for the empirical distribution function became the models for the striking generalizations that constitute modern empirical process theory. The first is the Vapnik–Červonenkis (V–C) (1971) law of large numbers. These authors gave very sharp combinatorial, non-random conditions on classes of sets $\mathcal{C}$ (classes of functions $\mathcal{F}$ in a subsequent article in 1981) for the empirical measure to converge to the probability law of the observations *uniformly* over the sets in $\mathcal{C}$ (or the functions in $\mathcal{F}$) almost surely, and necessary and sufficient random conditions on these classes for uniformity to take place. A little earlier, Blum and DeHardt had given other also very sharp general conditions (bracketing entropy) for the same result to hold over classes of functions in Euclidean space, see DeHardt (1971). Vapnik–Červonenkis also applied their work to pattern recognition (see ►Pattern Recognition, Aspects of and ►Statistical Pattern Recognition Principles).

The theory greatly gained in scope and applicability with the work of Dudley (1978), where Donsker's central limit theorem (also called invariance principle) was greatly generalized in the spirit of Vapnik–Červonenkis and Blum–deHardt by proving central limit theorems for empirical processes that hold uniformly in $f \in \mathcal{F}$ and by giving concrete meaningful examples of such classes, like classes of sets with differentiable boundaries, or positivity parts of finite dimensional sets of functions; measurability problems had to be overcome, and, more importantly, a relationship had to be provided between the V–C combinatorial quantities and metric entropy, which allowed proving asymptotic equicontinuity

conditions (hence, "uniform tightness") for the empirical process over large classes of functions in a way reminiscent of how Dudley proved his famous entropy bound for ►Gaussian processes.

The articles of Vapnik and Červonenkis and Dudley just mentioned originated the field of modern empirical processes. We now review some of its main developments. The most general form of the central limit theorem under bracketing conditions was obtained by Ossiander (1987). The connection between empirical processes and Gaussian and sub-Gaussian processes and empirical processes was made more explicit by means of random sign and Gaussian randomization (Pollard 1981; Koltchinskii 1986 and, more extensively, in Giné and Zinn (1984), where, moreover, the central limit theorem for classes of sets analogue to the V–C law of large numbers is proved; its conditions were proved later, in 1988, to be necessary by M. Talagrand, who continued some aspects of the work of Giné and Zinn and also gave a "structural" necessary and sufficient condition for the central limit theorem). Talagrand (1987) also has the last word on Glivenko–Cantelli type theorems.

The limit theory contains as well laws of the iterated logarithm, theorems on uniform and universal Donsker classes (very convenient for applications since in Statistics one does not know the law of the data; see Dudley 1987; Giné and Zinn 1991; Sheehy and Wellner 1992 and for the uniform in P Glvienko–Cantelli theorem, Dudley et al. 1991) and the bootstrap of empirical processes (Giné and Zinn 1990) among other types of results.

Particular mention should be made of exponential inequalities for empirical processes, which may be seen as Bernstein or Prokhorov exponential inequalities uniform over large collections of sums of independent random variables, indexed by functions. For the special but conspicuous Vapnik–Červonekis classes of functions, Alexander (1984) and Massart (1986) had versions of such inequalities in the mid 1980s, but Talagrand (1996) obtained a striking, completely general such inequality, just in terms of the supremum of $|f(X)|$ over $\mathcal{F}$ and of $E\sup_{f\in\mathcal{F}}\sum_{i=1}^{n} f^2(X_i)$. The constants in his inequality are not specific, but in a series of articles, P. Massart, T. Klein and E. Rio, and O. Bousquet obtained the best constants. This inequality has applications in many areas of Statistics and, in this author's view, is a true landmark.

Both, limit theorems and exponential inequalities have been extended to U-processes, not necessarily in the greatest generality. See Arcones and Giné (1993) and Major (2006).

Modern empirical process theory has already had and is having a deep impact in pattern recognition and learning theory, M-estimation, density estimation, and quite generally, asymptotic and non-parametric statistics. This will not be reviewed in this article.

The books to consult on modern empirical process theory are those of Pollard (1984), van der Vaart and Wellner (1996), Dudley (1999); for U-processes, de la Peña and Giné (1999); and for Talagrand's and other inequalities for empirical processes, Ledoux (2001).

About the Author

Dr. Evarist Giné is a Professor of Mathematics and (by courtesy) of Statistics at the University of Connecticut. He is an Elected member of the International Statistical Institute, a Fellow of the Institute of Mathematical Statistics and a corresponding member of the Institut d'Estudis Catalans. He has authored/coauthored more than 100 articles in Probability and Mathematical Statistics, as well as two books: *The Central Limit Theorem for Real and Banach Valued Random Variables* (with A. Araujo, Wiley, 1980), and *Decoupling, from Dependence to Independence* (with V. de la Peña, Springer, 1999). He has organized several meetings and special sessions in Probability and Mathematical Statistics, in particular he was the Program Chair of the 5th World Congress of the Bernoulli Society/IMS Annual Meeting. He has been Associate editor of *Annals of Probability*, *Stochastic Processes and Applications*, and *Revista Matemática Hispano-Americana*, and is presently Associate editor of *Bernoulli*, *Journal of Theoretical Probability*, *Electronic Journal and Electronic Communications in Probability*, *TEST* and *Publicacions Matemàtiques*.

Cross References

►Exact Goodness-of-Fit Tests Based on Sufficiency
►Khmaladze Transformation
►Strong Approximations in Probability and Statistics
►Testing Exponentiality of Distribution

References and Further Reading

Alexander K (1984) Probability inequalities for empirical processes and a law of the iterated logarithm. Ann Probab 12:1041–1067
Arcones A, Giné E (1993) Limit theorems for U-processes. Ann Probab 21:1494–1542
DeHardt J (1971) Generalizations of the Glivenko–Cantelli theorem. Ann Math Stat 42:2050–2055
de la Peña V, Giné E (1999) Decoupling, from dependence to independence. Springer, New York
Donsker MD (1952) Justification and extension of Doob's heuristic approach to the Kolmogorov–Smirnov theorems. Ann Math Stat 23:277–281
Dudley RM (1978) Central limit theorems for empirical measures. Ann Probab 6:899–929
Dudley RM (1987) Universal Donsker classes and metric entropy. Ann Probab 15:1306–1326
Dudley RM (1999) Uniform central limit theorems. Cambridge University Press, Cambridge
Dudley RM, Giné E, Zinn J (1991) Uniform and universal Glivenko–Cantelli classes. J Theoret Probab 4:485–510

Giné E, Zinn J (1984) Some limit theorems for empirical processes. Ann Probab 12:929–998
Giné E, Zinn J (1990) Bootstrapping general empirical measures. Ann Probab 18:851–869
Giné E, Zinn J (1991) Gaussian characterization of uniform Donsker classes of functions. Ann Probab 19:758–782
Koltchinskii VI (1986) Functional limit theorems and empirical entropy, I and II. Teor Veroyatn Mat Statist 33:31–42, bf 34 73–85
Komlós J, Major P, Tusnády G (1975) An approximation of partial sums of independent RV's and the sample DF. I. Z. Wahrsch Verw Gebiete 32:111–131
Ledoux M (2001) The concentration of measure phenomenon. American mathematical society, Providence, Rhode Island
Major P (2006) An estimate on the supremum of a nice class of stochastic integrals and U-statistics. Probab Theory Relat Fields 134:489–537
Massart P (1986) Rates of convergence in the central limit theorem for empirical processes. Ann Inst H Poincaré Probab Stat 22:381–423
Massart P (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. Ann Probab 18:1269–1283
Ossiander M (1987) A central limit theorem under metric entropy with L_2 bracketing. Ann Probab 15:897–919
Pollard D (1981) Limit theorem for empirical processes. Z Warsch Verb Gebiete 57:181–195
Pollard D (1984) Convergence of Stochastic Processes. Springer, New York
Sheehy A, Wellner JA (1982) Uniform Donsker classes of functions. Ann Probab 20:1983–2030
Shorack G, Wellner J (1986) Empirical Processes with Applications to Statistics. Wiley, New York
Talagrand M (1987a) The Glivenko–Cantelli problem. Ann Probab 15:837–870
Talagrand M (1987b) Donsker classes and random geometry. Ann Probab 15:1327–1338
Talagrand M (1988) Donsker classes of sets. Probab Theory Relat Fields 78:169–191
Talagrand M (1996) New concentration inequalities in product spaces. Invent Math 126:505–563
van der Vaart A, Wellner J (1996) Weak convergence and empirical processes: with applications to statistics. Springer, New York
Vapnik VN, Červonenkis AJa (1971) On the uniform convergence of relative frequencies of events to their probabilities. Theory Probab Appl 16:164–180
Vapnik VN, Červonenkis AJa (1981) Necessary and sufficient conditions for the convergence of means to their expectations. Theory Probab Appl 26:532–553

Entropy

Tarald O. Kvålseth
Professor Emeritus
University of Minnesota, Minneapolis, MN, USA

Introduction and Definition

Consider that $P_n = (p_1, \ldots, p_n)$, or $\{p_i\}$, is a finite discrete probability distribution with $p_i \geq 0$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} p_i = 1$. This distribution may be associated with a set of mutually exclusive and exhaustive events $\{E_1, \ldots, E_n\}$ or with a random variable X taking on a set of n discrete values with probabilities $p_i (i = 1, \ldots, n)$. For such a distribution, the classical definition of *entropy* is given by

$$H(P_n) = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

and generally referred to as *Shannon's entropy* after Shannon (Shannon 1948). Two different bases are typically used for the logarithm in (1). In the field of information theory, in which the entropy is of fundamental importance, the base -2 logarithm is the traditional choice, with the unit of measurement of entropy then being a *bit*. If the base $-e$ (natural logarithm) is used, which is mathematically more convenient to work with, the unit of measurement is sometimes called a *nat*. Also, $0 \log 0 = 0$ is the usual convention for (1).

Most generally, the entropy in (1) is considered to be a measure of the amount of *randomness* in a system or of a set of events or of a random variable. Alternatively, the entropy is considered to express the amount of *uncertainty, surprise*, or *information content* of a set of events or of a random variable. In information theory, the p_i's in (1) may be those of a set of messages (events) or those of a random variable X taking on the value i for the ith symbol of an alphabet for $i = 1, \ldots, n$. In thermodynamics and statistical mechanics, where entropy plays an important role, the p_i in (1) is the probability of a system being in the ith quantum state with $i = 1, \ldots, n$. Besides such fundamentally important applications, the entropy in (1) has proved to be remarkably versatile as a measure of a variety of attributes in different fields of study, ranging from psychology (e.g., Garner 1962) to ecology (e.g., Magurran 2004). The wide range of applications of the entropy formula is due in part to its desirable properties and extensions and partly perhaps because of the intrigue held by the entropy concept.

Properties of *H*

Some of the most important properties of the entropy function H in (1) for any $P_n = (p_1, \ldots, p_n)$ may be outlined as follows:

(**P1**) H is continuous in all its arguments $p_1, \ldots, p_n$.

(**P2**) H is (permutation) symmetric in the $p_i (i = 1, \ldots, n)$.

(**P3**) H is zero-indifferent (expansible), i.e., $H(p_1, \ldots, p_n, 0) = H(p_1, \ldots, p_n)$.

(**P4**) $H(0, \ldots, 0, 1, 0, \ldots, 0) = 0$ and $H(1/n, \ldots, 1/n)$ is strictly increasing in n.

(**P5**) H is strictly Schur-concave in P_n and also strictly concave (in the usual or Jensen sense).

These properties are all readily apparent from the definition in (1) and would seem to be reasonable for a measure of randomness. The strict Schur-concavity property ensures that $H(P_n)$ increases strictly as the p_i's become increasingly equal (even or uniform). Stated more precisely in terms of *majorization theory* (Marshall and Olkin 1979), if the probability distribution $P_n = (p_1, \ldots, p_n)$ is majorized by the distribution $Q_n = (q_1, \ldots, q_n)$, denoted as $P_n \prec Q_n$, then, because of the strict Schur-concavity of H, $H(P_n) \geq H(Q_n)$ with strict inequality unless P_n is a permutation of Q_n. From the majorization

$$P_n^1 = (1/n, \ldots, 1/n) \prec P_n = (p_1, \ldots, p_n) \prec P_n^0 = (0, \ldots, 0, 1, 0, \ldots, 0) \quad (2)$$

together with the strict Schur-concavity of H, it follows that

$$H\left(P_n^0\right) \leq H(P_n) \leq H\left(P_n^1\right) \quad (3)$$

with equalities if, and only if, $P_n = P_n^0$ or $P_n = P_n^1$ (Marshall and Olkin 1979: 7, 71, 410).

Another property of H is that of *additivity*. Thus, consider the case of two statistical experiments with their respective events, or, consider the pair of random variables (X, Y) with the joint probability distribution $\{p_{ij}\}$ and the marginal distributions $\{p_{i+}\}$ and $\{p_{+j}\}$ for X and Y, respectively, with $p_{i+} = \sum_{j=1}^{J} p_{ij}$ for all i and $p_{+j} = \sum_{i=1}^{I} p_{ij}$ for all j and with $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = \sum_{i=1}^{I} p_{i+} = \sum_{j=1}^{J} p_{+j} = 1$. Then, the following property follows from (1):

(**P6**) H is additive, i.e.,

$$H(\{p_{ij}\}) = H(\{p_{i+}p_{+j}\}) = H(\{p_{i+}\}) + H(\{p_{+j}\}) \quad (4)$$

under independence i.e., when the two experiments or when X and Y are independent.

For the case when $p_{i+} = 1/I$ and $p_{+j} = 1/J$ for all i and j, (4) reduces to

$$H\left(\left\{\frac{1}{IJ}\right\}\right) = H\left(\left\{\frac{1}{I}\right\}\right) + H\left(\left\{\frac{1}{J}\right\}\right) \quad (5)$$

which is sometimes referred to as the *weak additivity* property.

Two famous inequalities involving the entropy function H in (1) can be expressed as follows:

(**P7**) For any two probability distributions $P_n = (p_1, \ldots, p_n)$ and $Q_n = (q_1, \ldots, q_n)$,

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i \quad (6)$$

which is referred to as *Shannon's fundamental inequality* and is sometimes called the *Gibbs' inequality.*

(**P8**) As a consequence of (6),

$$H(\{p_{ij}\}) \leq H(\{p_{i+}\}) + H(\{p_{+j}\}) \quad (7)$$

which is called the *subadditivity* property of H and of which (4) is the only equality case, which occurs under independence when $p_{ij} = p_{i+}p_{+j}$ for all i and j.

The entropy in (1), as well as other entropy measures, may be derived by using some of these properties as axioms or postulates. Such *axiomatic characterization* approach starts off with a set of axioms specifying some properties that a measure should possess and then showing that a particular function satisfies those axioms (e.g., Aczél and Daróczy 1975; Mathai and Rathie 1975).

Cross-Entropy (Divergence)

The inequality in (6) can also be expressed as

$$D(P_n : Q_n) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right) \geq 0 \quad (8)$$

where $D(P_n : Q_n)$ is known as the *cross-entropy,* but is more commonly called the *divergence of P_n from Q_n*. Specifically, it is called the ▶Kullback–Leibler divergence or information (Kullback and Leibler 1951). Since $D(P_n : Q_n)$ is not symmetric in P_n and Q_n, the term *directed divergence* is sometimes used. It differs from the *conditional entropy* defined as $\sum_{i=1}^{I} p_{i+}\left[-\sum_{j=1}^{J}(p_{ij}/p_{i+})\log(p_{ij}/p_{i+})\right]$. To avoid mathematical difficulties with (8), the convention $0 \log 0 = 0$ is used together with the assumption that $q_i = 0$ whenever $p_i = 0$. From (8), $D(P_n : Q_n) = 0$ if, and only if, $P_n = Q_n$. It can also easily be seen that $D(P_n : Q_n)$ is strictly convex in P_n for any given Q_n and strictly convex in Q_n for any given P_n.

The $D(P_n : Q_n)$ in (8) and various other measures and generalizations of divergence (or "distance") are important in statistical inferences (see Pardo 2006 for a recent and extensive treatment of this topic). It is recognized from the form of the expression in (8) that it is related to the likelihood-ratio chi-square statistic G^2 for goodness-of-fit tests. In particular, if the p_i's are multinomial sample probabilities (proportions) $p_i = n_i/N$ for $i = 1, \ldots, n$ with sample size $N = \sum_{i=1}^{n} n_i$ and if $\Pi_{0n} = (\pi_{01}, \ldots, \pi_{0n})$ is the corresponding hypothesized distribution, with $m_i = N\pi_{0i}$ for $i = 1, \ldots, n$, then

$$G^2 = 2ND(P_n : \Pi_{0n}) = 2\sum_{i=1}^{n} n_i \log\left(\frac{n_i}{m_i}\right) \quad (9)$$

which has the asymptotic ▶chi-square distribution under the null hypothesis involving the Π_{0n}- distribution and with the proper degrees of freedom ($n-1$ if Π_{0n} involves no unknown parameters and $n-s-1$ if Π_{0n} involves s unknown parameters that are efficiently estimated).

One particular divergence (cross-entropy) is that from the joint distribution $\{p_{ij}\}$ for the random variables X and Y to their independence distribution $\{p_{i+}p_{+j}\}$, which, from (8) is given by

$$D(\{p_{ij}\}:\{p_{i+}p_{+j}\}) = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}\log\left(\frac{p_{ij}}{p_{i+}p_{+j}}\right) \qquad (10)$$

and is known as the *mutual information* between X and Y, but is also called *transinformation*. This measure reflects in a sense the statistical dependence between X and Y, with $D(\{p_{ij}\}:\{p_{i+}p_{+j}\}) = 0$ if, and only if, X and Y are independent. It can also clearly be expressed in terms of individual entropies from (1) as

$$D(\{p_{ij}\}:\{p_{i+}p_{+j}\}) = H(\{p_{i+}\}) + H(\{p_{+j}\}) - H(\{p_{ij}\}) \qquad (11)$$

which can also be seen to be non-negative from Property (P8) as expressed in (7).

In the case when X and Y are continuous random variables with the joint probability density function $h(x,y)$ and the marginal probability density functions $f(x)$ and $g(y)$, the equivalent of (10) becomes

$$\begin{aligned} D(h(x,y):f(x)g(y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) \\ &\quad \log\frac{h(x,y)}{f(x)g(y)}dxdy \\ &= -\int_{-\infty}^{\infty} f(x)\log f(x)dx \\ &\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) \\ &\quad \log\frac{h(x,y)}{g(y)}dxdy \end{aligned} \qquad (12)$$

provided, of course, that the integrals exist. The first right-side term in (12) is the continuous analog of the discrete entropy in (1). However, this continuous entropy may be arbitrarily large, positive or negative.

Parameterized Generalizations

A number of generalizations of the entropy in (1) have been proposed by introducing one or more arbitrary real-valued parameters. The best known such generalization is the *Rényi entropy of order* α (Rényi 1961) defined by

$$H_{R\alpha}(P_n) = \frac{1}{1-\alpha}\log_2\sum_{i=1}^{n} p_i^{\alpha}, \alpha > 0, \alpha \neq 1 \qquad (13)$$

with α being an arbitrary parameter. Rényi (1961) defined $\alpha \in (-\infty, \infty)$, but $\alpha > 0$ is used in (13) to avoid mathematical difficulties when some $p_i = 0$. While he used only the base -2 logarithm, the natural logarithm could obviously also be used in (13). Rényi's entropy has many of the same properties as Shannon's entropy in (1), including the additivity in property (P6). In the limit as $\alpha \to 1$, and using l'Hôpital's rule, (13) reduces to (1) with base -2 logarithm, showing that Shannon's entropy is a special case of Rényi's entropy.

The second early parameterized generalization of Shannon's entropy was that of Havrda and Charvát (1967) and given by

$$H_{HC\alpha}(P_n) = \frac{1}{2^{1-\alpha}-1}\left(\sum_{i=1}^{n} p_i^{\alpha} - 1\right), \alpha > 0, \alpha \neq 1 \qquad (14)$$

of which the entropy in (1) with base -2 logarithm is the limiting case as $\alpha \to 1$. If $1-\alpha$ is used as the denominator in (14), as is often done, then (1) with base $-e$ (natural) logarithm would be the limiting case of (14) as $\alpha \to 1$.

There have been numerous other generalized entropy formulations (e.g., Arndt 2004, Ch. 6; Kapur 1994; Pardo 2006, Ch. 1). In order to include the great majority of those formulations within a single parameterized generalization, one needs a four-parameter generalized entropy as introduced by Kvålseth (Kvålseth 1991, 1994), i.e.,

$$H_{\alpha\beta\delta}^{\lambda}(P_n) = \lambda\left[\left(\frac{\sum_{i=1}^{n} p_i^{\alpha}}{\sum_{i=1}^{n} p_i^{\delta}}\right)^{\beta} - 1\right] \qquad (15a)$$

$$0 < \alpha < 1 \le \delta,\ \beta\lambda > 0;\ \text{or},\ 0 \le \delta \le 1 < \alpha,\ \beta\lambda < 0 \qquad (15b)$$

with the parameter restriction in (15b) necessary for the entropy in (15a) to be non-negative and strictly Schur-concave. In order for the expression in (15a) to be non-negative, it is found to be sufficient that $(\delta-\alpha)\beta\lambda > 0$ for any real-valued α, β, δ, and λ (using the convention $0^c = 0$ for all real c), but (15b) is found to be required for strict Schur-concavity.

For the particular case when, for instance, $\lambda = \lambda_1 = [(1-\alpha)\beta]^{-1}$ and $\delta = 1$, it follows from (15a) that

$$H_{\alpha 01}^{\lambda_1}(P_n) = \lim_{\beta\to 0} H_{\alpha\beta 1}^{\lambda_1}(P_n) = \frac{1}{1-\alpha}\log_e\left(\frac{\sum_{i=1}^{n} p_i^{\alpha}}{\sum_{i=1}^{n} p_i}\right), \alpha > 0 \qquad (16)$$

which is Rényi's entropy of order α for a *possibly incomplete distribution* when $\sum_{i=1}^{n} p_i \le 1$ (Rényi 1961). The equivalent

form of Shannon's entropy becomes

$$H_{001}^{\lambda_1}(P_n) = \lim_{\alpha\to 0} H_{\alpha 01}^{\lambda_1}(P_n) = \frac{-1}{\sum_{i=1}^{n} p_i} \sum_{i=1}^{n} p_i \log_e p_i \qquad (17)$$

for $\sum_{i=1}^{n} p_i \le 1$. L'Hôpital's rule is used for the limits in (16)–(17). If, instead of setting $\lambda = \lambda_1 = [(1-\alpha)\beta]^{-1}$, $\lambda = \lambda_2 = (2^{(1-\alpha)\beta} - 1)^{-1}$ is used in (15a), then the special cases in (16)–(17) would involve base −2 logarithms. Two very simple members of (15) are the *quadratic entropy* $1 - \sum_{i=1}^{n} p_i^2$ and the *cubic entropy* $1 - \sum_{i=1}^{n} p_i^3$ for $\sum_{i=1}^{n} p_i = 1$ (Kapur and Kesavan 1992: 326–340). Note that, because of the term $\sum_{i=1}^{n} p_i^{\delta}$ in (15), this general formulation also includes incomplete distributions $\left(\sum_{i=1}^{n} p_i < 1\right)$ and hence also the entropy of a single event. Thus, since the generalized entropy, or four-parameter family of entropies, in (15) is zero-indifferent (see also Property (P3)), the entropy of a single event of probability p becomes

$$H_{\alpha\beta\delta}^{\lambda}(p, 0, \ldots, 0) = \lambda(p^{(\alpha-\delta)\beta} - 1)$$

subject to the parameter restriction in (15b) and with two simple particular cases being $1 - p$ (e.g., when $\lambda = 1, \alpha = 2, \delta = \beta = 1$) and the corresponding odds $1/p - 1$.

Various parameterized generalizations have also been proposed for additive and non-additive directed divergence measures. Arndt (2004), Kapur (1994), and Pardo (2006) provide overviews of such generalizations.

While such generalized entropy formulations provide interesting mathematical exercises and explorations, the practical utility of such efforts seem to be somewhat limited. Unless the added flexibility of such generalizations prove useful in particular situations, Shannon's entropy in (1) and (17) is the entropy of choice.

About the Author

Dr. Tarald O. Kvålseth (Ph.D. in Industrial Engineering and Operations Research, University of California, Berkely, 1971) is Professor Emeritus in mechanical engineering, industrial engineering, and biomedical engineering at the University of Minnesota, Minneapolis, USA, where he also served as Director of the Industrial Engineering Division. Before joining the University of Minnesota, he was first on the faculty at Georgia Institute of Technology, USA, then at the Norwegian Institute of Technology where he was Head of Industrial Management. He is the author of numerous papers and editor of three books on ergonomics and has been a member of various organizations, including the American Statistical Association, The Mathematical Association of America, the Institute of Ergonomics & Human Factors, and is a Fellow of the American Association for the Advancement of Science. He has served as chair and committee member of several international conferences and been a member of editorial boards and association boards, including Chair of the Nordic Ergonomics Committee and Vice President of the International Ergonomics Association. He is listed in Marquis Who's Who, including Who's Who in in Science and Engineering, Who's Who in America, and Who's Who in the World.

Cross References

►Bayesian Statistics
►Diversity
►Entropy and Cross Entropy as Diversity and Distance Measures
►Kullback-Leibler Divergence
►Maximum Entropy Method for Estimation of Missing Data
►Measurement of Uncertainty
►Radon–Nikodým Theorem
►Statistical View of Information Theory

References and Further Reading

Aczél J, Daróczy Z (1975) On measures of information and their characterization. Academic, New York
Arndt C (2004) Information measures: information and its description in science and engineering. Springer, Berlin
Garner WR (1962) Uncertainty and structure as psychological concepts. Wiley, New York
Havrda J, Charvát F (1967) Quantification method of classification processes. Kybernetika 3:30–35
Kapur JN (1994) Measures of information and their applications. Wiley, New York
Kapur JN, Kesavan HK (1992) Entropy optimization principles with applications. Academic, San Diego
Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86
Kvålseth TO (1991) On generalized information measures of human performance. Percept Mot Skills 72:1059–1063
Kvålseth TO (1994) Correction of a generalized information measure. Percept Mot Skills 79:348–350
Magurran AE (2004) Measuring biological diversity. Blackwell, Oxford
Marshall AW, Olkin I (1979) Inequalities: theory of majorization and its applications. Academic, San Diego
Mathai AM, Rathie PN (1975) Basic concepts in information theory and statistics. Wiley Eastern, New Delhi
Pardo L (2006) Statisctical inferences based on divergence measures. Chapman & Hall/CRC, Boca Raton
Rényi (1961) On measures of entropy and information. In: Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Berkeley, vol 1, pp 547–561
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656

Entropy and Cross Entropy as Diversity and Distance Measures

C. R. Rao
Distinguished Professor Emeritus and Adviser
C. R. RAO AIMCS, Hyderabad, India
Eberly Professor Emeritus in Statistics
Pennsylvania State University, University Park, PA, USA

> *We view this pile of data as an asset to be learned from. The bigger the pile, the better – if you have the tools to analyze it, to synthesize it, and make yourself more and more creative.*
> Britt Mayo

Introduction

Let $\mathcal{P}$ be a convex set of probability distributions defined on a measurable space $(\mathcal{X}, \mathcal{B})$. The classical definition of entropy as a measure of randomness in $p \in \mathcal{P}$ is what is known as Shannon entropy

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

in the case of a ▶multinomial distribution in n classes with class probabilities, $p_1, \ldots, p_n$, and

$$H(p) = -\int p \log p \, dv \tag{2}$$

in the case of a continuous distribution. These measures were used by Boltzman (1872) and Gibbs (1875–1878) in describing some equilibrium states in thermodynamics and Shannon (1948) in information theory.

The function (1) has been adopted by ecologists as a diversity measure in discussing relative abundance of different species of animals or plants in a locality. Some early references are Pielou (1975), Patil and Taillie (1982), and Rao (1982a, b, c).

While $H(p)$ is defined on $\mathcal{P}$, there is another function $C(q|p)$, defined on $\mathcal{P} \times \mathcal{P}$, called cross entropy (CE), not necessarily symmetric in p and q, designed to examine how close a surrogate model q is to a true model p. A well-known CE is Kullback and Leibler (1951) divergence measure

$$C(q|p) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \tag{3}$$

in the discrete case of n class multinomial distributions, and

$$C(q|p) = \int p \log \frac{p}{q} dv \tag{4}$$

in the case of continuous distributions.

In this paper, a general discussion of entropy and cross entropy measures, their characterizations, and applications to problems in statistics are given.

Entropy Functional

We state some general postulates governing an entropy function H on $\mathcal{P}$, conceived of as a measure of diversity (or randomness or uncertainty in prediction) as discussed in Rao (1982a, 1986).

$$A_1 : H(p) \geq 0 \quad \forall\, p \in \mathcal{P} \text{ and } H(p) = 0 \text{ iff } p \text{ is degenerate} \tag{5}$$

$$A_2 : H(\lambda p + \mu q) - \lambda H(p) - \mu H(q) = J(p, q : \lambda, \mu) \geq 0 \tag{6}$$

$$\forall\, p,\ q \in \mathcal{P},\ \lambda \geq 0,\ \mu \geq 0,\ \lambda + \mu = 1, \text{ and } = 0 \text{ iff } p = q$$

While A_1 requires H to be a non-negative function, A_2 implies that uncertainty increases if we contaminate p with another member q, i.e., $H(p)$ is a strictly concave function on $\mathcal{P}$. In a paper titled *Entropy* in this volume, Kvålseth gives a number of postulates governing H in addition to A_1 and A_2 and discusses the consequences of each postulate. One important difference is the additional postulate requiring symmetry of H in (5) with respect to $p_1, \ldots, p_n$. Some of the consequences of the postulate A_2 as stated in (6) are as follows:

1. In the class of multinomial distributions, $H(p)$ attains the maximum when all class probabilities are equal if H is symmetric in $p_1, \ldots, p_n$.
2. Dalton and Pielou used the following condition in characterizing a diversity measure:

$$H(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_n) \leq H(p_1, \ldots, p_i + \delta, \ldots, p_j - \delta, \ldots, p_n)$$
$$\text{if } p_i < p_i + \delta \leq p_j - \delta < p_j.$$

This is implied by A_2 if H is symmetric in $p_1, \ldots, p_n$.

Some examples of entropy functions in the case of multinomial distributions with $p_1, \ldots, p_n$ as class probabilities are:

1. $-\sum p_i \log p_i$, Shannon entropy
2. $1 - \sum {p_i}^2$, Gini-Simpson entropy
3. $(1-\alpha)^{-1} \log \sum p_i^{\alpha}$, $\alpha > 0$, $\alpha \neq 1$, Rényi entropy
4. $(\alpha - 1)^{-1} \left(1 - \sum p_i^{\alpha}\right)$, Havrda and Charvát entropy
5. $\sum\sum d_{ij} p_i p_j$, Rao's (1984) quadratic entropy where $d_{11} = \ldots = d_{kk}$ and the $(k-1) \times (k-1)$ matrix $(d_{ik} + d_{jk} - d_{ij} - d_{kk})$ is non-negative definite.

Properties of some of these entropy functions are discussed in Kvålseth in this volume. In the continuous case, some of the entropy functions are:

6. $-\int p \log p \, dv$, Shannon entropy, which may be negative for some p. For example, if

$$p = 2 \text{ for all } x \text{ in } (0, 1/2), \quad -\int p \log p \, dv$$
$$= -\int_0^{1/2} 2 \log 2 dv = -\log 2 < 0$$

7. $(1-\alpha)^{-1} \log \int p^{\alpha} dv$, $\alpha > 0$, $\alpha \neq 1$, Rényi entropy
8. $\int k(x,y)p(x)p(y)dv_x dv_y$, Rao's (1984) quadratic entropy where k is a conditionally negative definite kernel, i.e.,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j) a_i a_j < 0$$

for any n and $x_1, \ldots, x_n$ and $a_1, \ldots, a_n$ such that $\sum a_i = 0$.

Rao's quadratic entropy in the case of multinomial distributions is a function of both the class probabilities and possible differences in species in other aspects. It may not be a symmetric function of class probabilities. For a discussion on the use of Rao's quadratic entropy as an appropriate tool in the ecological studies, reference may be made to papers by Pavoine et al. (2005), Ricotta and Szeidl (2006) and Zoltán (2005). It may be noted that Rao's quadratic entropy reduces to the expression for the variance of the distribution p if $k(x,y) = (x-y)^2$.

Maxwell and Boltzmann obtained what is known as Maxwell-Boltzmann distribution of elementary particles (such as molecules) by maximizing Shannon entropy function, as in 6. above, subject to a restriction on p. Rao (1973, pp. 172–175) obtained the corresponding model by maximizing Rényi's entropy, as in 7. above, subject to the same restriction on p. It would be of interest to try other entropy functions and compare different models.

ANODIV

R. A. Fisher introduced the method of Analysis of Variance (ANOVA) for partitioning the variance of a set of measurements into several components such as "between" and "within" populations, "first order and higher order interactions" of factors, etc. Can a similar analysis be done with other measures of variability such as the mean deviation in the case of quantitative variables and a measure of diversity in the case of qualitative measurements? For instance we may have n populations of individuals and each individual in a population is scored for one of k possible skin colors. We may ask what is the average diversity of skin color within populations? What is the difference in skin color between populations as a whole?

The key to this lies in the choice of a suitable measure of diversity whether the measurements involved are qualitative or quantitative satisfying the postulate A_2

$$H(\lambda_1 p_1 + \cdots + \lambda_k p_k) - \sum_{i=1}^{k} \lambda_i H(p_i) = J_1(\{p_i\}, \{\lambda_i\}) \geq 0 \tag{7}$$

where the first term is the diversity of individuals in the mixed population, the second term is average diversity within populations and the difference is attributed to diversity between populations.

The function J in (7) is called Jensen difference of order 1, and H satisfying (7) as a diversity measure of order 1. Using (7) we have one-way ANODIV (Analysis of Diversity) as in Table 1.

where $P_\bullet = \sum \lambda_i p_i$. In practice, we estimate $H(P_\bullet)$ and $H(P_i)$ based on observed data and λ_i as proportional to sample size n_i of observations from the ith population. The ratio $G = B/T$, has been termed as genetic index of diversity in some population studies to interpret differences between populations as in Lewontin (1972), Nei (1973), and Rao (1982b). It may be noted that G depends on the choice of a suitable measure of diversity relevant to problem under consideration. For some applications reference may be made to Rao (1982a, b, 1984b).

Now, we investigate the condition on $H(\mathcal{P})$ for carrying out a two-way ANODIV. Let us consider populations denoted by $P_{ij}, i = 1, \ldots, r$ and $j = 1, \ldots, s$, where the first index i refers to a locality and the second index j to a specific community. Further, let $\lambda_i \mu_j$ be the relative strength of individuals in locality i and community j, where λ_i and μ_j are all non-negative and $\sum \lambda_i = \sum \mu_j = 1$. We may ask: how are the genetic contributions of individuals between localities and between communities as a whole different? Is the magnitude of genetic differences between the communities different in different localities? Such questions concerning the rs populations can be answered by a two-way ANODIV as in Table 2.

Entropy and Cross Entropy as Diversity and Distance Measures. Table 1 One-way ANODIV for k populations

Due to		Diversity
Between populations	(B)	$J_1(\{p_i\}, \{\lambda_i\})$
Within populations	(W)	$\sum \lambda_i H(p_i)$
Total	(T)	$H(P_\bullet)$

Entropy and Cross Entropy as Diversity and Distance Measures. Table 2 Two-way ANODIV

Due to		Diversity
Localities	(L)	$H(P_{\cdot\cdot}) - \sum_{i=1}^{r}\lambda_i H(P_{i\cdot})$
Communities	(C)	$H(P_{\cdot\cdot}) - \sum_{j=1}^{s}\mu_j H(P_{\cdot j})$
Interaction	(LC)	* by subtraction
Between populations	(B)	$H(P_{\cdot\cdot}) - \sum\lambda_i\mu_j H(P_{ij})$
Within populations	(W)	$\sum\sum\lambda_i\mu_j H(P_{ij})$
Total	(T)	$H(P_{\cdot\cdot})$

In Table 2, $P_{\cdot\cdot} = \sum\sum\lambda_i\mu_j P_{ij}, P_{i\cdot} = \sum\mu_j P_{ij}, P_{\cdot j} = \sum\lambda_i P_{ij}$. What are the conditions under which the entries in Table 2 are non-negative? For B, L, and C to be non-negative, the function $H(\cdot)$ defined on $\mathcal{P}$ should be strictly concave. For the interaction LC to be non-negative

$$\begin{aligned}(LC) &= -\left[H(P_{\cdot\cdot}) - \lambda_i H(P_{i\cdot})\right] + \sum\mu_j\left[H(P_{\cdot j}) - \lambda_i H(P_{ij})\right] \\ &= -J_1\left(\{P_{i\cdot}\} : \{\lambda_i\}\right) + \sum\mu_j J_1\left(\{P_{ij}\} : \{\lambda_i\}\right) \\ &= J_2\left(\{P_{ij}\} : \{\lambda_i\mu_j\}\right) \geq 0 \end{aligned} \tag{8}$$

or J_1 as defined in (7) on $\mathcal{P}^r$ is convex. We represent this condition as C_2 on a diversity measure. If C_1 and C_2 hold, we call such a diversity measure as of order 2. Note the (LC) can also be expressed as

$$\begin{aligned}(LC) &= -\left[H(P_{\cdot\cdot}) - \sum\mu_j H(P_{ij})\right] + \sum\lambda_i\Big[H(P_{i\cdot}) \\ &\qquad -\sum\mu_j H(P_{ij})\Big] \\ &= -J_1\left(\{P_{\cdot j}\} : \{\mu_j\}\right) + \sum\lambda_i J_1\left(\{P_{ij}\} : \{\mu_j\}\right) \end{aligned} \tag{9}$$

or J_1 as a function on $\mathcal{P}^s$ is convex.

We can recursively define higher order Jensen differences J_3 from J_2, J_4 from J_3 and so on, and call $H(\cdot)$ for which $J_0, J_1, \ldots, J_{i-1}$ are convex as the i^{th} order diversity measure. With such a measure we can carry out ANODIV for i-way classified data. A diversity measure for which Jensen differences of all orders are convex is called a perfect diversity measure.

Burbea and Rao (1982a, b, c) have shown that Shannon entropy satisfies the conditions C_0, C_1, and C_2, but not $C_3, C_4, \ldots$ The Havrda and Charvát entropy satisfies C_0, C_1, and C_2 for α in the range $(1, 2]$ when $k \geq 3$ and for α in the range $[1, 2] \cup (3, 11/3)$ when $k = 2$ and $C_3, C_4, \ldots$ do not hold except when $\alpha = 2$ in which case it reduces to Gini-Simpson index. Rényi's entropy satisfies C_0, C_1, and C_2 only for α in $(0, 1)$. Most of the well known entropy functions can be used only for two-way ANODIV but not for higher order ANODIV.

In the case of Rao's quadratic entropy, Jensen differences of all orders are convex, so that ANODIV can be carried out for any order classified data. It is shown by Lau (1985) that an entropy with this property is necessarily Rao's Quadratic Entropy.

Asymptotic distribution of ratios of entries in the ANODIV table, which can be used for tests of hypotheses as in ANOVA, can be obtained by bootstrapping as illustrated in Liu and Rao (1995).

Entropy Differential Metric

Using Fisher Information matrix, Rao (1945) introduced a quadratic differential metric, known as Fisher-Rao metric, over the parameter space of a family of probability distributions and proposed the geodesic distance (Rao distance) induced by the metric as a measure of dissimilarity between probability distributions. Burbea and Rao (1982c) introduced ϕ-entropy functional

$$H_\phi(p) = -\int \phi[p(x)]dv \tag{10}$$

and derived the quadratic differential metric as the Hessian of ϕ, assuming a parametric family of probability densities

$$p(x) = p(x, \theta_1, \ldots, \theta_n).$$

This opens up a wide variety of differential metrics. For instance, the choice of Shannon entropy with

$$\phi(p) = p\log p \tag{11}$$

in (10) gives Fisher-Rao metric. The reader is referred to Burbea and Rao (1982b, c) for the details. Maybank (2008) has written a number of papers on the application of Fisher-Rao metric, Rao distance and the associated Rao Measure. The possibility of using Burbea-Rao metric, which offers wider possibilities may be explored.

Cross Entropy

Characterization

There are situations where the true probability distribution p is not known but we us a surrogate distribution q for an analysis of data arising from p, or p is known but it is easy to generate observations from q to estimate some quantities such as the probability of large deviations in p, by a technique known in statistics as importance sampling as explained in section "▶Some Applications of CE". See also Rubinstein and Kroese (2004). In such a case we need to select q from $\mathcal{Q}$, a chosen family of distributions, such that q is close to an optimum distribution, using a suitable measure of closeness. One such measure, called

cross entropy (CE), used in some problems is Kullback-Leibler (1951) measure of divergence

$$C(q|p) = \int p \log \frac{p}{q} dv. \tag{12}$$

We suggest a few postulates for the choice of CE:

$$B_1 : \; C(q|p) \geq 0 \;\; \forall \; p, q \in \mathcal{P} \text{ and } C(q|p) = 0 \text{ only if } p = q. \tag{13}$$

$$B_2 : \; C(q|\lambda p + \mu q) \leq C(q|p), \;\; \lambda \geq 0, \;\; \mu \geq 0, \;\; \lambda + \mu = 1. \tag{14}$$

The postulate B_2 is a natural requirement as the mixture $\lambda p + \mu q$ has some component of q which brings q closer to $\lambda p + \mu q$.

There are several ways of constructing CEs depending on its use in solving a given problem. A loss function approach is as follows: let p be true probability density and we wish to find q such that

$$\int l(p(x), q(x)) p(x) dx \tag{15}$$

is a minimum, where l is a measure of difference between the likelihoods $p(x)$ and $q(x)$. The choice

$$l(p(x), q(x)) = \log \frac{p(x)}{q(x)} \tag{16}$$

leads to ▶Kullback-Leibler divergence measure

$$C(q|p) = \int p \log \frac{p}{q} dv. \tag{17}$$

The choice

$$l(p(x), q(x)) = \frac{[p(x) - q(x)]^2}{p(x) q(x)} \tag{18}$$

introduced by Rao (2010) leads to

$$\int \frac{(p-q)^2}{pq} p dv = \int \frac{p^2}{q} dv - 1, \tag{19}$$

which has some interesting properties.

A general version of CE, known as Csizar generalized measure, is

$$\int l\left(\frac{q}{p}\right) p dv$$

where

$$l \geq 0, \;\; l(1) = 0 \text{ and } l''(0) > 0. \tag{20}$$

Rao and Nayak (1985) gave a general method of deriving a CE from a given entropy function $H(p)$ as

$$C(q|p) = \lim_{\lambda \to 0} \frac{H(q + \lambda(p-q)) - H(q)}{\lambda} + H(q) - H(p). \tag{21}$$

The choice of $H(p)$ as Shannon entropy in (21) leads to (17), the Kullback-Leibler divergence measure.

Some Applications of CE

A comprehensive account of the use of CE in estimating probability of large deviations and a variety of stochastic and non-stochastic optimization problems is given in Rubinstein and Kroese (2004).

An example of how cross entropy is used is as follows. Suppose the problem is that of estimating $\gamma = E_p[\Phi(x)]$, the expectation of a function $\Phi(x)$ with respect to a given probability distribution $p(x)$. For instance, if we want to find the probability of $x \geq a$, as in the problem of large deviations. we can express it as the expectation of the function $I_{x \geq a}$, where I is the indicator function.

A general Monte Carlo technique of estimating γ is to draw a sample $x_1, \ldots, x_n$ from $p(x)$ and estimate γ by

$$\hat{\gamma} = n^{-1} \sum \Phi(x_i). \tag{22}$$

Observing that

$$\int \Phi(x) p(x) \, dx = \int \Phi(x) \frac{p(x)}{q(x)} q(x) \, dx \tag{23}$$

and

$$\gamma = E_q \left[\Phi(x) \frac{p(x)}{q(x)} \right] \tag{24}$$

we may draw a sample $(x_1', \ldots, x_n')$ from q, known as an importance sampling distribution, and estimate γ by

$$\hat{\gamma} = n^{-1} \sum \Phi(x_i') \, p(x_i') / q(x_i'). \tag{25}$$

The best choice of q which reduces the variance of $\hat{\gamma}$ to zero is

$$q^*(x) = \frac{\Phi(x) p(x)}{\gamma}. \tag{26}$$

However, the solution depends on the unknown γ. An alternative is to choose a family of sampling distributions indexed by a number of parameters

$$q(x, \theta), \;\; \theta = (\theta_1, \ldots, \theta_s) \tag{27}$$

and estimate θ by minimizing $C[q(x, \theta)|q^*(x)]$ with respect to θ. If we are using KL divergence measure, the problem reduces to

$$\max_{\theta} \int q^*(x) \log q(x, \theta) dv \tag{28}$$

which can be solved analytically or by maximizing with respect to the stochastic counterpart

$$\sum_{i=1}^{n} \frac{q^*(x_i)}{q(x_i, \overline{\theta})} \log q(x, \theta) \tag{29}$$

with respect to θ, where, $x_1, \ldots, x_n$ are observations drawn from $q(x, \overline{\theta})$, choosing a fixed value $\overline{\theta}$ of θ. Use of (19) as CE reduces the variance of the estimate (25).

The CE method can also be used to find the maximum or minimum of a mathematical function $f(x)$ defined over a region $\mathcal{R}$, subject to some equality and inequality constraints as in the Travelling Salesman, pattern recognition (see ►Pattern Recognition, Aspects of and ►Statistical Pattern Recognition Principles), and clustering problems. The problem is converted to a stochastic problem by choosing a probability distribution $p(x)$ over the given region $\mathcal{R}$ and a constant γ, and applying the algorithm used in the problem of large deviations, Prob $[f(x) > \gamma]$. The probability distribution $p(x)$ and the constant γ are altered step by step till the large deviation probability becomes negligibly small. For details, the reader is referred to Rubinstein and Kroese (2004).

Epilogue

In his paper in this volume on *The Future of Statistics*, Efron mentions that the nature of statistical methods is constantly changing with the availability of large data sets and increase in computing power. Early development of statistical theory and practice during the first half of the last century was based on parametric models for observed small data sets, and a set of prescribed rules for testing given hypotheses and estimating parameters of the chosen model. A critical discussion of these methods is given in the paper, *Has statistics a future*? *If so, in what form*? (*with discussion*), by Rao (2003).

Efron's bootstrap (see ►Bootstrap Methods) and new computer intensive methods such as boosting, bagging, CART, Lasso, Lars, projection pursuit, machine learning methods as in training ►neural networks, and random forests, have put statistical methodology in a different perspective without use of specific models for data. To these may be added the CE approach described in this paper, which provides a comprehensive computational algorithm for solving a variety of statistical problems, such as estimation of large deviations (see ►Large Deviations and Applications), pattern recognition and clustering, DNA sequence alignment and non-stochastic optimization problems such as The Travelling Salesman Problem.

I would like to mention that the aim of statistical analysis is not just to answer specific questions posed by the customer, but to find hidden patterns in given data, described as ►exploratory data analysis emphasized by Tukey, or ►data mining or bottom-up analysis to use modern terminology. Such an analysis will enable us to frame new hypotheses leading to possible expansion of knowledge. Some of these hypotheses could be tested with the existing data and some of them may need further acquisition of data. For further discussion on the need of statistics in solving problems of the real world, development of statistical methodology in view of technological advances in data collection, availability of large data sets, and emergence of new disciplines like ►bioinformatics raising new problems for data analysis, and training of statisticians, reference may be made to Efron's paper in this volume and Rao (2003).

About the Author

Calyampudi Radhakrishna Rao (born on September 10, 1920, in India) is among the most important statisticians in the past 60 years. He was the eighth in a family of 10 children. Following the usual custom, he was named after the God Krishna, who was also the eighth child of his parents. Rao received MA degree in mathematics from Andhra University (1941) and MA degree in statistics from Calcutta University (1943) with a first class, first rank and a gold medal. He started working in the Indian Statistical Institute (ISI), at Calcutta in 1941. In 1944 while giving a course on estimation at the Calcutta University to the senior students of the master's class a student asked him whether a result similar to the Fisher's information inequality for the asymptotic variance of a consistent estimate exists in small samples. During the same night Rao discovered the result that is today widely known as Cramér-Rao inequality. He obtained his Ph.D. in 1948 from the King's College, Cambridge University, UK, with R. A. Fisher as his thesis advisor (Fisher requested that Rao choose a research problem for his Ph.D. thesis himself). On his return to India Rao was appointed at the ISI as Professor at the age of 29. In India he held several important positions, as the Director of the ISI (1972–1976), Jawaharlal Nehru Professor and National Professor. He was the founder of Indian Econometric Society and the Indian Society for Medical Statistics. Rao took mandatory retirement at the age of 60 and moved to USA where he was appointed as a professor at the University of Pittsburgh (1979–1988). In 1988 he moved to the Pennsylvania State University accepting the newly established Chaired Professorship as Eberly Professor of Statistics (1988–2001). Since 2001 he is Eberly Professor Emeritus of Statistics. Professor Rao has been the President of the International Biometric Society (1973–1975), Institute of Mathematical Statistics (1976–1977) and International Statistical Institute (1977–1979). He has (co-)authored about 350 research papers, 14 books, including *Linear Statistical Inference and its Applications* (John Wiley 1965) and *Statistics and Truth: Putting Chance to Work* (World Scientific Publishing, 1989), both translated into six languages. Professor Rao

is the only living statistician whose two single-authored papers (*Information and the accuracy attainable in the estimation of statistical parameters* and *Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems in Estimation*) which had a high impact on the development of statistical theory were included in the book *Breakthroughs in Statistics* (Eds. S. Kotz and N. L. Johnson, Springer, 1993/1997). His name is associated with many statistical terms like: Cramér-Rao inequality, Rao-Blackwellization, Rao's Score Test, Fisher-Rao and Rao Theorems on second order efficiency of an estimator, Rao's U test, Fisher-Rao metric, and Rao distance. Other technical terms bearing his name appearing in specialized books are Rao's Quadratic Entropy, Cross Entropy, Rao's Paradoxes, Rao-Rubin, Lau-Rao, Lau-Rao-Shanbhag and Kagan-Linnik-Rao theorems on characterization of probability distributions. For his pioneering contributions to statistical theory and applications Professor Rao has won numerous awards and honors. Just to mention a few, he is a Fellow of Royal Society, UK (1967), a member of the Academy of Arts and Science (1975), and National Academy of Science (1995), US, and a member of three National Academies in India. Professor Rao was awarded numerous medals, including: Guy Medal in Silver of the Royal Statistical Society (1965), Jagdish Chandra Bose Gold Medal (1979), Samuel S. Wilks Memorial Medal (1989), Mahalanobis Birth Centenary Gold Medal (1996), Army Wilks Award (2000). On June 12, 2002, he received the National Medal of Science, the highest scientific honor by the United States. Professor Rao has also received other prestigious awards, such as: Padma Vibhushan award (2001), International Mahalanobis Prize (2003), and India Science Award (2010). For his outstanding achievements, C.R. Rao has been honored with the establishment of an institute named after him: C.R. Rao Advanced Institute for Mathematics, Statistics and Computer Science (C.R. Rao AIMSCS), located in the campus of the University of Hyderabad, India (http://aimscs.org). Nine special issues of several leading international journals were published (1991–2002) and six international conferences were organized in honor of Professor Rao (1980–2010). He was the doctoral thesis advisor for over 50 students who in turn produced more than 350 Ph.D.'s. Professor Rao holds 32 honorary doctorates from universities in 18 countries spanning six continents.

"The first half of the century was the golden age of statistical theory, during which our field grew from ad hoc origins, similar to the current state of computer science, into a firmly grounded mathematical science. Men of the intellectual caliber of Fisher, Neyman, Pearson, Hotelling, Wald, Cramér and *Rao* were needed to bring statistical theory to maturity." (Bradley Efron (1995). The Statistical Century, *Royal Statistical Society News*, **22**, 5, 1–2.)

Cross References

►Bayesian Statistics
►Diversity
►Entropy
►Kullback-Leibler Divergence
►Maximum Entropy Method for Estimation of Missing Data
►Measurement of Uncertainty
►Radon–Nikodým Theorem
►Statistical View of Information Theory

References and Further Reading

Boltzmann, Ludwig (1872) Further Studies on the Thermal Equilibrium of Gas Molecules (Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen), In: Sitzungsberichte der Akademie der Wissenschaften, Mathematische-Naturwissenschaftliche Klasse (pp. 275–370), Bd. 66, Dritte Heft, Zweite Abteilung, Vienna, Gerold

Burbea J, Rao CR (1982a) On the convexity of divergence measures based on entropy functions. IEEE Trans Inf Theory 28:489–495

Burbea J, Rao CR (1982b) On the convexity of higher order Jensen differences based on entropy functions. IEEE Trans Inf Theory 28:961–963

Burbea J, Rao CR (1982c) Entropy differential metric distance and divergence measures in probability spaces: a unified approach. J Multivariate Anal 12:575–596

Burbea J, Rao CR (1984) Differential metrics in probability spaces. Probab Math Stat 3:241–258

Efron, B (2010) The future of statistics, this volume

Friedman JH, Popescu BE (2003) Importance sampled learning ensembles. Technical Report, Stanford University

Gibbs JW (1875–1878) On the equilibrium of heterogeneous substances. Connecticut Academy of Sciences. (Reprinted in The Scientific Papers of J. Willard Gibbs. Dover, New York (1961))

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Kvålseth TO (2010) Entropy, this volume

Lewontin RC (1972) The apportionment of human diversity. Evol Biol 6:381–398

Liu ZJ, Rao CR (1995) Asymptotic distribution of statistics based on quadratic entropy and bootstrapping. J Stat Plan Infer 43:1–18

Maybank SJ (2008) The Fisher-Rao metric. Math Today 44:255–257

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Nat Acad Sci 70:3321–3323

Pavoine S, Ollier S, Pontier D (2005) Measuring diversity with Rao's quadratic entropy: are any dissimilarities suitable? Theor Popul Biol 67:231–239

Patil GP, Taillie C (1982) Diversity as a concept and its measurement. J Am Stat Assoc 77:548–567

Pielou EC (1975) Ecological diversity. Wiley, New York

Rao CR (1945) Information and accuracy attainable in the estimation of statistical parameters. Bull Calcutta Math Soc 37:81–91

Rao CR (1973) Linear statistical inference and its applications. Wiley, New York, pp 172–175

Rao CR (1982a) Diversity and dissimilarity coefficients: a unified approach. Theor Popul Biol 21:24–43
Rao CR (1982b) Diversity, its measurement, decomposition, apportionment and analysis. Sankhya 44:1–21
Rao CR (1982c) Gini-Simpson index of diversity: a characterization, generalization and applications. Utilitas Math 21:273–282
Rao CR (1984a) Convexity properties of entropy functions and analysis of diversity. In: Tong YL (ed) Inequalities in statistics and probability, IMS Lecture notes, vol 5, IMS, Hayward, pp 68–77
Rao CR (1984b) Use of diversity and distance measures in the analysis of qualitative data. In: Vark GN, Howels WW (eds) Multivariate statistical methods in physical anthropology: a review of recent advances and current developments, D. Reidel Pub. Co., Boston, pp 44–67
Rao CR (1986) Rao's axiomatization of diversity measures. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 7. Wiley, New York, pp 614–617
Rao CR (2003) Has statistics a future? If so in what form? (with discussion). In: Gulati C, Lin Y-X, Mishra S, Rayner J (eds) Advances in statistics, combinatorics and related areas, World Scientific, Singapore, pp 211–246
Rao CR (2010) Cross entropy with applications to problems of large deviations, combinatorial optimization and machine learning, University at Buffalo, Lecture delivered at the Bio Stat Dept
Rao CR, Nayak T (1985) Cross entropy, dissimilarity measures and quadratic entropy. IEEE T Inform Theory 31:589–593
Ricotta C, Szeidl L (2006) Towards a unifying approach to diversity measures: bridging the gap between Shannon entropy and Rao's quadratic entropy. Theor Popul Bio 70:237–243
Rubinstein RV, Kroese DP (2004) The cross entropy method: a unified approach to combinatorial optimization, monte carlo simulation and machine learning. Springer, New York
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656
Zoltán BD (2005) Rao's quadratic entropy as a measure of functional diversity based on multiple traits. J Veg Sci 16:533–540

Environmental Monitoring, Statistics Role in

Jennifer Brown
President of the New Zealand Statistics Association, Head
University of Canterbury, Christchurch, New Zealand

Environmental monitoring is conducted to provide information on status, and changes in status, of an environmental system. Often monitoring is associated with an impact, such as a proposed land development activity or rehabilitation of a habitat. At other times, environmental monitoring is conducted to assess the success (or failure) of a new environment management strategy or change in strategy. Monitoring can also be carried out to provide information on the overall status of a land or water area of special interest in fields such as biodiversity, the welfare of an endangered species or the abundance of a pest species.

In all these examples, the common theme is that monitoring is conducted to provide information. This information may be used in reports and articles that are created to bring about change in management. Typically, such information is numerical – a summary statistic, or a set of data – or some type of numerical measure. This is the role of statistics – statistics is the process used to collect and summarize data to provide relevant information for environmental monitoring.

Some underlying principles apply to information collected from environmental monitoring. The first is that in any monitoring design, the aims and objectives need to be clearly stated, both in a temporal and a spatial scale. The most successful monitoring programmes have aims and objectives that can be quantified to guide development of the survey design and data analysis (Gilbert 1987).

The survey design for the monitoring programme should specify how information is to be collected. Various survey designs can be used, and the important criterion is that they should provide a sample that is representative of the population and provide information relevant to the survey objective. The population can be considered to be an area of land or water that has fixed and delineated boundaries. A reserve or national park is an example of such a population. Other populations may be a species of interest, e.g., a bird population. In the example of a bird species, the population may be finite, although of unknown size, but the spatial boundaries may be unknown if the birds are highly mobile. In other applications, defining the population may be very difficult. For example, when monitoring the impact of a new industrial development in a rural area, delineating the area beyond which there is unlikely to be an effect may be very difficult.

Sample designs that use an element of probability (probability sampling) are recommended so that sample survey theory can be used to estimate sample precision (Thompson 2002). Examples of designs are simple random sampling, stratified sampling, grid-based sampling and spatially balanced designs (Stevens and Olsen 2004).

Statistics is used when a modeling approach is needed to describe the environmental system or for measuring the size of an environmental effect. Examples of an environmental effect are the changes (measured in some way) in environmental status of an area over time, or difference in environmental quality among sites receiving different management treatments (Manly 2008).

The very nature of environmental monitoring is that data are collected over time to allow assessment of change. A special type of statistical analysis is used for such data to account for the temporal correlation. Repeated measures analysis and longitudinal analysis are two terms used to describe different analyses for when environmental samples are collected from the same population over time (Manly 2008, Diggle et al. 2002). A consideration in surveys that are repeated over time is whether the same sites should be visited on each survey occasion or whether new sites should be selected and visited at each occasion. Designs that combine aspects of both are often used, where on each survey occasion, a mix of previous and new sites are surveyed (See, for example, Skalski 1990).

A feature of environmental monitoring that is designed to detect any impact resulting from an event is that information on the environment in the absence of the impact needs to be collected. An environmental impact describes events such as a new industrial development or the implementation of a new management strategy to protect an endangered species. The statistical term for these designs is before/after control/impact (BACI) (Underwood 1994).

In a BACI study, information is collected before the impact both at the control sites and at the site(s) of the future impact. Control sites are those that will not be affected by the impact event. Information is then collected from control sites and from the actual impact sites. This design provides information on the environment in the absence of the impact temporally (i.e., before the impact) and spatially (i.e., the control sites). The BACI statistical analysis considers whether the difference between the control and impact sites increases or decreases after the impact event. The analysis provides a way of quantifying this change in differences. Clearly, for some events, the complete BACI design cannot be used, e.g., unplanned impacts such as oil spills.

A growing area of statistics is spatial analysis and the use of geographic information systems (GIS) to map and describe environmental systems. Environmental monitoring can be designed to provide information on geographic changes in spatial patterns and distributions. Often maps and other well-designed graphical displays can be very informative. A final comment must be made about the logistical support for environmental monitoring. Monitoring programmes rely on (and often assume) data being collected accurately in the field. Whether this is true or not depends heavily on whether the field team have appropriate training and support. It is very important to have consistency in data collection protocols to ensure that any observed variation in summary statistics over time is a result of changes in the environmental system and not simply changes in field staff or their ability. Other considerations are that data need to be recorded accurately and stored in an accessible way.

Statistics plays a vital role in environmental monitoring because the essential ingredient to monitoring is data. A well planned, designed and executed monitoring programme can provide a wealth of information to guide future management and help ensure we maintain healthy environments.

About the Author

Professor Brown is President of the New Zealand Statistical Association, a position she has held since 2008. She is the Head of the Department of Mathematics and Statistics at University of Canterbury. Professor Brown is currently an Associate Editor of *Journal of Agricultural, Biological and Environmental Statistics*, of the *Australian and New Zealand Journal of Statistics*, and of the *International Journal of Ecological Economics and Statistics*.

Cross References

►Geostatistics and Kriging Predictors
►Mathematical and Statistical Modeling of Global Warming
►Spatial Statistics
►Statistics of Extremes

References and Further Reading

Diggle PJ, Heagerty PJ, Liang KY, Zeger SL (2002) Analysis of longitudinal data, 2nd edn. Oxford University Press, Oxford
Gilbert RO (1987) Statistical methods for environmental pollution monitoring. Wiley, New York
Manly BFJ (2008) Statistics for environmental science and management, 2nd edn. Chapman & Hall/CRC
Skalski JR (1990) A design for long-term status and trend. J Environ Manage 30:139–144
Stevens DL Jr, Olsen AR (2004) Spatially balanced sampling of natural resources. J Am Stat Assoc 99:262–278
Thompson SK (2002) Sampling, 2nd edn. Wiley, New York
Underwood AJ (1994) On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecol Appl 4:3–15

Equivalence Testing

Dieter Hauschke
Professor
University Medical Centre Freiburg, Freiburg, Germany

Many clinical trials have the objective of showing equivalence between two treatments, usually a test treatment

under development and an existing reference treatment. In such studies the aim is no longer to detect a difference, as in the case when comparing a new treatment with placebo, but to demonstrate that the two active treatments are equivalent within a priori stipulated acceptance limits. Let Y_T and Y_R designate the primary clinical outcome of interest for the test and reference treatment, respectively. A two-sample situation is considered where it is assumed that the outcomes are mutually independent and normally distributed with unknown but common variance σ^2,

$$Y_{Tj} \sim N(\mu_T, \sigma^2),\ j = 1, \ldots, n_1, \text{ and } Y_{Rj} \sim N(\mu_R, \sigma^2),\ j = 1, \ldots, n_2.$$

For equivalence testing it is reasonable to assume that the signs of the corresponding population means μ_T and μ_R are the same and, without loss of generality, positive. Let the interval (δ_1, δ_2), $\delta_1 < 0 < \delta_2$, denote the pre-specified equivalence range, so that the corresponding test problem can be formulated as follows:

$$H_0 : \mu_T - \mu_R \le \delta_1 \text{ or } \mu_T - \mu_R \ge \delta_2 \quad \text{vs}$$
$$H_1 : \delta_1 < \mu_T - \mu_R < \delta_2.$$

A split of the above two-sided test problem into two one-sided test problems (Schuirmann 1987) results in

$$H_{01} : \mu_T - \mu_R \le \delta_1 \quad \text{vs} \quad H_{11} : \mu_T - \mu_R > \delta_1$$

and

$$H_{02} : \mu_T - \mu_R \ge \delta_2 \quad \text{vs} \quad H_{12} : \mu_T - \mu_R < \delta_2.$$

According to the intersection–union principle (Berger and Hsu 1996), H_0 is rejected at significance level α in favor of H_1 if both hypotheses H_{01} and H_{02} are rejected at significance level α:

$$T_{\delta_1} = \frac{\overline{Y}_T - \overline{Y}_R - \delta_1}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{1-\alpha, n_1+n_2-2} \text{ and}$$
$$T_{\delta_2} = \frac{\overline{Y}_T - \overline{Y}_R - \delta_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{1-\alpha, n_1+n_2-2},$$

where $\overline{Y}_T$ and $\overline{Y}_R$ denote the corresponding sample means, $t_{1-\alpha, n_1+n_2-2}$ is the $(1-\alpha)$ quantile of the central Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom and

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2 - 2}\left(\sum_{j=1}^{n_1}(Y_{Tj} - \overline{Y}_T)^2 + \sum_{j=1}^{n_2}(Y_{Rj} - \overline{Y}_R)^2\right)$$

is the pooled estimator of σ^2.

This is equivalent to

$$\overline{Y}_T - \overline{Y}_R - t_{1-\alpha, n_1+n_2-2}\,\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} > \delta_1 \text{ and}$$

$$\overline{Y}_T - \overline{Y}_R + t_{1-\alpha, n_1+n_2-2}\,\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \delta_2,$$

and hence to the inclusion of the two-sided $(1-2\alpha)100\%$ confidence interval for $\mu_T - \mu_R$ in the equivalence range

$$\left[\overline{Y}_T - \overline{Y}_R - t_{1-\alpha, n_1+n_2-2}\,\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \overline{Y}_T - \overline{Y}_R + t_{1-\alpha, n_1+n_2-2}\,\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] \subset (\delta_1, \delta_2).$$

In clinical practice the equivalence limits δ_1 and δ_2 are often expressed as fractions of the unknown reference mean $\mu_R \neq 0$, i.e., $\delta_1 = f_1\mu_R$ and $\delta_2 = f_2\mu_R$, $-1 < f_1 < 0 < f_2$. For example $f_1 = -f_2 = -0.2$ corresponds to the common ±20% criterion. The test problem for equivalence can then be formulated as:

$$H_0 : \frac{\mu_T}{\mu_R} \le \theta_1 \text{ or } \frac{\mu_T}{\mu_R} \ge \theta_2 \quad \text{vs} \quad H_1 : \theta_1 < \frac{\mu_T}{\mu_R} < \theta_2,$$

where (θ_1, θ_2), $\theta_1 = 1 + f_1$, $\theta_2 = 1 + f_2$, $0 < \theta_1 < 1 < \theta_2$, is the corresponding equivalence range for the ratio of the expected means μ_T and μ_R. The null hypothesis can be rejected in favor of equivalence, if

$$T_{\theta_1} = \frac{\overline{Y}_T - \theta_1\overline{Y}_R}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{\theta_1^2}{n_2}}} > t_{1-\alpha, n_1+n_2-2} \text{ and}$$
$$T_{\theta_2} = \frac{\overline{Y}_T - \theta_2\overline{Y}_R}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{\theta_2^2}{n_2}}} < -t_{1-\alpha, n_1+n_2-2}.$$

Hauschke et al. (1999) have shown that rejection of H_0 by the two tests T_{θ_1} and T_{θ_2} each at level α is equivalent to inclusion of the $(1-2\alpha)100\%$ confidence interval for μ_T/μ_R, given by Fieller (1954), in the equivalence range (θ_1, θ_2), with

$$[\theta_l, \theta_u] \subset (\theta_1, \theta_2) \text{ and } \overline{Y}_R^2 > a_R,$$

where

$$\theta_l = \frac{\overline{Y}_T\overline{Y}_R - \sqrt{a_R\overline{Y}_T^2 + a_T\overline{Y}_R^2 - a_Ta_R}}{\overline{Y}_R^2 - a_R},$$
$$\theta_u = \frac{\overline{Y}_T\overline{Y}_R + \sqrt{a_R\overline{Y}_T^2 + a_T\overline{Y}_R^2 - a_Ta_R}}{\overline{Y}_R^2 - a_R},$$
$$a_T = \frac{\hat{\sigma}^2}{n_1}t_{1-\alpha, n_1+n_2-2}^2, \quad a_R = \frac{\hat{\sigma}^2}{n_2}t_{1-\alpha, n_1+n_2-2}^2.$$

Note that the condition $\overline{Y}_R^2 > a_R$ implies that $\mu_R \neq 0$

$$\overline{Y}_R^2 > \frac{\hat{\sigma}^2}{n_2}t_{1-\alpha, n_1+n_2-2}^2 \Leftrightarrow \frac{|\overline{Y}_R|}{\hat{\sigma}\sqrt{\frac{1}{n_2}}} > t_{1-\alpha, n_1+n_2-2}.$$

It should be noted that in clinical trials, a significance level of $\alpha = 0.025$ is required for equivalence testing and this refers to the calculation of two-sided 95% confidence intervals (CPMP 2000). Hence, equivalence can be concluded at level $\alpha = 0.025$ if the corresponding two one-sided test problems can be rejected each at level $\alpha = 0.025$.

About the Author

Dr. Dieter Hauschke is a Professor in Biometry, Department of Statistics, University of Dortmund, Germany. Currently, he is working as a scientist at the Institute for Medical Biometry and Medical Informatics, University Medical Center Freiburg, Germany. He was a member of the Advisory Council of the International Biometric Society – German Region (2001–2005) and council member of the International Biometric Society (2006–2010). He is incoming Head (2008–2010), Head (2011–2012) of the committee biometry within German Association of Medical Informatics, Biometry and Epidemiology. He is member of the Editorial Board of the International Journal of Clinical Pharmacology and Therapeutics and he was Associate Editor of the *Biometrical Journal* (2006–2010). Since 2009, he is statistical consultant of the Drug Commission of the German Medical Association and of the Treat-NMD Neuromuscular Network. Prof. Hauschke is author or co-author of more than 75 research articles and book-chapters. He is also co-author of the textbook *Bioequivalence in Drug Development: Methods and Applications* (with V. Steinijans and I. Pigeot, Wiley, 2007).

Cross Reference

►Clinical Trials: An Overview
►Pharmaceutical Statistics: Bioequivalence
►Significance Testing: An Overview

References and Further Reading

Berger RL, Hsu JC (1996) Bioequivalence trials, intersection union tests and equivalence confidence sets. Stat Sci 11: 283–319

Committee for Proprietary Medicinal Products (2000) Points to consider on switching between superiority and non-inferiority. EMEA, London

Fieller E (1954) Some problems in interval estimation. J R Statist Soc B 16:175–185

Hauschke D, Kieser M, Diletti E, Burke M (1999) Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. Stat Med 18:93–105

Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 15: 657–680

Ergodic Theorem

Steven P. Lalley
Professor
University of Chicago, Chicago, IL, USA

Birkhoff's Ergodic Theorem

Birkhoff's theorem (see Birkhoff 1931) extends the strong law of large numbers to stationary processes. The theorem is most easily formulated in terms of *measure-preserving transformations*: If $(\Omega, \mathcal{F}, P)$ is a probability space then a measurable transformation $T : \Omega \to \Omega$ is *measure-preserving* if $EX = EX \circ T$ for every bounded random variable X defined on Ω. In this case P is said to be *T-invariant*. Let T be a measure-preserving transformation of a probability space $(\Omega, \mathcal{F}, P)$; then a random variable X defined on $(\Omega, \mathcal{F})$ is *T-invariant* if $X = X \circ T$ almost surely, and an event $F \in \mathcal{F}$ is *T-invariant* if its indicator function is. The collection of all T-invariant events $F \in \mathcal{F}$ is a σ-algebra, denoted by $\mathcal{I}$. The measure-preserving transformation T is *ergodic* if the only bounded, T-invariant random variables are almost surely constant. Denote by T^i the ith iterate of T, that is, $T^{i+1} = T \circ T^i$.

Birkhoff's Theorem. Let T be a measure-preserving transformation of a probability space $(\Omega, \mathcal{F}, P)$. Then for every integrable random variable X defined on $(\Omega, \mathcal{F})$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X \circ T^i = E(X \,|\, \mathcal{I}) \quad \text{almost surely.} \qquad (1)$$

If T is ergodic, then $E(X|\mathcal{I}) = EX$ almost surely, so in this case the sample averages converge almost surely to the expectation EX.

It is not difficult to see that if T is a measure-preserving transformation of a probability space $(\Omega, \mathcal{F}, P)$ then for every random variable X defined on $(\Omega, \mathcal{F})$ the sequence $X_n := X \circ T^n$ is a stationary sequence, that is, the joint distribution of $X_1, X_2, \ldots$ is the same as that of $X_2, X_3, \ldots$. Conversely, if $X_1, X_2, \ldots$ is a stationary sequence of real random variables defined on some probability space and if μ is the joint distribution of $X_1, X_2, \ldots$, viewed as a random element of $\mathbb{R}^\infty$, then the forward shift operator σ is a measure-preserving transformation of the probability space $(\mathbb{R}^\infty, \mathcal{B}^\infty, \mu)$. Birkhoff's theorem implies that if this shift is ergodic – in which case the stationary sequence $X_1, X_2, \ldots$ is said to be ergodic – then the sequence $X_1, X_2, \ldots$ obeys the strong law of large numbers. Kolmogorov's 0 – 1 Law implies that if the sequence $X_1, X_2, \ldots$ consists of i.i.d. random variables then it is ergodic, when viewed as a stationary sequence. Thus, the

usual strong law of large numbers for sample averages of independent, identically distributed random variables is a corollary of Birkhoff's theorem. In fact, Birkhoff's theorem implies much more: If the sequence $X_1, X_2, \ldots$ is stationary and ergodic, then for any Borel measurable function $F : \mathcal{R}^\infty \to \mathbb{R}$ the sequence

$$Y_n := F(X_n, X_{n+1}, \ldots)$$

is stationary and ergodic, and therefore obeys the strong law of large numbers provided the first moment $|EY_1|$ is finite.

Consequences of Birkhoff's Theorem

Ergodic Markov Chains

An immediate consequence of Birkhoff's theorem is a strong law of large numbers for additive functionals of *ergodic Markov chains*. A Markov chain (see ▶Markov Chains) X_n on a finite or denumerable state space $\mathcal{X}$ is said to be *ergodic* if all states communicate and all states are positive recurrent. Each ergodic Markov chain has a unique stationary probability distribution. If the initial state X_0 is distributed according to the stationary distribution then the sequence $X_0, X_1, X_{2,\ldots}$ is stationary and ergodic, and so the law of large numbers for real-valued functionals $F(X_n)$ follows directly. Moreover, even if the Markov chain is started in an initial distribution ν other than the stationary distribution π, the strong law of large numbers must hold, because the probability measure P^ν governing the non-stationary chain is absolutely continuous relative to the probability measure P^π governing the stationary chain. See Revuz (1984) or Meyn and Tweedie (2009) for similar results concerning Markov chains on non-denumerable state spaces.

Birkhoff's Theorem and Dynamical Systems

Birkhoff's theorem has important implications for deterministic dynamical systems, especially those governed by Hamilton's equations where the Hamiltonian has no explicit time dependence (see Arnold and Avez 1967). According to a fundamental theorem of Liouville, if $\{\Phi_t\}_{t\geq 0}$ is the phase flow of a Hamiltonian system on the phase space $\mathbb{R}^{2N}$ then for each $t > 0$ Lebesgue measure on $\mathbb{R}^{2N}$ is Φ_t-invariant. Lebesgue measure cannot be renormalized to be a probability measure; however, if the level surfaces $H = E$ of the Hamiltonian are *compact*, as is often the case, then Lebesgue measure induces on each energy surface a Φ_t-invariant probability measure, called the *Liouville measure*. The *ergodic hypothesis* of Boltzmann asserts that the Liouville measure-preserving transformation Φ_t is ergodic. When true this implies, by Birkhoff's theorem, that "time averages equal space averages." To date the ergodic hypothesis has been established only for some very special Hamiltonian systems.

The Shannon–MacMillan–Breiman Theorem

Let $\{X_n\}_{n\in\mathbb{Z}}$ be a two-sided stationary process valued in a finite alphabet A. For each finite sequence $x_1x_2\cdots x_n$ in A, denote by $p(x_1x_2\cdots x_n)$ the probability that $X_i = x_i$ for each index $i \in [1, n]$. Similarly, denote by $p(x_0|x_{-1}x_{-2}\cdots x_{-n})$ the conditional probability that $X_0 = x_0$ given that $X_i = x_i$ for every $-n \leq i \leq -1$. The martingale convergence theorem implies that for almost every sequence $x_{-1}x_{-2}\cdots$ (relative to the joint distribution of the process $\{X_n\}_{n\in\mathbb{Z}}$) the limit

$$p(x_0|x_{-1}x_{-2}\cdots) := \lim_{n\to\infty} p(x_0|x_{-1}x_{-2}\cdots x_{-n}) \tag{2}$$

exists. The *Kolmogorov–Sinai entropy* of the stationary process $\{X_n\}_{n\in\mathbb{Z}}$ is defined to be

$$h := -E\log p(X_0|X_{-1X_{-2}\cdots}). \tag{3}$$

Shannon-MacMillan-Breiman Theorem. If the stationary process $\{X_n\}_{n\in\mathbb{Z}}$ is ergodic then with probability one,

$$\lim_{n\to\infty} p(X_1X_2\cdots X_n)^{1/n} = e^{-h} \tag{4}$$

This can be deduced from Birkhoff's theorem, which implies directly that with probability one

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \log p(X_k|X_{k-1}X_{k-2}\cdots) = -h. \tag{5}$$

The Shannon–MacMillan–Breiman theorem is of fundamental importance in information theory, for it implies that entropy limits the "compressibility" of a "message" generated by a stationary, ergodic source. See Shannon and Weaver (1949) and Cover and Thomas (2006) for further information.

Kingman's Subadditive Ergodic Theorem

In the 80 years since Birkhoff's paper, dozens of extensions, generalizations, and other progeny of Birkhoff's theorem have been discovered. See Krengel (1985) for an excellent review. One of these subsequent extensions has proved to be of singular importance: this is Kingman's *subadditive ergodic theorem* (see Kingman 1973). Let T be an ergodic, measure-preserving transformation of a probability space $(\Omega, \mathcal{F}, P)$. A double array $\{S_{k,m}\}_{0\leq k\leq m}$ of real random variables defined on $(\Omega, \mathcal{F})$ is called a *subadditive process* relative to T if

$$S_{k,m} \circ T = S_{k+1,m+1}, \tag{6}$$

$$S_{k,n} \leq S_{k,m} + S_{m,n}, \quad \text{and} \tag{7}$$

$$\gamma := \inf_{n\geq 1} n^{-1} E S_{0,n} > -\infty. \tag{8}$$

Observe that if $S_{k,m} = \sum_{i=k+1}^{m} X \circ T^i$ where X satisfies the hypotheses of Birkhoff's theorem then $\{S_{k,m}\}$ is a subadditive process.

Kingman's Theorem. If $\{S_{k,m}\}$ is a subadditive process relative to an ergodic, measure-preserving transformation T of a probability space $(\Omega, \mathcal{F}, P)$, then

$$\lim_{n\to\infty} S_{0,n}/n = \gamma \quad \mu - \text{almost surely.} \tag{9}$$

Kingman's theorem has myriad uses in the study of percolation processes and interacting particle systems. See Liggett (1985) for some of the basic applications.

Subadditive processes also arise naturally in connection with random walks (see ►Random Walk) on groups and semigroups. A *right random walk* on a semigroup G started at the identity $X_0 = 1$ is the sequence X_n of partial products of a sequence $\xi_1, \xi_2, \ldots$ of i.i.d. G-valued random variables, with multiplication on the right:

$$X_n = \xi_1 \xi_2 \cdots \xi_n. \tag{10}$$

The special case where G is a matrix group or semigroup is of particular importance. Let d be an invariant metric on G, that is, a metric such that $d(x, y) = d(xz, yz)$ for all $x, y, z \in G$. For instance, if G is a discrete, finitely generated group then one might choose d to be the natural distance in the *Cayley graph*; if G is a matrix group then one might take d to be the distance induced by the Riemannian metric on G. In any case, the process

$$S_{k,m} := d(\xi_{k+1}\xi_{k+2}\cdots\xi_m, 1)$$

is subadditive, and so Kingman's theorem implies that $d(X_n, 1)/n$ converges almost surely to a constant γ (possibly $+\infty$). Similarly, if $G = \mathcal{M}_d$ is the semigroup of all $d \times d$ real matrices and $\|g\|$ denotes the usual matrix norm of a matrix $g \in G$, then

$$S_{k,m} := \log \|\xi_{k+1}\xi_{k+2}\cdots\xi_m\|$$

is subadditive, and so Kingman's theorem implies that if $\log \|\xi_1\|$ has finite first moment then

$$\lim_{n\to\infty} \|X_n\|^{1/n} = e^{\gamma} \tag{11}$$

almost surely. This is the celebrated *Furstenberg–Kesten* theorem (Furstenberg and Kesten 1960) of random matrix theory. See Bougerol and Lacroix (1985) for further development of the theory of random matrix products.

About the Author

Steven Lalley is a Professor of Statistics and Mathematics at the University of Chicago. He has also held positions at Columbia University, Purdue University, and l'Universite de Paris VI. He is a former Editor of the *Annals of Probability*, and is a Fellow of the Institute of Mathematical Statistics.

Cross References

►Almost Sure Convergence of Random Variables
►Glivenko-Cantelli Theorems
►Random Matrix Theory
►Random Walk

References and Further Reading

Arnold VI, Avez A (1967) Problèmes ergodiques de la mécanique classique. Monographies Internationales de Mathématiques Modernes, 9. Éditeur. Gauthier-Villars, Paris

Birkhoff GD (1931) Proof of the ergodic theorem. Proc Natl Acad Sci USA 17:656–660

Bougerol P, Lacroix J (1985) Products of random matrices with applications to Schrödinger operators. Progress in probability and statistics, vol 8. Birkhäuser Boston, Boston

Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley-Interscience, Hoboken

Furstenberg H, Kesten H (1960) Products of random matrices. Ann Math Stat 31:457–469

Kingman JFC (1973) Subadditive ergodic theory. Ann Probab 1:883–909. With discussion by Burkholder DL, Daryl Daley, Kesten H, Ney P, Frank Spitzer and Hammersley JM, and a reply by the author

Krengel U (1985) Ergodic theorems. de Gruyter studies in mathematics, vol 6. Walter de Gruyter, Berlin. With a supplement by Antoine Brunel

Liggett TM (1985) Interacting particle systems. Grundlehren der Mathematischen Wissenschaften [Fundamental principles of mathematical sciences], vol 276. Springer, New York

Meyn S, Tweedie RL (2009) Markov chains and stochastic stability, 2nd edn. Cambridge University Press, Cambridge. With a prologue by Peter W. Glynn

Erlang's Formulas

M. F. Ramalhoto
Professor
Technical University of Lisbon, Lisbon, Portugal

Mathematical calculations on probabilities have been put in writing mainly since the seventeenth century with the work of mathematicians like Fermat (1601–1665), Pascal (1623–1662), Huygens (1629–1695), Bernoulli (1654–1705), Moivre (1667–1754), Laplace (1749–1827), Gauss (1777–1855), Poisson (1781–1840), and Tchébychev (1821–1894). In the twentieth century the work of Markov, Liapounov,

Khintchine, Kolmogorov, Palm, Wiener, Fisher, Doob, Neumann, Cramer, Takács, Itó, and Santaló, among many others, on probability foundations, ▶stochastic processes, and mathematical statistics provided the basis for what it is called here "*Stochastics*."

There are two stochastic processes that are fundamental, and occur over and over again, often in surprising ways. The deepest results in "*Stochastics*" seem to be concerned with their interplay. One, the Wiener process (the Wiener model of Brownian motion, see ▶Brownian Motion and Diffusions), has been the subject of many books. The other, the Poisson process (the simplest renewal process, i.e., exponential inter-renewal times), has been comparatively a bit more neglected. Kingman (1993) redresses the balance and provides an enjoyable and clearly written introduction to the structure and properties of ▶Poisson processes in one and more dimensions (the book simultaneously addresses the beginner and the expert).

When, in 1875 (June 2) in Boston, Alexander Graham Bell accidentally discovered the possibility of telephone communication he could hardly foresee that by this discovery he had also created the inspiration for ▶queueing theory. In fact, the pioneer work on queueing theory is in Erlang (1909). A. K. Erlang was a Danish engineer and mathematician dealing with the problems and worries of the telephone communication system of the time, whose fundamental papers appeared between 1909 and 1920. He was responsible for the concept of statistical equilibrium, for the introduction of the so-called balance-of-state equations, and for the first consideration of the optimization of a queueing system. Rapid progress was made when the use of queueing theory spread out to many other fields in addition to telephone theory, and mathematicians like William Feller and David Kendall, among many others, became interested in the mathematics of queues. It would be difficult to give a brief outline of the development of the subject with a proper assignment of credits. Many of the most meritorious papers responsible for new methods are now rarely mentioned in the literature, in spite of the progress that they initiated; the D. V. Lindley's integral equation of the single-server queueing system, published in (1952), is an example. Queueing theory (including queueing networks and complex particle systems) is one of the most relevant branches of "*Stochastics*." Indeed, some of the today's most significant problems can be reduced to resource allocation and resource sharing. When the customers are human beings, very often patterns of human behavior and responses have to be brought into the mathematical analysis of the corresponding queueing systems. For instance, Little's law, a celebrated queueing property, does not necessarily hold due to the non-linear human's perception of waiting. Particularly in the queueing systems of the service industry, issues linked to quality improvement and innovation are now more than ever very relevant; an introduction to these issues can be founded in Ramalhoto (2000).

The power of the structure and the mathematical properties of the Poisson process (a few key results often produce surprising consequences) have been recognized in the queueing theory since its foundation in Erlang (1909).

First Erlang Formula

In 1917, A.K. Erlang (who also laid the foundations of modern teletraffic theory) from the analysis of the concept of statistical equilibrium obtained his famous formula (one of the most used formulas in practice even today) for the loss probability of the $M/G/r/r$ queueing system (i.e., Poisson arrival process, arbitrary service time distribution, r servers and zero waiting positions) in steady state:

$$B(r,r\rho) = \left[(r\rho)^r/r!\right]\left[\sum_{i=0}^{r}(r\rho)^i/i!\right]^{-1};\ r \in \mathbb{N},\ \rho \in \mathbb{R}^+ \quad (1)$$

where N and R^+ represent the natural numbers and the positive real numbers, respectively. Here, $r\rho = \lambda E[S]$ is the offered load, ρ is the intensity of traffic, and $E[S] = \mu^{-1}$ is the mean service time. The problem considered by Erlang can be phrased as follows. Calls arrive at a link as a Poisson process of rate λ. The link comprises r circuits, and a call is blocked and lost if all r circuits are occupied. Otherwise, the call is accepted and occupies a single circuit for the holding period of the call. Call holding periods are independent of each other and of arrival times, and are identically distributed with unit mean. This formula gives the proportion of calls that are lost in the $M/G/r/r$ queueing system in steady state. It is called the 1st *Erlang formula*, the Erlang's loss formula, or Erlang B formula. In several telecommunication studies the need arose to extend the definition of the 1st *Erlang formula* to non-integral values of r and to evaluate the derivatives of $B(r,r\rho)$ with respect to r and ρ. The most commonly used extension of the 1st *Erlang formula* is the analytic continuation

$$B(r,a)^{-1} = \int_0^{+\infty} e^{-u}(1-u/a)^r du \quad (2)$$

where $a = r\rho$ (offered load). This is known as the "*continued* 1st *Erlang function*." Jagerman (1984) presents improved computations of the 1st *Erlang formula* and its derivatives, with practical computer programs for immediate application.

Second Erlang Formula

The 2nd *Erlang formula* (also called the Erlang's delay formula or Erlang C formula), first published by Erlang in 1917, is defined in the $M/M/r$ queueing system (i.e., Poisson arrivals, exponential service time, r servers and infinite waiting positions) in steady state as follows:

$$C(r,r\rho) = \left[(r\rho)^r/r!(1-\rho)\right]\left[\sum_{i=0}^{r-1}(r\rho)^i/i! + (r\rho)^r/r!(1-\rho)\right]^{-1}; r \in \mathbb{N}, 0 < \rho < 1. \quad (3)$$

It gives the proportion of customers (calls) that are delayed in the $M/M/r$ queueing system in steady state. Unlike the 1st *Erlang formula*, the 2nd *Erlang formula* is not valid for an arbitrary service time distribution. From (1) and (3), it is clear that the 1st and 2nd *Erlang formulas* are related as follows:

$$C(r,r\rho) = \left[\rho + (1-\rho)B^{-1}(r,r\rho)\right]^{-1}; r \in \mathbb{N}, 0 < \rho < 1. \quad (4)$$

Further results and generalizations of the 1st and 2nd *Erlang formulas* may be found, for example, in the proceedings of international telecommunication and teletraffic conferences.

Third Erlang Formula

Let $C_d(r,r\rho)$ mean the probability that an arbitrary customer on its arrival finds r or more customers in the $M/M/r/r+d$ queueing system (i.e., Poisson arrivals, exponential service time distribution, r servers and d waiting positions) in steady state. That is to say, $C_d(r,r\rho)$ is the stationary probability that an arriving customer is blocked (i.e., not immediately served). It gives the proportion of customers that are delayed or lost in the $M/M/r/r+d$ queueing system in steady state. It is easy to show that $C_d(r,r\rho)$ can be written as follows:

$$C_d(r,r\rho) = \left[((r\rho)^r/r!)\sum_{i=0}^{d}\rho^i\right] \times \left[\sum_{i=0}^{r-1}(r\rho)^i/i! + ((r\rho)^r/r!)\sum_{i=0}^{d}\rho^i\right]^{-1}; \quad r \in \mathbb{N}, d \in \mathbb{N}_0, \rho \in \mathbb{R}^+. \quad (5)$$

From (1) and (5) it is clear that when $d = 0, C_d(r,r\rho)$ coincides with $B(r,r\rho)$, the 1st *Erlang formula*. From (3) and (5) it is clear that the limit of $C_d(r,r\rho)$ as d goes to infinity is the 2nd *Erlang formula* (which is defined only for $0 < \rho < 1$). Therefore, it is reasonable to call $C_d(r,r\rho)$ the 3rd *Erlang formula* (or Erlang's loss-delay formula). From (1) and (5) it is clear that the 3rd *Erlang formula* can be rewritten in terms of the 1st *Erlang formula* as follows:

$$C_d(r,r\rho) = \left[B^{-1}(r,r\rho)\left(\sum_{i=0}^{d}\rho^i\right)^{-1} + 1 - \left(\sum_{i=0}^{d}\rho^i\right)^{-1}\right]^{-1}; \quad r \in \mathbb{N}, d \in \mathbb{N}_0, \rho \in \mathbb{R}^+. \quad (6)$$

The relation (6) shows the ability of the 3rd *Erlang formula* to make use of the results available for the 1st *Erlang formula*; namely through the relation (2) it may be extended to non-integral values of r and d (i.e., the "*continued* 3rd *Erlang function*"). Monotonicity and convexity properties of the 3rd *Erlang formula* in terms of its parameters r, d, ρ, and $r\rho$ (offered load) are also easy to obtain. All these properties are useful in the study of the probabilistic behavior of the $M/M/r/r+d$ queueing system (also called Markovian multi-server finite-capacity queue or multi-server Erlang loss-delay queue). Mainly because most of the relevant system's characteristics can be rewritten as very simple functions of the 3rd *Erlang formula*. For instance, it is easy to show that in steady state both the loss probability and the delay probability depend on r only through the 3rd *Erlang formula*.

Exact Decomposition Formulas for the Multi-Server Erlang Loss-Delay Queue

For the $M/M/r/r+d$ queueing system in steady state each random variable (r. v.) $N'_{r,d}$ (the number of customers waiting in the queue), $NS_{r,d}$, (the number of occupied servers), $N_{r,d}$ (the number of customers in the system; waiting or being served), $T'_{r,d}$ (the waiting time in the queue), and $T_{r,d}$ (the total sojourn time in the system; waiting or being served) is distributed as the sum of two r. v.'s. weighted by the 3rd *Erlang formula*, for $r \in \mathbb{N}, d \in \mathbb{N}_0$ (where the symbol $\sim$ means "distributed as").

$$N'_{r,d} \sim (1 - C_d(r,r\rho))O + C_d(r,r\rho)N_{1,d-1}, \quad (7)$$

where $P(0 = o) = 1$ (i.e., a degenerated r. v.) and $N_{1,d-1}$ is the number of customers in the $M/M/1/1+(d-1)$ queueing system in steady state with the same ρ (which has a truncated geometric distribution with parameters d and $1-\rho$, denoted here by the symbol $G(d,1-\rho)$).

$$NS_{r,d} \sim (1 - C_d(r,r\rho))(X|X < r) + C_d(r,r\rho)R, \quad (8)$$

where $P(R = r) = 1$ and X is a Poisson r. v. of parameter $r\rho$.

$$N_{r,d} \sim (1 - C_d(r,r\rho))(X|X < r) + C_d(r,r\rho)(R + N_{1,d-1}), \quad (9)$$

$$T'_{r,d} \sim (1 - C_{d-1}(r,r\rho))O + C_{d-1}(r,r\rho)E_d, \quad (10)$$

where E_d is a "*generalized*" *Erlang* r. v. with parameters $G(d-1,1-\rho)$ and μ, respectively.

$$T_{r,d} \sim (1 - C_{d-1}(r, r\rho))S_r + C_{d-1}(r, r\rho)(E_d + S_r), \quad (11)$$

where S_r is the service time distribution of each server. The proof is essentially based on rewriting the respective distributions (that can be found in Gross and Harris (1985, pp. 93-102) or Gross et al. (2008)) in terms of the 3rd *Erlang formula*. Therefore, in the steady state, the probabilistic behavior of the $M/M/r/r + d$ queueing system can be obtained in terms of the corresponding probabilistic behavior of the $M/M/r/r$ queueing system and of the $M/M/1/1 + (d-1)$ queueing system, respectively; and its closeness to each one of these two queueing systems is measured by the 3rd *Erlang formula*. Taking limits as d goes to infinity in (7) to (11) (for $0 < \rho < 1$) similar results follow for the $M/M/r$ queueing system, in the steady state, in terms of the $M/M/r/r$ queueing system, the $M/M/1$ queueing systems and the 2nd Erlang formula, respectively.

Further Models

Due to their analytical tractability the queueing systems $M/M/\infty, M/M/1$, and $M/M/1/1 + d$ are by far the most extensively studied queueing systems in the steady state as well as in the time-dependent regime. Moreover, the use of the $M/M/\infty$ queueing system to approximate the $M/M/r/r$ queueing system (and their generalized cases for arbitrary service time distributions) has been studied by several authors, see, for instance, Ramalhoto (1999) and the references therein.

An extension of the decomposition formulas for the $M/M/r/r + d$ queueing system with constant retrial rate is provided in Ramalhoto and Gomez-Corral (1998). This type of queueing system (which, in fact, is a queueing network) is characterized by the feature that if on its arrival the customer finds all servers and waiting positions occupied the customer leaves the service area and enters an orbit, i.e., the retrial group. In the case $r + d < 3$, analytical results were obtained exactly in the same way, as before, by rewriting the corresponding distribution function in terms of the *probability of entering the orbit*. It is shown that the distributions of the "number of customers in orbit" (i.e., the orbit size) and the "total number of servers and waiting positions occupied" can be expressed also as mixtures of two r. v.'s weighted by the *probability of entering the orbit* (which in the retrial case corresponds to the 3rd *Erlang formula*). Another relevant property of these novel formulas is that they display the exact influence of the $M/M/r/r$ queue (which is well approximated by the $M/M/\infty$) and the probability of entering the orbit (in the retrial case) which somehow captures the effect of traffic intensity. Because the physics of the decomposition structures presented do not seem to have much to do with the values of r and d themselves empirical decomposition formulas (of similar kind of the $r + d < 3$ case) were proposed and heuristically explored. Based on them, numerical results for the stationary distribution of the orbit size were obtained for the queueing systems $M/M/2/2 + 1, M/M/3/3, M/M/3/3 + 1$, and $M/M/4/4$ with constant retrial, respectively. A similar methodology may be applicable to more complex types of retrials

The idea behind this decomposition concept is intuitive and simple to understand and use. And it seems to be suitable for obtaining bounds and other approximations for time-dependent regime and non-Markovian queueing systems as well as for new simulation strategies in queueing systems.

About the Author

M. F. Ramalhoto (Maria Fernanda Neto Ramalhoto) is Professor of Probability and Statistics, Department of Mathematics, at Instituto Superior Técnico (IST) of the Technical University of Lisbon since 1987. She was Vice-Rector of the Portuguese Open University (1990–1993). She was a Member of the Executive Council of IST (1987–1988) and the ERASMUS Institutional Coordinator at IST (1987–1990). She is a Fellow of the "Société Européenne pour la Formation des Engénieurs", and Associate Editor of the *European Journal for Engineering Education*. She was Visiting Professor in several universities worldwide, including MIT, Berkeley and Princeton, and Programme Committee Member of several international conferences in Europe, Israel and India. She was a Co-Founder and a Vice-President of the European Network for Business and Industrial Statistics (which now has over a thousand members and a successful annual conference). She was Expert/Evaluator for the Fifth and Sixth Frameworks of the European Union. She is Expert for the Seventh Framework. She is author and co-author (with colleagues from the US, UK, Germany, Italy, and Spain) of research papers in stochastic processes, statistics, quality control, and reliability. She has been Associate Editor of the international journal *Quality Technology and Quantitative Management* since its foundation.

Cross References

►Queueing Theory

References and Further Reading

Erlang AK (1909) The theory of probabilities and telephone conversations. Nyt Tidsskrift Matematik B 20:33–39

Erlang AK (1917) Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. Electroteknikeren (Danish) 13:5–13 (English translation in the PO Elec Eng J 10:189–197 (1917–1918))
Gross D, Harris CM (1985) Fundamentals of queueing theory, 2nd edn. Wiley, New York
Gross D, Shortle JF, Thompson JM, Harris CM (2008) Fundamentals of queueing theory, 4th edn. Wiley, New York
Jagerman DL (1984) Methods in traffic calculations. AT&T Bell Labs Tech J 63:1283–1310
Kingman JFC (1993) Poisson processes. Oxford studies in probability 3. Clarendon Press, Oxford
Lindley DV (1952) The theory of queues with a single server. Proc Camb Philos Soc 48:277–289
Ramalhoto MF (1999) The infinite server queue and heuristic approximations to the multi-server with and without retrials. TOP 7:333–350 (December)
Ramalhoto MF (2000) Stochastic modelling for quality improvement in processes. In: Park SH, Geoffrey Vining G (eds) Statistical process monitoring and optimization. Marcel Dekker, New York (Chapter 26), pp 435–456
Ramalhoto MF, Gomez-Corral A (1998) Some decomposition formulae for M/M/r/r + d queues with constant retrial rate. Commun Stat-Stoch Models 14(1):123–145

Estimation

Nancy Reid
Professor
University of Toronto, Toronto, ON, Canada

Introduction

Statistical models involve unknown quantities, usually called parameters of the model, and inference about these parameters provides, in principle, understanding about plausible data-generating mechanisms. Inference about these parameters is also a first step in assessing the adequacy of the proposed model, in predicting future patterns likely to be observed from this model, and as a basis for making decisions. Estimation is the process of using the data to make conclusions about the values of unknown quantities in a statistical model.

The theory of point estimation is concerned with a narrower problem: given a parametric model $f(y;\theta)$, that is a density on a sample space $\mathcal{Y}$, with unknown parameter θ taking values in a parameter space Θ, and a set of observations from this model, how do we use these observations to provide a good guess for the true value of θ? In ►nonparametric estimation, Θ is an infinite-dimensional space, for example the space of all distribution functions on $\mathbb{R}$; in parametric estimation Θ is usually a subspace of a well-behaved parameter space, such as $\mathbb{R}^p$ for fixed and known dimension p.

A point estimate of a parameter or a function θ is not often very useful without an associated statement about the accuracy of the point estimate, usually provided as an estimated standard error. More formally, the theory of interval estimation aims to estimate a set of plausible values of θ consistent with the data.

In the early development of statistical science, methods for constructing a point estimate were often developed in the context of particular problems, and some statistical properties of the resulting point estimate were then studied. This led to a number of strategies for point estimation, including the method of moments, the method of ►least squares, and, with Fisher (1922), the method of maximum likelihood estimation; see Aldridge (1997). Least squares estimates were used in the early 19th century in problems in astronomy (Stigler, 1981). The method of moments was very popular in the early 20th century, but except in very complex problems has been superseded by maximum likelihood estimation.

A method for estimation is evaluated by assessment of its statistical properties. Historically, there developed a very large literature on deciding which properties of an estimation method were most relevant, and which methods of estimation satisfied those properties. Some properties commonly proposed for "good" estimators are: unbiased, admissible, minimax, minimum variance, minimum mean-squared error, equivariant, consistent, efficient, asymptotically unbiased, asymptotically optimal, and robust. In the 1940s and 1950s statistical theory emphasized optimality of various methods of point estimation according to some criterion to be specified, such as variance, mean squared error, or risk. The conditions for optimal estimation, by any criterion, are rarely satisfied in a wide enough range of models to inform statistical practice, so attention turned to asymptotic optimality. The maximum likelihood estimator was shown to be consistent and asymptotically efficient, and is now generally computable with standard software, so is usually the first choice of estimation method, at least in parametric families of models. In the 1970s a theory of robust point estimation was developed, to formalize the notion that a "good" method of estimation should be stable under perturbations of the assumed model.

Theory and methods for nonparametric estimation of a wide variety of functional, or infinite-dimensional, parameters has developed very intensively over the past twenty years and have a prominent role in theoretical and methodological statistics. In this note however we will concentrate on finite-dimensional parametric estimation. The classic

text on the theory of point estimation is Lehmann (1983), subsequently revised in Lehmann and Casella (2003). Casella and Berger (1990) and Knight (2000) provide excellent treatments at a somewhat less advanced level. The exposition here draws heavily on Chapter 7 of Davison (2003) and Chapter 8 of Cox and Hinkley (1974).

Defining Point Estimators

Most methods of points estimation are based on minimizing some notion of distance between the data and some aspect of the fitted model. For example, least squares estimators are defined to minimize $\sum_{i=1}^{n}\{y_i - \mu_i(\theta)\}^2$, where we assume that $E(y_i) = \mu_i(\theta)$ has a known form, for example $x_i^T\theta$ for a given p-vector x_i^T. Least absolute deviation estimators minimize $\sum_{i=1}^{n}|y_i - \mu_i(\theta)|$. There may be other unknown parameters in the model; for example if y_i is normally distributed with mean $\mu_i(\theta)$ the variance may also be unknown, and a suitable estimator for the variance is usually taken to be the residual mean square $\Sigma\{y_i - \mu_i(\hat{\theta})\}^2/(n-p)$, where $\hat{\theta}$ is the estimate of the parameters in the mean, and θ is assumed to have dimension p. This estimator of σ^2 is unbiased, as defined below, and widely used, but not optimal under conventional distance measures.

Maximum likelihood estimators are defined to maximize the probability of the observed data, or equivalently to minimize $-2\log f(y;\theta)$. They are defined via

$$L(\hat{\theta}_{ML};y) = \sup_{\theta} L(\theta;y),$$

where $L(\theta;y) \propto f(y;\theta)$ is the likelihood function. In models with smooth differentiable likelihoods that have just one maximum,

$$\nabla_\theta \ell(\hat{\theta}_{ML};y) = 0, \tag{1}$$

where $\ell(\theta;y) = \log L(\theta;y)$. The maximum likelihood estimator can also be defined as an estimate of the value that minimizes the ▶Kullback–Leibler divergence between the parametric model and the true density for y:

$$D(f,g) = \int \log\left\{\frac{g(y)}{f(y;\theta)}\right\} g(y)dy,$$

where we temporarily assume that the true density is $g(\cdot)$ but we are fitting the parametric model $f(y;\theta)$. In the case of independent observations $y_1,\ldots,y_n$, the score equation (1) can be expressed as

$$n^{-1}\sum_{i=1}^{n}(\partial/\partial\theta)\log f(y_i;\hat{\theta}_{ML}) = 0,$$

which is an empirical version of the derivative of $D(f,g)$ with respect to θ. In machine learning applications maximum likelihood estimators are often derived from this point of view, and D is referred to as the negative log-entropy loss function.

The method of moments defines point estimators by equating the theoretical moments of the presumed model to the observed sample moments. These estimators have been largely superseded by maximum likelihood estimators, but are often useful starting points, especially in rather complex models defined, for example, for stochastic systems modeled through sets of partial differential equations. The other class of models for which method of moments estimators are sometimes used is in estimating components of variance in normal theory linear models.

Example 1 Suppose the vector of observations y is taken in k groups of size n, and modeled as

$$y_{ij} = \mu + b_i + \epsilon_{ij}, \quad j = 1,\ldots,n; i = i,\ldots,k,$$

where we assume that $b_i \sim N\left(0,\sigma_b^2\right)$, $\epsilon_{ij} \sim N(0,\sigma^2)$, and the b's and ϵ's are mutually independent. The analysis of between and within group sums of squares leads to the following two statistics:

$$SS_{\text{within}} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2,$$
$$SS_{\text{between}} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_{i.} - \bar{y}_{..})^2,$$

where $\bar{y}$ indicates averaging over the appropriate subscript. It is not difficult to show that

$$E(SS_{\text{within}}) = k(n-1)\sigma^2,$$
$$E(SS_{\text{between}}) = (k-1)\left(\sigma^2 + n\sigma_b^2\right),$$

which leads to simple method of moments estimators of σ^2 and σ_b^2. If there is very little sample variation among groups, then the estimate of σ_b^2 could be negative, as method of moments does not automatically obey constraints in the parameter space. This example can be extended to much more complex structure, with several levels of variation, and possible dependence among some of the random effects.

Point estimators can also be derived from a Bayesian point of view, although in a fully Bayesian context it is not necessary and may not even be wise to summarize the posterior distribution of the parameters by a point estimate. However, it is relatively straightforward to proceed, assuming the availability of a prior density $\pi(\theta)$, to construct the

posterior density

$$\pi(\theta \mid y) = L(\theta; y)\pi(\theta) / \int L(\theta; y)\pi(\theta)\mathrm{d}\theta,$$

and then to compute the posterior mode, $\hat{\theta}_\pi$, which maximizes $\pi(\theta \mid y)$, or the posterior median, or mean, or some other summary statistic. In most practical applications there will be much more information in the data than in the prior, and $\hat{\theta}_\pi$ will not be very different from the maximum likelihood estimator. This can be made more precise by considering asymptotic properties of Bayesian posterior distributions, under the sampling model. Empirical Bayes estimators can be obtained by using the data to estimate the parameters in the prior distribution, as well as the parameters in the model.

A class of estimators often called *shrinkage* estimators can be derived by minimizing a ►loss function, but imposing a penalty on the resulting estimator of some form. The simplest example is the ridge regression estimator (see ►Ridge and Surrogate Ridge Regressions) associated with ►least squares.

Example 2 Suppose y is an $n \times 1$ vector that follows a normal distribution with mean $X\beta$ and covariance matrix $\sigma^2 I$, where X is an $n \times p$ matrix and β is a $p \times 1$ vector of regression coefficients of interest. The least squares estimator of β minimizes $(y - X\beta)^T(y - X\beta)$, with solution, if X is of full column rank,

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

If the least squares problems is changed to

$$\min_{\beta} (y - X\beta)^T (y - X\beta), \qquad \text{subject to } \beta^T \beta \le t,$$

then the solution, which also minimizes the penalized sum of squares

$$(y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta,$$

is

$$\hat{\beta}_R = (X^T X + \lambda I)^{-1} X^T y;$$

the individual components $\hat{\beta}_j$ are shrunk towards zero. This is sometimes used if number of components of β is quite large relative to n, as in most settings this would suggest that the individual components cannot all be well determined from the data. There are a great many extensions of this idea, especially in the literature on fitting "smooth" functions, rather than well specified parametric functions, to data. For example, we might replace $X\beta$ by $m_1(x_1) + \cdots + m_p(x_p)$, without specifying very much about the form of m_j. One approach to estimating these functions is to model them as piecewise polynomial, with a large number of parameters, and then to shrink the polynomial coefficients. Empirical Bayes estimators are also usually shrinkage estimators; typically shrinking to some central data value, such as the sample mean, rather than a particular point of interest, such as 0, in the parameter space.

Evaluating Point Estimators

It is conventional to refer to a point estimate as the value computed from a given set of data, and the point estimator as the function of the observations that will be used in practice. Many texts distinguish these two ideas by using upper case letters for random variables, and lower case letters for observed values of these random variables.

Example 3 Suppose we assume that we have a sample $y = (y_1, \ldots, y_n)$ of independent observations of a random vector $Y = (Y_1, \ldots, Y_n)$ with joint density

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{(2\pi)^n \sigma^n}} \exp\left\{-\frac{1}{2\sigma^2} \sum (y_i - \mu)^2\right\}. \quad (2)$$

The sample mean $\hat{\mu} = (1/n)\Sigma y_i$ is an estimate of μ, and the estimator $\hat{\mu} = (1/n)\Sigma Y_i$ has a distribution determined by (2); for example $E(\hat{\mu}) = \mu$ and $\mathrm{var}(\hat{\mu}) = \sigma^2/n$. Note that the notation for $\hat{\mu}$ does not distinguish between the estimate and the estimator, but this will usually be clear from the context.

Suppose more generally that we have a vector of observations y of a random vector Y from a parametric model with density $f(y; \theta)$ where $y \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^p$.

Definition 1 An estimator $\hat{\theta} = \hat{\theta}(y)$ is *unbiased* for θ if

$$E\{\hat{\theta}(Y)\} = \theta.$$

The estimator is a minimum variance unbiased estimator if, for any other unbiased estimator $\tilde{\theta}$, we have

$$\mathrm{var}(\hat{\theta}) \le \mathrm{var}(\tilde{\theta}).$$

The expectation and variance are calculated under the model $f(y; \theta)$. When θ is a vector, $\mathrm{var}(\theta)$ is a matrix, and the variance condition above is generalized to $a^T \mathrm{var}(\hat{\theta}) a \le a^T \mathrm{var}(\tilde{\theta}) a$ for all real vectors a of length p. Minimum variance unbiased estimators rarely exist; an exception is in regular exponential family models, where the minimal sufficient statistic is a minimum variance unbiased estimator of its expectation.

An estimator $\hat{\theta}$ is (weakly) *consistent* for θ if it converges to θ in probability, under the model $f(y; \theta)$. A consistent estimator $\hat{\theta}$ is *asymptotically efficient* if the

asymptotic variance of its limiting distribution is as small as possible. The ▶Cramér–Rao inequality establishes that maximum likelihood estimators in regular models are asymptotically efficient, and the lower bound on the asymptotic variance for any consistent estimator is $i^{-1}(\theta)$, where $i(\theta)$ is the expected Fisher information in the model, $E\{-\partial^2 \ell(\theta; Y)/\partial\theta^2\}$.

Bayesian treatments of point estimation often emphasize that estimates based on posterior distributions obtained via proper priors are admissible, meaning that no non-Bayesian estimators have smaller Bayes risk. Bayes risk is defined with respect to a loss function, so depends in an intrinsic way on the units of measurement for the parameter. Admissible estimators are sometimes useful in decision theoretic contexts, but not often in scientific work.

In machine learning contexts where there is the possibility to fit rather complex models to large data sets, performance is often measured by mean-squared error, as biased point estimators such as the shrinkage estimator mentioned above are widely used. The phrase "bias-variance tradeoff" refers to the fact that a biased point estimator can have smaller variance than an unbiased estimator, but that bias and variance cannot be simultaneously minimized. Estimates of functional parameters, such as densities or regression functions, are often compared by means of integrated mean-squared error.

Example 4 Suppose Y is a sample of independent, identically distributed observations each from an unknown density $f(\cdot)$, and the goal is to estimate this density. The kernel density estimator of f is defined as

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right),$$

where $h > 0$ is a tuning parameter and $K(\cdot)$ is a kernel function, usually a symmetric probability density function, such as the density for a $N(0,1)$ distribution. The tuning parameter h controls the smoothness of the estimated function $\hat{f}(\cdot)$, by controlling the number of sample points that enter the average in estimating $f(y)$. It can be shown that the squared bias of $\hat{f}$ at a single point y is $O(h^4)$, and the variance is $O(nh)^{-1}$, so the optimal bandwidth can be shown to be $h \propto n^{-1/5}$, with the constant of proportionality depending on the point y and the kernel function $K(\cdot)$.

Robust estimators are defined, following Huber (1981), to be stable under perturbations of the postulated model. From a data analytic point of view, robust estimators are not much affected by ▶outliers, or extremely unusual data points. The motivation for this is that such outliers may be coding mistakes, or may be irrelevant for the inference desired. However this varies widely by application area.

The theory of estimating equations considers the primary object of interest as the equation defining the point estimator, rather than the point estimator itself. For an independent, identically distributed sample $y = (y_1, \ldots, y_n)$, from a model with parameter θ, an estimating equation for θ is given by

$$\sum_{i=1}^{n} g(y_i; \theta) = 0,$$

for a function $g(\cdot)$ to be chosen. An estimating function is unbiased if $E\{g(Y; \theta)\} = 0$, and it would be rare to start with a biased estimating function, as the resulting estimator would not be consistent for θ without further modification. The maximum likelihood estimator is, in regular models, defined by the estimating equation with $g(y; \theta) = (\partial/\partial\theta) \log f(y; \theta)$, but more general classes of estimating equations arise naturally in studies of robustness; see Davison (2003, Ch. 7).

About the Author

For biography *see* the entry ▶Likelihood.

Cross References

▶Asymptotic Relative Efficiency in Estimation
▶Bayesian vs. Classical Point Estimation: A Comparative Overview
▶Best Linear Unbiased Estimation in Linear Models
▶Cramér–Rao Inequality
▶Estimation: An Overview
▶Methods of Moments Estimation
▶Nonparametric Estimation
▶Properties of Estimators
▶Robust Regression Estimation in Generalized Linear Models
▶Statistical Inference
▶Statistical Inference: An Overview
▶Sufficient Statistical Information

References and Further Reading

Aldridge J (1997) R.A. Fisher and the making of maximum likelihood 1912–1922. Stat Sci 12:162–176
Casella G, Berger RL (1990) Statistical inference. Pacific Grove, California
Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London
Davison AC (2003) Statistical models. Cambridge University Press, Cambridge

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Phil Trans R Soc Lond A 22:309–368
Huber PJ (1981) Robust statistics. Wiley, New York
Knight K (2000) Mathematical statistics. Chapman and Hall, London
Lehmann EL (1983) Theory of point estimation. Wiley, New York
Lehmann EL, Casella G (2003) Theory of point estimation, 2nd edn. Springer-Verlag, New York
Stigler SM (1981) Gauss and the invention of least squares. Ann Statist 9:465–474

Estimation Problems for Random Fields

Mikhail P. Moklyachuk
Professor
Kyiv National Taras Shevchenko University, Kyiv, Ukraine

Estimation problems for random fields $X(t)$, $t \in \mathbb{R}^n$ (estimation of the unknown mathematical expectation, estimation of the correlation function, estimation of regression parameters, extrapolation, interpolation, filtering, etc) are similar to the corresponding problems for ▶stochastic processes (random fields of dimension 1). Complications usually are caused by the form of domain of points $\{t_j\} = D \subset \mathbb{R}^n$, where observations $\{X(t_j)\}$ are given, and by the dimension of the field. The complications can be overcome by considering specific domains of observations and particular classes of random fields.

Say in the domain $D \subset \mathbb{R}^n$ there are given observations of the random field

$$X(t) = \sum_{i=1}^{q} \theta_i g_i(t) + Y(t),$$

where $g_i(t), i = 1, \ldots, q$, are known non-random functions, $\theta_i, i = 1, \ldots, q$, are unknown parameters, and $Y(t)$ is a random field with $EY(t) = 0$. The problem is to estimate the regression parameters $\theta_i, i = 1, \ldots, q$. This problem includes as a particular case ($q = 1, g_1(t) = 1$), the problem of estimation of the unknown mathematical expectation. Linear unbiased least squares estimates of the regression parameters can be found by solving the corresponding linear algebraic equations or linear integral equations determined with the help of the correlation function. For the class of isotropic random fields, formulas for estimates of the regression parameters are proposed by M. I. Yadrenko (Yadrenko 1983). For example, the estimate $\hat{\theta}$ of the unknown mathematical expectation θ of an isotropic random field $X(t) = X(r,u)$ from observations on the sphere $S_n(r) = \{x \in \mathbb{R}^n, \|x\| = r\}$ is of the form

$$\hat{\theta} = \frac{1}{\omega_n} \int_{S_n(r)} X(r,u) m_n(du), n \geq 2,$$

where $m_n(du)$ is the Lebesgue measure on the sphere $S_n(r)$, ω_n is the square of the surface of the sphere, (r,u) are spherical coordinates of the point $t \in \mathbb{R}^n$.

Consider the extrapolation problem.

1. Observations of the mean-square continuous homogeneous and isotropic random field $X(t)$, $t \in \mathbb{R}^n$ are given on the sphere $S_n(r) = \{x \in \mathbb{R}^n, \|x\| = r\}$. The problem is to determine the optimal mean-square linear estimate $\hat{X}(s)$ of the unknown value $X(s)$, $s \notin S_n(r)$, of the random field. It follows from the spectral representation of the field that this estimate is of the form

$$\hat{X}(s) = \sum_{m=0}^{\infty} \sum_{l=1}^{h(m,n)} c_m^l(s) \int_0^{\infty} \frac{J_{m+(n-2)/2}(r\lambda)}{(r\lambda)^{(n-2)/2}} Z_m^l(d\lambda),$$

where coefficients $c_m^l(s)$ are determined by a special algorithm (Yadrenko 1983). For practical purposes it is more convenient to have a formula where observations $X(t)$, $t \in S_n(r)$, are used directly. The composition theorem for spherical harmonics gives us this opportunity. We can write

$$\hat{X}(s) = \int_{S_n(r)} c(s,t) X(t) dm_n(t),$$

where the function $c(s,t)$ is determined by the spectral function $\Phi(\lambda)$ of the field $X(t)$ (Yadrenko 1983).

2. An isotropic random field $X(t)$, $t = (r,u) \in \mathbb{R}^n$ is observed in the sphere $V_R = \{x \in \mathbb{R}^n, \|x\| \leq R\}$. The optimal liner estimate $\hat{X}(s)$ of the unknown value $X(s)$, $s = (\rho, v) \notin V_R$, of the field has the form

$$\hat{X}(s) = \int_{V_R} C(s,t) X(t) dm_n(t),$$

$$C(s,t) = \sum_{m=0}^{\infty} \sum_{l=1}^{h(m,n)} c_m^l(r) S_{in}^l(u),$$

where coefficients $c_m^l(r)$ are determined via special integral equations

$$b_m(\rho,q) S_m^l(v) = \int_0^R b_m(r,q) c_m^l(r) r^{n-1} dr, \qquad m = 0, 1, \ldots;$$
$$l = 1, 2, \ldots, h(m,n), \quad q \in [0, R].$$

If, for example, $X(t)$, $t = (r,u)$, is an isotropic random field where $b_m(r,q) = a^{|m|} \exp\{-\beta|r-q|\}$, then it is easy to see that $\hat{X}(\rho, v) = \exp\{-\beta|\rho - R|\} X(R, v), v \in S_n$.

For methods of solutions of other estimation problems for random fields (extrapolation, interpolation, filtering, etc.) see Grenander (1981), Moklyachuk (2008), Ramm

(2005), Ripley (1981), Rozanov (1982), Yadrenko (1983), and Yaglom (1987).

About the Author

For biography *see* the entry ▶Random Field.

Cross References

▶Estimation
▶Random Field

References and Further Reading

Grenander U (1981) Abstract inference. Wiley series in probability and mathematical statistics. Wiley, New York

Moklyachuk MP (2008) Robust estimates of functionals of stochastic processes. Vydavnycho-Poligrafichnyĭ Tsentr, Kyïvskyĭ Universytet, Kyïv

Ramm AG (2005) Random fields estimation. World Scientific, Hackensack, NJ

Ripley BD (1981) Spatial statistics. Wiley series in probability and mathematical statistics. Wiley, New York

Rozanov YA (1982) Markov random fields. Springer-Verlag, New York

Yadrenko MI (1983) Spectral theory of random fields. Translation series in mathematics and engineering. (Optimization Software Inc., New York) Springer-Verlag, New York

Yaglom AM (1987) Correlation theory of stationary and related random functions, volume I and II. Springer series in statistics. Springer-Verlag, New York

Estimation: An Overview

Sara van de Geer
Professor
ETH Zürich, Zürich, Switzerland

Introduction

Let P be a probability distribution on some space $\mathcal{X}$. A *random sample* from P is a collection of independent random variables $X_1, \ldots, X_n$, all with the same distribution P. We then also call $X_1, \ldots, X_n$ independent copies of a population random variable X, where X has distribution P. In statistics, the probability measure P is not known, and the aim is to estimate aspects of P using the observed sample $X_1, \ldots, X_n$. Formally, an estimator, say T, is any given known function of the data, i.e., $T = T(X_1, \ldots, X_n)$.

Example 1. *Suppose the observations are real-valued, i.e., the sample space $\mathcal{X}$ is the real line $\mathbb{R}$. An estimator of the population mean $\mu := EX$ is the sample mean $\hat{\mu} := \sum_{i=1}^n X_i/n$.*

A statistical model is a collection of probability measures $\mathcal{P}$. If the true distribution is in the model class $\mathcal{P}$, we call the model *well-specified*.

In many situations, it is useful to parametrize the distributions in $\mathcal{P}$, i.e., to write

$$\mathcal{P} := \{P_\theta : \theta \in \Theta\},$$

where Θ is the *parameter space*. The *parameter of interest* is some function of θ, say

$$\gamma := g(\theta) \in \Gamma.$$

Example 2.
Case (i) *Suppose X is known to be normally distributed with unknown mean μ and unknown variance σ^2 (we write this as $X \sim \mathcal{N}(\mu, \sigma^2)$). The parameter space is then the collection of 2-dimensional parameters $\theta := (\mu, \sigma^2)$. If one is only interested in estimating the mean μ, one lets $g(\mu, \sigma^2) = \mu$. The second parameter σ^2 is then called a nuisance parameter.*
Case (ii) *If actually the distribution of X is completely unknown, we may take*

$$\mathcal{P} := \{\text{all distributions on } \mathbb{R}\}.$$

We can let $\mathcal{P}$ itself be the parameter space, and the parameter of interest is written as function $g(P)$ of the probability measure P. For example, with $\mu = EX$ being the parameter interest, we have $g(P) := \int x dP(x)$.

How to Construct Estimators?

From a mathematical point of view, the construction of an estimator as function of the data $X_1, \ldots, X_n$ can be based on *loss* and *risk*. Let $g(\theta)$ be the parameter of interest, and $T = T(X_1, \ldots, X_n)$ be an estimator. With this estimator, we associate the *loss* $L(g(\theta), T)$, where L is a given ▶loss function. As T is a random variable, the loss is generally random as well. The *risk* of the estimator T is now defined as

$$R(\theta, T) := E_\theta L(g(\theta), T),$$

where the expectation E_θ is with respect to the probability measure of an i.id. sample $X_1, \ldots, X_n$ from P_θ. When the parameter of interest is real-valued, an important special case is quadratic loss

$$L(g(\theta), T) = |T - g(\theta)|^2.$$

The risk is then the mean square error

$$MSE_\theta(T) = E_\theta |T - g(\theta)|^2.$$

The mean square error can be separated into a squared bias term and a variance term

$$MSE_\theta(T) = \text{bias}_\theta^2(T) + \text{var}_\theta(T),$$

where

$$\text{bias}_\theta(T) := E_\theta T - g(\theta),$$

and

$$\text{var}_\theta(T) := E_\theta T^2 - (E_\theta T)^2.$$

An estimator T is called *unbiased* if

$$E_\theta T = g(\theta), \ \forall\ \theta.$$

It is called *uniform minimum variance unbiased* (UMVU) if it has the smallest variance among all unbiased estimators, for all values of the parameter.

Once the loss function $L(g(\theta), T)$ is specified, one may try to find the estimator T which has smallest risk $R(\theta, T)$. However, the risk depends on θ, which is unknown! This generally means that estimators are not directly comparable. One way to overcome this problem is to require the estimator to be the best one in the worst possible case:

$$\min_{\text{estimators } T} \max_{\theta\in\Theta} R(\theta, T).$$

This is called the *minimax approach*. Another way to treat the dependence on θ is to assign a priori weights to the various parameter values. This is called the Bayesian approach. Formally, let Π be some probability measure on Θ, the so-called prior. The Bayes' risk of the estimator T is then defined as

$$R(T) := \int R(\vartheta, T) d\Pi(\vartheta).$$

In other words, the unknown parameter is integrated out. The minimizer T_{bayes} of $R(T)$ over all estimators $T = T(X_1, \ldots, X_n)$ is called Bayes' estimator.

The (minimax, Bayes) risk generally heavily depends on the model class $\mathcal{P}$. This means that if the model is misspecified, the estimator may not have the optimality properties one hoped for. A related problem is the robustness of an estimator: how much does it change if some of the data points in the sample are perturbed? One often decides to stay on the conservative side, i.e. instead of believing in the model and aiming at mathematical optimality, one prefers to be sure to be doing something reasonable. And, last but not least, there may be practical cost considerations that prevent one from applying a particular estimator. Here, cost can also be computation time, e.g., if one needs online answers.

The asymptotic approach is to try to construct estimators that are "approximately" "optimal" under mild assumptions. For example, unbiased estimators often do not even exist. However, for large sample size n, many estimators are "approximately" unbiased. One then looks for the one with the smallest asymptotic variance, i.e., the highest asymptotic efficiency. Rigorous asymptotic theory can be mathematically involved, as a certain *uniformity* in the parameter θ is required.

The Plug in Principle

In this section, we assume that the sample size allows for an *asymptotic* approach, i.e., that n is large enough to base ideas on e.g. ▶laws of large numbers and ▶central limit theorems. We then see $X_1, \ldots, X_n$ as the first n of an infinite sequence. The *plug in* principle is mainly based on the law of large numbers. Consider for example a subset $A \subset \mathcal{X}$. The probability of this set is $P(X \in A)$, or shortly $P(A)$. As, by the law of large numbers, probabilities can be approximated by frequencies, a sensible estimator of $P(A)$ is the proportion of observations that fall into A:

$$\hat{P}_n(A) := \frac{1}{n}\#\Big\{i \in \{1, \ldots, n\} :\ X_i \in A\Big\}.$$

Note that $\hat{P}_n$ corresponds to a probability measure that puts mass $1/n$ at each observation. One calls $\hat{P}_n$ the *empirical distribution*.

Consider now a parameter of interest $\gamma = g(\theta)$. We write it as

$$\gamma := Q(P),$$

where Q is a given function on the probability measures $P \in \mathcal{P}$. In other words, $g(\theta) = Q(P_\theta)$. The plug in principle is now to use the estimator

$$\hat{\gamma}_n := Q(\hat{P}_n),$$

i.e., to replace the unknown probability measure P by the empirical measure. We remark however that is may happen that $Q(\hat{P}_n)$ is not defined (as generally $\hat{P}_n$ is not in the model class $\mathcal{P}$). In that case, one takes

$$\hat{\gamma}_n := Q_n(\hat{P}_n),$$

where

$$Q_n(P) \approx Q(P).$$

Example 3. *Suppose* $X \in \mathbb{R}$. *Let* $\mu = EX = \int x dP(x)$ *be the parameter of interest. The sample mean is*

$$\hat{\mu}_n = \int x d\hat{P}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i.$$

Example 4. *Let the sample space again be the real line. The distribution function of* X *is*

$$F(x) := P(X \le x), \qquad x \in \mathbb{R}.$$

The empirical distribution function is

$$\hat{F}_n(x) = \frac{1}{n}\#\Big\{i \in \{1, \ldots, n\} :\ X_i \le x\Big\}.$$

The Glivenko–Cantelli Theorem tells us that the law of large numbers hold, uniformly in x (In fact, by Donsker's Theorem, $\sqrt{n}(\hat{F}_n - F)$ converges in distribution to $B \circ F$, where B is a Brownian bridge.)

$$\sup_x |\hat{F}_n(x) - F(x)| \to 0, \qquad n \to \infty,$$

with probability one. Suppose we know that F has a density f with respect to Lebesgue measure, and that we aim at estimating f at a fixed point x. Then

$$Q(F) = f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}.$$

Note that $Q(\hat{F}_n)$ does not make sense in this case. Instead, we take

$$Q_n(\hat{F}_n) := \frac{\hat{F}_n(x+h_n) - \hat{F}_n(x-h_n)}{2h_n},$$

where the *bandwidth* h_n is "small." The choice of h_n can theoretically be based on a trade-off between bias and variance.

Maximum Likelihood

Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a dominated family, i.e., there exists a σ-finite dominating measure ν, such that the densities

$$p_\theta := \frac{dP_\theta}{d\nu}, \qquad \theta \in \Theta$$

exist. The *maximum likelihood estimator* (MLE) $\hat{\theta}_n$ of θ is then defined as

$$\hat{\theta}_n := \arg\max_{\vartheta \in \Theta} \sum_{i=1}^n p_\vartheta(X_i),$$

where "arg" is "argument," i.e, the location where the maximum is achieved. It is to be checked in particular cases that the maximum actually exists.

Example 5. Suppose the densities w.r.t. Lebesgue measure are

$$p_\theta(x) = \theta(1+x)^{-\theta+1}, \ x > 0,$$

where $\theta \in \Theta = (0, \infty)$. Then

$$\log p_\vartheta(x) = \log \vartheta - (\vartheta + 1)\log(1+x).$$

We denote the derivative w.r.t. ϑ by

$$s_\vartheta(x) := \frac{d}{d\vartheta} \log p_\vartheta(x) = \frac{1}{\vartheta} - \log(1+x).$$

We put the derivative $\sum_{i=1}^n s_\vartheta(X_i)$ to zero:

$$\frac{n}{\hat{\theta}_n} - \sum_{i=1}^n \log(1+X_i) = 0.$$

This gives $\hat{\theta}_n = 1/\left[\sum_{i=1}^n \log(1+X_i)/n\right]$.

When Θ is finite-dimensional, say $\Theta \subset \mathbb{R}^p$, then under general regularity conditions, the MLE is approximately unbiased and each component has the smallest asymptotic variance. In fact, then $\sqrt{n}(\hat{\theta}_n - \theta)$ is approximately normally distributed, with mean zero and covariance matrix $I(\theta)^{-1}$, where $I(\theta)$ is the $p \times p$ Fisher information matrix

$$I(\theta) = E_\theta s_\theta(X) s_\theta^T(X),$$

with s_θ the *score function*

$$s_\theta := \frac{\partial}{\partial \theta} \log p_\theta.$$

Moreover, $I(\theta)$ can be approximated by the matrix of second derivatives

$$-\frac{\partial^2}{\partial \vartheta \partial \vartheta^T} \sum_{i=1}^n \log p_\vartheta(X_i)/n \Big|_{\vartheta = \hat{\theta}_n}.$$

These results however are not for free as they do rely on regularity conditions.

The MLE can be seen as plug in estimator:

$$\hat{\theta}_n = \arg\max_\vartheta \int \log p_\vartheta d\hat{P}_n,$$

and

$$\theta = \arg\max_\vartheta \int \log p_\vartheta dP_\theta.$$

M-Estimators

An M-estimator is of the form

$$\hat{\gamma}_n := \arg\min_{c \in \Gamma} \frac{1}{n} \sum_{i=1}^n \rho_c(X_i).$$

Here, for each $c \in \Gamma$, $\rho_c : \mathcal{X} \to \mathbb{R}$, $c \in \Gamma$ is a given function, called a *loss function* (generally to be distinguished from the loss function considered in Section ►"How to Construct Estimators?"). The M-estimator targets at the parameter

$$\gamma := \arg\min_{c \in \Gamma} E\rho_c(X).$$

This is indicated by the plug-in principle: the mean $E\rho_c(X)$ is approximated by the average $\sum_{i=1}^n \rho_c(X_i)/n$.

Example 6. Let $\mathcal{X} = \mathbb{R}$ and also $\Gamma = \mathbb{R}$. The sample mean $\hat{\mu} = \sum_{i=1}^n X_i/n$ is an M-estimator with $\rho_c(x) = (x-c)^2$, the squared error. In other words, the sample mean is the least squares estimator. In the same spirit, the sample median (the middle observation) can be seen as an M-estimator with $\rho_c(x) = |x-c|$. More generally, the empirical $(1-\alpha)$-quantile $\hat{F}_n^{-1}(1-\alpha)$ is an M-estimator with

$$\rho_c(x) = \rho(x-c), \ \rho(x) = \alpha|x|\{x > 0\} + (1-\alpha)|x|\{x < 0\}.$$

Example 7. The MLE is an M-estimator, with $\Gamma = \Theta$, and

$$\rho_\vartheta = -\log p_\vartheta.$$

Non-Parametric and High-Dimensional Problems

The non-parametric case is the situation where the parameter space Θ is infinite-dimensional. Closely related is the case where $\Theta \subset \mathbb{R}^p$ with $p > n$ (or $p \gg n$). Construction of estimators in high-dimensional or nonparametric models often - but not always - requires some complexity regularization. This is because a famous statistical rule-of-thumb, which says that the number of parameters should not exceed the number of observations, is violated.

Example 8. *Consider a linear regression model* (see ▶Linear Regression Models)

$$Y = \sum_{j=1}^{p} Z_j \beta_j + \epsilon,$$

where Y is a real-valued response variable, $Z_1, \ldots, Z_p$ are p co-variables, and where ϵ is mean-zero measurement error. The coefficients $\beta = (\beta_1, \ldots, \beta_p)^T$ form a p-dimensional unknown parameter. Let $Z = (Z_1, \ldots, Z_p)$ be the p-dimensional co-variable. Suppose we observe n independent copies $\{(Y_i, Z_i)\}_{i=1}^n$ of (Y, Z). Ideally, we would like $n \gg p$, in order to estimate the p parameters. If however $p \geq n$, one generally needs a regularization penalty. For example, when one believes that only a few of the variables are relevant, one may use the so-called Lasso

$$\hat{\beta}_n = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} |Y_i - (Z\beta)_i|^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|,$$

where $\lambda_n > 0$ is a tuning parameter.

Note that we used the standard quadratic loss function here. Other choices may also be applied, for instance quantile regression:

$$\hat{\beta}_n = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i - (Z\beta)_i) + \lambda_n \sum_{j=1}^{p} |\beta_j|,$$

where, as in Example 6

$$\rho(x) = \alpha |x| \{x > 0\} + (1 - \alpha)|x| \{x < 0\}.$$

Example 9. *The estimator*

$$\hat{f}_n(x) := \frac{\hat{F}_n(x + h_n) - \hat{F}_n(x - h_n)}{2h_n},$$

considered in Example 4 is a nonparametric estimator of the density f at the point x. In this case, the bandwidth h_n is the tuning parameter. More generally, one may apply kernel estimators

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),$$

where K is a given kernel. Another possibility is to use a regularized M-estimator. For example, when $\mathcal{X} = [0,1]$, one may consider using

$$\hat{f}_n = \arg\min_{\mathrm{f}} \left\{ \int_0^1 \mathrm{f}^2(x)dx - 2\sum_{i=1}^{n} \mathrm{f}(X_i)/n + \lambda_n \int_0^1 |\ddot{\mathrm{f}}(x)|^2 dx \right\},$$

where λ_n is again a tuning parameter.

Regularization is not always necessary, it can often be replaced by qualitative constraints such as monotonicity or convexity.

Example 10. *Suppose X has density f with respect to Lebesgue measure, and suppose we know that f is increasing. Let $\hat{f}_n$ be the Grenander estimator, i.e., the maximum likelihood estimator*

$$\hat{f}_n = \arg\max_{\mathrm{f\ increasing}} \frac{1}{n} \sum_{i=1}^{n} \log \mathrm{f}(X_i).$$

Then under moderate conditions, $\hat{f}_n$ is a good approximation of f for large sample size n.

Further Reading

Estimation is a very broad area is statistics. Classical books covering this area are Bickel and Doksum (2001) and Rice (1994), and for Bayesian methods, Berger (1985). The details for the asymptotic approach (also non- and semi-parametric) are in van der Vaart (2000). In van de Geer (2000) one finds a treatment of M-estimators, mostly in a nonparametric context. One can read more about density estimation in Wand and Jones (1995), and Silverman (1986). The Lasso is explained in Hastie et al. (2001), and the estimator of a monotone density is handled in Grenander (1981).

About the Author

Dr. Sara Ana van de Geer obtained her PhD in 1987 under supervision of Prof. R.D. Gill and Prof. W.R. van Zwet. She is Full Professor at the ETH Zürich (since 2005). She was Chair of Probability and Mathematical Statistics, Leiden University (1999–2005), and currently is Chair of the Seminar for Statistics, ETH Zürich. Dr. van de Geer is Elected Fellow of the Institute of Mathematical Statistics, Elected member of the International Statistical Institute, and Correspondent of the Dutch Royal Academy of Sciences. She was Associate Editor of *Bernoulli* (2004–2008), *Annals of Statistics* (1997–2007), *Statistica Sinica* (2005–2008), and *Statistica Neerlandica* (1996–2000). Currently, she is Editor

of Collections Volume (Springer, 2010) and Associate Editor of *The Scandinavian Journal of Statistics* (2010–). Professor van de Geer has (co-)authored over 40 papers and three books, including *Empirical Processes in M-Estimation* (Cambridge University Press, 2000). She was awarded the Medallion Lecturer at the Joint Statistical Meetings, San Francisco (August 2003) and the International Statistical Institute Award.

Cross References

►Absolute Penalty Estimation
►Advantages of Bayesian Structuring: Estimating Ranks and Histograms
►Approximations for Densities of Sufficient Estimators
►Asymptotic Relative Efficiency in Estimation
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Best Linear Unbiased Estimation in Linear Models
►Complier-Average Causal Effect (CACE) Estimation
►Decision Trees for the Teaching of Statistical Estimation
►Estimation
►Estimation Problems for Random Fields
►Hazard Ratio Estimator
►Hodges-Lehmann Estimators
►Horvitz–Thompson Estimator
►James-Stein Estimator
►Kaplan-Meier Estimator
►Maximum Entropy Method for Estimation of Missing Data
►Methods of Moments Estimation
►Nonparametric Density Estimation
►Nonparametric Estimation
►Nonparametric Estimation Based on Incomplete Observations
►Optimal Designs for Estimating Slopes
►Optimal Shrinkage Estimation
►Optimal Shrinkage Preliminary Test Estimation
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Properties of Estimators
►Robust Regression Estimation in Generalized Linear Models
►Small Area Estimation
►Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts
►Statistical Inference
►Statistical Inference: An Overview
►Target Estimation: A New Approach to Parametric Estimation
►Trend Estimation
►Unbiased Estimators and Their Applications

References and Further Reading

Berger JO (1985) Statistical decision theory and Bayesian analysis. Springer, New York
Bickel PJ, Doksum KA (2001) Mathematical statistics, vol 1. Prentice Hall, Upper Saddle River, NJ
Grenander U (1981) Abstract inference. Wiley, New York
Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, New York
Rice JA (1994) Mathematical statistics and data analysis. Duxbury Press, Belmont, CA
Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall/CRC, Boca Raton, FL
van de Geer S (2000) Empirical processes in M-estimation. Cambridge University Press, Cambridge, UK
van der Vaart AW (2000) Asymptotic statistics. Cambridge University Press, Cambridge, UK
Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall/CRC, Boca Raton, FL

Eurostat

Walter J. Radermacher
Chief Statistician of the European Union and
Director-General of Eurostat
Luxembourg

Eurostat is the Statistical Office of the European Union. Established in 1953 and based in Luxembourg, it is part of the European Commission. Its role is to provide the European Union with high-quality statistics that enable comparisons between countries and regions.

Responsibilities

Eurostat has two main areas of responsibility. Firstly, in close cooperation with the national statistical authorities of the EU Member States, it develops harmonized methods for the production of statistics.

Secondly, Eurostat calculates and publishes total figures, aggregates, for the European Union and the euro area. As Eurostat does not conduct its own surveys in EU Member States, statistics from specific areas are transmitted to Eurostat by individual countries.

The methodology applied by EU Member States in the data collection as well as the delivery deadlines for transmission to Eurostat are regulated to a large degree by EU legislation.

Eurostat's statistics enable straightforward comparisons between European countries and regions and are crucial for politicians, economists and business people active at European, country and local government levels.

Increasingly, statistics made available by Eurostat are also consulted by members of the European public and media.

The Eurostat website gives access to a huge selection of statistical information. Over 300 million figures are available in its data base, around 4,500 Eurostat publications; many of them in English, German and French, and 1,000 tables can be accessed online and free of charge.

Internal Organization

Eurostat is part of the European Commission. It is headed by Director-General Walter Radermacher, the Chief Statistician of the European Union. Eurostat reports to the European Commissioner for Economic and Monetary Affairs.

In 2009, Eurostat had 900 staff. It is divided into seven directorates. They are:

- Cooperation in the European Statistical System; Resources
- Quality, methodology and information systems
- National and European Accounts
- External cooperation, communication and key indicators
- Sectoral and regional statistics (agriculture, fisheries, environment, transport, energy and regional statistics)
- Social and information society statistics (population, labor market, living conditions and social protection, education, science, culture, health and food safety, crime, information society and tourism statistics)
- Business statistics (business statistics, international trade and price statistics)

European Statistical System

Eurostat does not work alone. Together with the national statistical authorities of the EU-27 Member States (national statistical institutes, or NSIs, ministries, central banks etc.) it forms the European Statistical System (ESS). Its legal basis is the Regulation on European Statistics which came into force in April 2009, replacing older legislation.

The main decision-making body of the ESS is the European Statistical System Committee (ESSC), chaired by the Director-General of Eurostat. The ESSC is made up of Heads of the NSIs. It develops the ESS strategy, in accordance with the so-called European Statistics Code of Practice from 2005.

In 2008, the European Statistical Governance Advisory Board, or ESGAB, was created. Its role is to enhance the professional independence, integrity and accountability of the European Statistical System as well as to improve the quality of European statistics.

Also created in 2008, the European Statistical Advisory Committee (ESAC) consists of 24 representatives of institutional and private statistics users, who give their opinion on various statistical projects developed by the ESS.

Short History

The origins of Eurostat go back to the early 1950s, when a "statistical service" of the Coal and Steel Community was created in 1952, charged with conducting "a continuous study of market and price trends."

Over the next few years the responsibilities of the Statistical Division, renamed in 1954, diversified and, when the European Community was founded in 1958, it became a separate Directorate-General of the European Commission. Eurostat received its present name in 1959.

Today, Eurostat's key role is to supply the EU institutions and policy makers as well as the research community and all European citizens with a wide range of statistics, instrumental for planning, development and implementation of joint European Community policies.

For More Information

Eurostat website: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home

About the Author

Walter Radermacher was born in Walheim (near Aachen) on 10 June 1952. From 1970 to 1975 he studied business administration in Aachen and in Münster. After obtaining the academic degree of Diplom-Kaufmann (economist), he worked as member of academic staff at the University of Münster from 1975 to 1977. At the beginning of 1978, Walter Radermacher joined the Federal Statistical Office. Mr Radermacher has held a wide variety of posts at the German Statistical Office during his thirty-year career there, notably in the fields of environmental and economic statistics. He has been Vice-President of the German Federal Statistical Office (2003–2006), and President since 2006. During the 2007 German Presidency, he was Chair of the Council's working group on statistics. He has also chaired the UN Committee on Environmental–Economic Accounting. During his career, Mr. Radermacher has had teaching assignments in statistics and environmental economy at the specialised college of higher education (Fachhochschule) in Wiesbaden and the University of Lüneburg in Germany. As of 2008, he was appointed Director General of Eurostat and chief statistician of the European Union.

Cross References

►Business Surveys
►Comparability of Statistics
►Integrated Statistical Databases
►National Account Statistics
►Statistical Publications, History of

Event History Analysis

Hans-Peter Blossfeld
Professor
Otto-Friedrich-Universität Bamberg, Bamberg, Germany

Introduction

Event history analysis studies a collection of units, each moving among a finite (usually small) number of states. An example is individuals who are moving from unemployment to employment. In this article various examples of event history applications are presented, the concepts and advantages of event history analysis are demonstrated, and practical complications are addressed.

Examples of Continuous-Time, Discrete-State Processes

Event history analysis is the study of processes that are characterized in the following general way: (1) there is a collection of units (which may be individuals, organizations, societies, etc.), each moving among a *finite* (usually small) *number of states*; (2) these changes (or *events*) may *occur at any point in time* (i.e., they are not restricted to predetermined points in time); and (3) there are *time-constant and/or time-varying factors* influencing the timing of events. Examples are workers who move between unemployment and employment; men and women who enter into consensual unions or marriages; companies that are founded or closed down; governments that break down; people who are mobile between different regions or nation states; consumers who switch from one brand to another; prisoners who are released and commit another crime; people who show certain types of verbal and non-verbal behaviors in social interactions; students who drop out of school; incidences of racial and ethnic confrontation, protest, riot, and attack; people who show signs of psychoses or neuroses; patients who switch between the states "healthy" and "diseased," and so on. In event history analysis, the central characteristics of the underlying stochastic process are mirrored in the specific way theoretical and mathematical models are built, data are collected, and the estimation and evaluation of models is done.

Different Observation Plans

In event history analysis the special importance of broader research design issues has been stressed. In particular, different observation plans have been used to collect information on the *continuous-time, discrete-state substantive process*. These various schemes produce different types of data that constrain the statistical analysis in different ways: cross-sectional data, event count data, event sequence data, ►panel data, and event history data.

The *cross-sectional observation design* is still the most common form of data collection in many fields. A cross-sectional sample is only a "snapshot" taken on the continuous-time, discrete-state substantive process (e.g., the jobs of people at the time of the interview). Event history analysis has shown that one must be very cautious in drawing inferences about explanatory variables on the basis of such data because, implicitly or explicitly, social researchers have to assume that the process under study is in some kind of *equilibrium*. Equilibrium means that the *state probabilities* are fairly trendless and the relationships among the variables are quite stable over time. Therefore, an equilibrium of the process requires that the inflows to and the outflows from each of the discrete states be equal over time to a large extent. This is a strong assumption that is not very often justified in the social and behavioral sciences (see Blossfeld and Rohwer 2002; Blossfeld et al. 2007).

A comparatively rare type of data on changes of the process are *event count data*. They record the number of different types of events for each unit in an interval (e.g., the number of racial and ethnic confrontations, protests, riots, or attacks in a period of 10 years). *Event sequence data* provide even more information. They record the sequence of specific states occupied by each unit over time. Today, the temporal data most often available to the scientist are *panel data*. They provide information on the same units of analysis at a series of discrete points in time. In other words, there is only information on the states of the units at predetermined survey points and the course of the process between the survey points remains unknown. Panel data certainly contain more information than cross-sectional data, but involve well-known distortions created by the method itself (e.g., panel bias, attrition of sample). In addition, causal inferences based on panel approaches are much more complicated than has been generally acknowledged (Blossfeld and Rohwer 2002; Blossfeld et al. 2007). With regard to the continuous-time, discrete-state substantive process, panel analysis is particularly sensitive

to the length of the time intervals between the waves relative to the speed of the process. They can be too short, so that very few state transitions will be observed, or too long, so that it is difficult to establish a time-order between the events.

The *event oriented observation design* records the complete sequence of states occupied by each unit and the timing of changes among these states. For example, an event history of job mobility consists of more or less detailed information about each of the jobs and the exact beginning and ending dates of each job. Thus this is the most complete information one can get on the continuous-time, discrete-state substantive process. Such *event history data* are often collected retrospectively with *life history studies*. Life history studies are normally cheaper than panel studies and have the advantage that they code the data into one framework of codes and meaning. But retrospective studies also suffer from several limitations (see Blossfeld and Rohwer 2002; Blossfeld et al. 2007). In particular, data concerning motivational, attitudinal, cognitive, or affective states are difficult (or even impossible) to collect retrospectively because the respondents can hardly recall the timing of changes in these states accurately. Also the retrospective collection of behavioral data has a high potential for bias because of its strong reliance on autobiographic memory. To reduce these problems of data collection, modern panel studies (e.g., the PSID in the US, the BHPS in the UK, or the SOEP in Germany) use a mixed design that provides traditional panel data and retrospectively collects event history data for the period before the first panel wave and between the successive panel waves.

Event History Analysis

Event history analysis, originally developed as independent applications of mathematical probability theory in ►demography (e.g., the *classical life table analysis* and the *product-limit estimator*), reliability engineering, and ►biostatistics (e.g., the path-breaking *semiparametric regression* model for ►*survival data*), has gained increasing importance in the social and behavioral sciences (see Tuma and Hannan 1984; Blossfeld et al. 1989; Lancaster 1990; Blossfeld and Rohwer 2002; Blossfeld et al. 2007) since the 1980s. Due to the development and application of these methods in various disciplines, the terminology is normally not easily accessible to the user. Central definitions are therefore given first.

Basic Terminology

Event history analysis studies *transitions* across a set of discrete states, including the length of *time intervals* between entry to and exit from specific states. The basic analytical framework is a state space and a time axis. The choice of the *time axis* or *clock* (e.g., age, experience, and marriage duration) used in the analysis must be based on theoretical considerations and affects the statistical model. Dependent on practical and theoretical reasons, there are event history methods using a *discrete-* (Yamaguchi 1991; Vermunt 1997) or *continuous-time axis* (Tuma and Hannan 1984; Blossfeld et al. 1989; Lancaster 1990; Courgeau and Lelièvre 1992; Blossfeld and Rohwer 2002; Blossfeld et al. 2007). Discrete-time event history models are, however, special cases of continuous-time models, and are therefore not further discussed here. *An episode, spell, waiting time, or duration* – terms that are used interchangeably in the literature – is the time span a unit of analysis (e.g., an individual) spends in a specific state. The *states are discrete* and usually small in number. The definition of a set of possible states, called the *state space*, is also dependent on substantive considerations. Thus, a careful, theoretically driven choice of the time axis and design of state space are crucial because they are often serious sources of misspecification.

The most restricted event history model is based on a process with only a *single episode* and two *states* (an *origin state* j and a *destination state* k). An example may be the duration of first marriage until the end of the marriage, for whatever reason (separation, divorce, or death). In the *single episode* case each unit of analysis that entered into the origin state is represented by one episode. Event history models are called *multistate models*, if more than one destination state exists. Models for the special case with a single origin state but two or more destination states are also called *models with competing events or risks*. For example, a housewife might become "unemployed" (meaning entering into the state "looking for work"), or start being "full-time" or "part-time employed." If more than one event is possible (i.e., if there are repeated events or transitions over the observation period), the term *multi-episode modes* is used. For example, an employment career normally consists of a series of job shifts.

Censoring

Observations of event histories are often *censored*. Censoring occurs when the information about the duration in the origin state is incompletely recorded. An episode is *fully censored on the left*, if starting and ending times of a spell are located before the beginning of an observation period (e.g., before the first panel wave) and it is *partially censored on the left*, if the length of time a unit has already spent in the origin state is unknown. This is typically the case in panel studies, if individuals' job episodes at the time of a first panel wave are known but no further information about the history of the job is collected. Left censoring is

normally a difficult problem because it is not possible to take the effects of the unknown episodes into account. It is only without problems if the assumption of a *Markov process* (see ►Markov Processes) is justified (i.e., if the transition rates do not depend on the duration in the origin state).

The usual kind of censoring, however, is *right censoring*. In this case the end of the episode is not observed but the observation of the episode is terminated at an arbitrary point in time. This type of censoring typically occurs in life course studies at the time of the retrospective interviews or in panel studies at the time of the last panel wave. Because the timing of the end of the observation window is normally determined independently from the substantive process under study, this type of right censoring is unproblematic and can easily be handled with event history methods. Finally, episodes might be *completely censored on the right*. In other words, entry into and exit from the duration occurs after the observation period. This type of censoring normally happens in retrospective life history studies in which individuals of various birth cohorts can only be observed over very different spans of life. To avoid sample selection bias, such models have to take into account variables controlling for the selection, for example, by including birth cohort dummy variables (see Yamaguchi 1991).

The Transition Rate

The central concept of event history analysis is the *transition rate*. Because of the various origins of event history analysis in the different disciplines, the transition rate is also called the *hazard rate, intensity rate, failure rate, transition intensity, risk function, or mortality rate*:

$$r(t) = \lim_{t' \to t} Pr\left(t \le T < t' | T \ge t\right) / \left(t' - t\right).$$

The transition rate provides a local, time-related description of how the process evolves over time. It can be interpreted as the propensity (or *intensity*) to change from an origin state to a destination state, at time t. But one should note that this propensity is defined in relation to a *risk set* $(T \ge t)$ at t, i.e., the set of units that still can experience the event because they have not yet had the event before t.

Statistical Models

The central idea in event history analysis is to make the transition rate, which describes a process evolving in time, dependent on time (t) and on a set of covariates, x:

$$r(t) = g(t, x).$$

The causal interpretation of the transition rate requires that we take the temporal order in which the processes evolve very seriously. In other words, at any given point in time t, the transition rate $r(t)$ can be made dependent on conditions that happened to occur in the past (i.e., before t), but not on what is the case at t or in the future after t.

There are several possibilities to specify the functional relationship $g(.)$ (see Blossfeld and Rohwer 2002; Blossfeld et al. 2007). For example, the *exponential model*, which normally serves a baseline model, assumes that the transition rate can vary with different constellations of covariates x, but that the rates are time constant. The *piecewise-constant exponential model* allows the transition rate to vary across fixed time periods with period-specific effects of covariates. There are also different parametric models that are based on specific shapes of the time dependence (e.g., the *Gompertz-Makeham*, *Weibull*, *Sickle*, *Log-logistic*, *Log-normal*, or *Gamma model*). If the time shape of the *baseline hazard rate* is unspecified, and only possible effects of covariates x are estimated, the model is called a *semi-parametric or partial likelihood model*. Finally, there are also *models of unobserved heterogeneity* in which the transition rate can be made dependent on the observed covariates x, the duration t, and a stochastic error term.

Parallel and Interdependent Processes

The most important scientific progress permitted by event history analysis is based on the opportunity to include explicitly measured *time-varying covariates* in transition rate models (see Blossfeld and Rohwer 2002; Blossfeld et al. 2007). These covariates can change their values over process time in an analysis. Time-varying covariates can be qualitative or quantitative, and may stay constant for finite periods of time or change continuously. From a substantive point of view, time-varying covariates can be conceptualized as observations of the sample path of parallel processes. These processes can operate at different levels. In sociology, for example, (1) there can be parallel processes at the level of the individual's different domains of life (e.g., one may ask how upward and downward moves in an individual's job career influence his/her family trajectory), (2) there may be parallel processes at the level of some few individuals interacting with each other (e.g., one might study the effect of the career of the husband on his wife's labor force participation), (3) there may be parallel processes at the intermediate level (e.g., one can analyze how organizational growth influences career advancement or changing household structure determines women's labor force participation), (4) there may be parallel processes at the macro level (e.g., one may be interested in the effect of changes in the business cycle on family formation or career

advancement), and (5) there may be any combination of such processes of type (1) to (4). For example, in the study of life course, cohort, and period effects, time-dependent covariates at different levels (see below) must be included simultaneously (►*multilevel analysis*).

In dealing with such systems of parallel processes, the issue of *reverse causation* is often addressed in the methodological literature (see, e.g., Tuma and Hannan 1984; Blossfeld et al. 1989; Yamaguchi 1991). Reverse causation refers to the (direct or indirect) effect of the dependent process on the independent covariate process(es). Reverse causation is often seen as a problem because the effect of a time-dependent covariate on the transition rate is confounded with a feedback effect of the dependent process on the values of the time-dependent covariate. However, Blossfeld and Rohwer (2002) have developed a *causal approach* to the analysis of interdependent processes that also works in the case of interdependence. For example, if two interdependent processes, $Y_t{}^A$ and $Y_t{}^B$, are given, a change in $Y_t{}^A$ at any (specific) point in time t' may be modeled as being dependent on the history of both processes up to, but not including t'. Or stated in another way: What happens with $Y_t{}^A$ at any point in time t' is conditionally independent of what happens with $Y_t{}^B$ at t', conditional on the history of the joint process $Y_t = \left(Y_t{}^A, Y_t{}^B\right)$ up to, but not including, t ("*principle of conditional independence*"). Of course, the same reasoning can be applied if one focuses on Y_t^B instead of Y_t^A as the "dependent variable." Beginning with a transition rate model for the joint process, $Y_t = \left(Y_t^A, Y_t^B\right)$, and assuming the principle of conditional independence, the likelihood for this model can then be factorized into a product of the likelihood's for two separate models: a transition rate model for $Y_t{}^A$, which is dependent on $Y_t{}^B$ as a time-dependent covariate, and a transition rate model for $Y_t{}^B$, which is dependent on $Y_t{}^A$ as a time-dependent covariate. From a technical point of view, there is therefore no need to distinguish between *defined, ancillary, and internal covariates* because all of these time-varying covariate types can be treated in the estimation procedure. Estimating the effects of time-varying processes on the transition rate can easily be achieved by applying the *method of episode splitting* (see Blossfeld and Rohwer 2002; Blossfeld et al. 2007).

Unobserved Heterogeneity

Unfortunately, researchers are not always able to include all the important covariates into an event history analysis. These unobserved differences between subpopulations can lead to *apparent time dependence* at the population level and additional identification problems. There exists a fairly large literature on misspecified models in general. In particular, there have been several proposals to deal with *unobserved heterogeneity* in transition rate models (so-called *frailty models*) (see, e.g., Blossfeld and Rohwer 2002; Blossfeld et al. 2007).

About the Author

Dr. Hans-Peter Blossfeld is Professor of Sociology at Bamberg University. He is also Director of the Institute of Longitudinal Studies in Education (Institut für bildungswissenschaftliche Längsschnittstudien – INBIL), Principal Investigator of the National Educational Panel Study (NEPS), and Director of the State Institute for Family Research at Bamberg University (Staatsinstitut für Familienforschung - ifb). He is an Elected member of the German Academy of Sciences (Leopoldina), the Bavarian Academy of Sciences and the Berlin-Brandenburg Academy of Sciences. He is also a Elected member of the European Academy of Sociology, London, and Elected Chairman of the European Consortium of Sociological Research (ECSR). He serves as member of the Steering Committees of the European Science Foundation (ESF) Programes "Quantitative Methods in the Social Sciences II" and "TransEurope". He is also member of the ESRC National Strategy Committee on Longitudinal Studies in the UK. He has authored and co-authored more than 200 papers, mostly in refereed journals, and has written as well as edited 22 books, including *Techniques of Event History Modeling: New Approaches to Causal Analysis* (with G. Rohwer, Lawrence Erlbaum Associates, 2002). Professor Blossfeld has received the Descartes award for his Globalife project (European Commission, 2007). He is Editor of the following journals: *European Sociological Review*, *Zeitschrift für Erziehungswissenschaft*, and *Zeitschrift für Familienforschung*. Currently, he is also an Associate editor of *International Sociology*.

Cross References

►Censoring Methodology
►Demography
►Frailty Model
►Survival Data

References and Further Reading

Blossfeld H-P, Hamerle A, Mayer KU (1989) Event history analysis. Lawrence Erlbaum Associates, Hillsdale, NJ

Blossfeld H-P, Rohwer G (2002) Techniques of event history modeling. New approaches to causal analysis. Lawrence Erlbaum Associates, Mahwah, NJ/London

Blossfeld H-P, Golsch K, Rohwer G (2007) Event history analysis with stata. Lawrence Erlbaum Associates, Mahwah, NJ/London

Courgeau D, Lelièvre E (1992) Event history analysis in demography. Clarendon, Oxford

Lancaster T (1990) The econometric analysis of transition data. Cambridge University Press, Cambridge
Tuma NB, Hannan MT (1984) Social dynamics. Models and methods. Academic, New York
Vermunt JK (1997) Log-linear models for event histories. Sage, Newbury Park, CA
Yamaguchi K (1991) Event history analysis. Sage, Newbury Park, CA

Exact Goodness-of-Fit Tests Based on Sufficiency

Federico J. O'Reilly, Leticia Gracia-Medrano
Professors
IIMAS, UNAM, México City, México

Introduction and Simple Case

In the univariate continuous goodness-of-fit problem, the probability integral transformation (PIT) is used to transform the sample observations into the unit interval when the distribution to be tested is completely specified (the simple case); thus reducing the problem to that of testing uniformity of the transforms. The asymptotic theory for the induced empirical process in (0,1) is well known and so, that of functionals of it, including the so called EDF tests that are based on the Empirical Distribution Function (see ►Tests of fit based on the empirical distribution function).

The empirical process (see ►Empirical Processes), labeled with the original x is the process

$$\psi_n(x) = \sqrt{n}\{F_n(x) - F_0(x)\},$$

for $x \in R$, with F_n being the empirical distribution function constructed from the observed sample $x_1, \cdots, x_n$ and with F_0 the known distribution function to be tested. This process is in a 1:1 correspondence with the corresponding process labeled with $u \in (0,1)$,

$$\eta_n(u) = \sqrt{n}\{G_n(u) - u\},$$

with $u = F_0(x)$, where G_n stands for the empirical distribution function of the transformed sample $u_1, \cdots, u_n$ with the u_i's given by $u_i = F_0(x_i)$.

Functionals of the empirical process, that appeared in literature (mostly) for the continuous goodness-of-fit problem are Kolmogorov's statistic, D_n, Cramér-von Mises, $W^2{}_n$, Anderson-Darling's, $A^2{}_n$ and others like Kuiper's V_n, Watson's $U^2{}_n$, and Pearson's chi square, χ^2.

These may be described in the $(0,1)$ transformed range as:

$$D_n = sup_u|\eta_n(u)|$$
$$W^2{}_n = \int_0^1 (\eta_n(u))^2 du$$
$$A^2{}_n = \int_0^1 (\eta_n(u))^2 \frac{1}{u(1-u)} du,$$

having well known computable formulas in terms of the $u_{(i)}$'s, the now ordered transformed u_i's, from smallest to largest, with $u_{(0)} = 0$ and $u_{(n+1)} = 1$, if needed.

$$D_n = max\{D^+{}_n, D^-{}_n\},$$
$$D^+{}_n = max_i\left\{G_n(u_{(i)}) - u_{(i)}\right\},$$
$$D^-{}_n = max_i\{u_{(i)} - G_n(u_{(i-1)})\}, G_n(u_{(i)}) = i/n,$$
$$W^2{}_n = \sum_{i=1}^{n}[u_{(i)} - (2i-1)/(2n)]^2 + 1/(12n),$$
$$A^2{}_n = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[log(u_{(i)}) + log(1 - u_{(n+1-i)})],$$

with similar expressions for the other EDF tests (like Kuiper's $V_n = D^+{}_n + D^-{}_n$).

If the continuous distribution to be tested, has unknown parameters, the established procedure (using asymptotics) is to estimate first and substitute them in place of the parameters before employing the PIT, to transform to the unit interval. Under quite general conditions, asymptotics for the resulting empirical process in $(0,1)$ have been reported in literature for many distributions (see Durbin 1973; D'Agostino and Stephens 1986 and references given there).

If the distribution is discrete, the above approach using asymptotics, has not been followed much. To start, in the simple case, the PIT does not yield uniform transforms. There are however several proposed tests for the discrete case; some based on the empirical characteristic function, others on the empirical probability generating function and others. See for example Kocherlakota and Kocherlakota (1986), Rueda et al. (1991), Nakamura and Pérez-Abreu (1993), Rueda and O'Reilly (1999), Gurtler and Henze (2000) and references therein. In Spinelli and Stephens (1997) discrete versions of EDF tests are studied. All these procedures, yield tests related to the "observed and expected counts." Further mention of the discrete case is made only for "on site" evaluation of ►*p*-values.

With fast computers available now, the *simple* goodness-of-fit problem, either continuous or discrete, may be solved by computing the *p*-value of the selected statistic to test the fit, simply by simulating many independent samples from

the null hypothetical distribution. This is done by observing first the value of the test statistic with the initial sample and then looking the corresponding simulated values of the test statistic, then finding the proportion of simulated values exceeding the initial value of the test statistic. That is the p-value.

This procedure is easier than having to use tables, and possible finite-n corrections. Moreover, this simulation procedure applies to the continuous as well as the discrete case, because the inverse of the PIT (IPIT) is a proper generator for any distribution.

The IPIT maps a uniformly distributed random variable U with the "inverse" of a distribution function F^{-1} yielding a new random variable $X = F^{-1}(U)$, whose distribution function is precisely, F. With the inverse $F^{-1}(u)$, say, taken as the *supremum* of the x-values such that $F(x) \leq u$.

So in *any* simple goodness-of-fit problem (continuous or discrete), given

$$\underline{x} = (x_1, x_2, \cdots, x_n),$$

a sample of independent identically distributed random variables with common distribution function F, to test the null hypothesis,

$$H_0 : F = F_0,$$

with F_0 totally known, one proceeds as follows:

For $S_n(\underline{x}) = S_n(x_1, \cdots, x_n)$ the test statistic to be used for testing H_0, evaluated at the observed sample, simulate from the null distribution F_0, say 10,000 samples $(\underline{x}_j, j = 1, \cdots, 10{,}000)$ of the same size n and with each, compute $S_n(\underline{x}_j)$. Finally look at the proportion of simulated values (the $S_n(\underline{x}_j)$'s) exceeding $S_n(\underline{x})$. This is the p-value obtained with the 10,000 simulations.

Composite Case: Group Model

Many "composite" goodness-of-fit problems may be tackled with a strikingly similar Monte-Carlo simulation. Those problems arise when the null hypotheses corresponds to a parametric family with a group structure, where the test statistic S_n is invariant; typically, most families with location and/or scale parameters only. It was noticed that when finding the limiting distribution of the "empirical process with estimated parameters," the limiting process was a zero mean Gaussian process, whose covariance function did not depend on the parameter values (generally depends on the particular family to be tested). This result allowed the use of the limiting distribution for the EDF test statistic; perhaps using some finite-n correction (see e.g., Chap. 4 in D'Agostino and Stephens 1986). This is the case of most location/scale families.

Under a group model, not only the asymptotic, but the finite-n distributions of invariant tests S_n are independent of the value of the parameters, so one may simply assign any arbitrary value to them, and simulate tens of thousands of samples and repeat as in the simple case to evaluate an exact p-value in situ.

Composite Case: Non-Group Model

There are important parametric families where there is no group structure, and limiting distributions depend also on the parameter values. These include those where a "shape" parameter is present, as in the case of the ▶gamma distribution, the inverse-Gaussian and is the case of most discrete distributions with unknown parameters. For these, an attempt has been made to compute *not* the p-value, since in general that quantity depends on the parameter, but rather use the conditional distribution of the test statistic S_n given some suitable sufficient statistic to compute a *conditional* p-value. This is explained next.

Assume that in the setting $H_0 : F(\cdot) = F_0(\cdot; \theta)$, where $\theta \in \Theta$, T_n is the minimal sufficient statistic. Not being a group model the test statistic S_n will have a distribution, under H_0, that will in general depend on θ. But having T_n sufficient means that the conditional distribution of S_n given T_n, will not depend on θ, so in order to base the procedure in a simulation, it will be enough to have a simulator of samples which are *conditionally to* T_n, independent identically distributed. One way to simulate in that fashion appears in O'Reilly and Gracia-Medrano (2006), and is illustrated for the ▶inverse Gaussian distribution. In González-Barrios et al. (2006), the results provide means to simulate for the Poisson, binomial and negative binomial, with unknown parameters (besides the test derived there that uses enumeration for the conditional distribution). In Lockhart et al. (2007) a simulator is provided for the two parameter gamma distribution. The first reference bases the simulator on the explicit use of the Rao-Blackwell estimator of the distribution $F_0(\cdot; \theta)$. The third reference uses the Gibbs sampler to achieve an exact simulation method.

The inverse-Gaussian simulator is sketched to illustrate. Let the null hypothesis correspond to the inverse-Gaussian distribution, $F_0(x; \mu, \lambda)$. Denote the Rao-Blackwell distribution estimate, based on the sample $\underline{x}$, by $\check{F}_n(x)$, which is a function of the minimal sufficient statistic,

$$T_n(\underline{x}) = \left(\sum_i x_i, \sum_i \left(\frac{1}{x_i} \right) \right),$$

and exists for $n > 2$. Its expression appears (corrected) in Chhikara and Folks (1977).

Observe that the statistic T_n has the property of double transitivity, which means that the pair (T_n, X_n) is in a 1:1 correspondence with (T_{n-1}, X_n). It implies in particular that knowing the value of the statistic with a sample of size n and knowledge of, say, the last observation, one can find the value of the statistic with the corresponding smaller sample of size $(n-1)$.

After observing the sample, denote by s the value of the goodness of fit test and by t_n the value of the minimal sufficient statistic; that is, $S_n(\underline{x}) = s$ and $T_n(\underline{x}) = t_n$. In order to simulate, conditionally to $T_n = t_n$, a conditional sample (denoted a $*$-sample) proceed as follows:

Obtain randomly u_n from the $U(0,1)$ distribution and set $x^\star{}_n = \tilde{F}_n^{-1}(u_n)$, then compute $t^\star_{n-1}$ from the knowledge of t_n and taking away form its evaluation the (imposed) n-th value $x^\star{}_n$, i.e., obtain $t^\star_{n-1}$ from the pair $\left(t_n, x^\star{}_n\right)$. Next with an independently selected u_{n-1} from the $U(0,1)$, use the Rao-Blackwell estimate given $t^\star_{n-1}$ and define $x^\star{}_{n-1} = \tilde{F}_{n-1}^{-1}(u_{n-1})$. Continue finding $t^\star_{n-2}$ from $\left(t^\star_{n-1}, x^\star{}_{n-1}\right)$, then looking at the Rao-Blackwell estimate given $t^\star_{n-2}$, use its inverse to get $x^\star{}_{n-2}$ with an independently selected u_{n-2}; and so on until finding $x^\star{}_3$.

Finally, get $x^\star{}_2$ and $x^\star{}_1$ as the solution to yield the same value for T_n when evaluated at the new $*$-sample (that is t_n).

The procedure is repeated for tens of thousands of times getting tens of thousands of $*$-samples which in turn produces the desired simulated values of the statistic S_n. The proportion of simulated values that exceed the original s is the conditional p-value, and is exact. With 10,000 $*$-samples and $n = 20$, evaluation of the p-value takes around 13 s on a conventional PC.

Parametric Bootstrap

Simulations based on parametric bootstrap to get a p-value, provide an *exact* procedure in the case of a group model. In the non-group model, bootstrapping is not theoretically exact but very good. Ongoing research is being done on the closeness of the conditional p-value and the one found with bootstrap (see ▶Bootstrap Methods). Articles where bootstrap has been proposed to evaluate the p-value for particular distributions, show that for large n, the procedure provides a very good approximation.

About the Author

Professor O'Reilly is past President of the Mexican Statistical Association (1999–2000), and former Director of the "Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas" (IIMAS, 2000–2004). He is member of the Mexican Academy of Science and also a member of the ISI.

Cross References

▶Anderson-Darling Tests of Goodness-of-Fit
▶Bootstrap Methods
▶Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
▶Chi-Square Test: Analysis of Contingency Tables
▶Cramér-Von Mises Statistics for Discrete Distributions
▶Jarque-Bera Test
▶Kolmogorov-Smirnov Test
▶Tests of Fit Based on The Empirical Distribution Function

References and Further Reading

Chhikara RS, Folks JL (1977) The inverse Gaussian distribution as a lifetime model. Technometrics 19:461–468
D'Agostino RB, Stephens MA (eds) (1986) Goodness of fit techniques. Marcel Decker, New York
Durbin J (1973) Distribution theory for tests based on the sample distribution function. In: Regional Conference Series in Applied Mathematics, vol 9. SIAM, Philadelphia
González-Barrios JM, O'Reilly F, Rueda R (2006) Goodness of fit for discrete random variables using the conditional density. Metrika 64:77–94
Gurtler N, Henze N (2000) Recent and classical goodness of fit tests for the Poisson distribution. J Stat Plan Infer 90:207–225
Kocherlakota S, Kocherlakota K (1986) Goodness of fit tests for discrete distributions. Commun Stat Theor Meth 15:815–829
Lockhart RA, O'Reilly F, Stephens MA (2007) Use of the Gibbs sampler to obtain conditional tests, with applications. Biometrika 94:992–998
Nakamura M, Pérez-Abreu V (1993) Empirical probability generating function. An overview. Insur Math Econ 12:287–295
O'Reilly F, Gracia-Medrano L (2006) On the conditional distribution of goodness-of-fit-tests. Commun Stat Theor Meth 35:541–549
Rueda R, Pérez-Abreu V, O'Reilly F (1991) Goodness of fit for the Poisson distribution based on the probability generating function. Commun Stat Theor Meth 20(10):3093–3110
Rueda R, O'Reilly F (1999) Tests of fit for discrete distributions based on the probability generating function. Commun Stat Sim Comp 28(1):259–274
Spinelli J, Stephens MA (1997) Cramér-von Mises tests of fit for the Poisson distribution. Can J Stat 25(2):257–268

Exact Inference for Categorical Data

Roshini Sooriyarachchi
Professor
University of Colombo, Colombo, Sri Lanka

In practice statistical inference for categorical data is based mainly on large sample approximations of test statistics.

This asymptotic theory is valid only if the sample sizes are reasonably large and well balanced across populations. To make valid inferences in the presence of small, sparse or unbalanced data, exact p-values and confidence intervals, based on the permutational distribution of the test statistic (Good 1993) needs to be computed. The most common approach to exact inference for categorical data has been a conditional one (Agresti, 1992) thus this approach is mainly discussed here. This utilizes the distribution of the sufficient statistic for the parameter of interest, conditional on the sufficient statistics for the other model parameters (Agresti 2001). The sufficiency principle used here is explained in more detail in Mehta (1998).

Settings and Notation

The rest of the article presents a variety of exact methods for categorical data. The required settings and notation are defined in this section. Consider a two-way contingency table having I rows and J columns cross–classifying a row variable X and a column variable Y. Let $\{n_{ij}\}$ correspond to the cell counts $n_{i+} = \sum_j n_{ij}$, $n_{+j} = \sum_i n_{ij}$, and $n = \sum_{i,j} n_{ij}$. Agresti (1992) and Mehta (1998) list three sampling schemes (settings) that can give rise to cross-classified categorical data. These schemes are namely:

1. Full multinomial sampling scheme where the cell counts $\{n_{ij}\}$ have a ▶multinomial distribution generated by n independent trials with IJ cell probabilities $\{\pi_{ij}\}$.
2. Product multinomial sampling where counts $\{n_{ij}\}$ in row I have a multinomial distribution for all j, with counts in different rows being independent. Here $n_{i+} = \sum_j n_{ij}$, $i = 1, \ldots, I$ are fixed.
3. Poisson sampling, where $\{n_{ij}\}$ have a Poisson distribution with expected value $E(n_{ij}) = m_{ij}$ where $n = \sum_{i,j} n_{ij}$ is random.

As Agresti (1992) indicates that all three sampling models lead to the same inference, for the sake of simplicity the following results are based on sampling scheme (2). Under this scheme, the conditional probability of belonging to the jth column given that observation is in the ith row is denoted by $\pi_{j|i} = P(Y = j|X = i)$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

2 × 2 Table

Let π_1 and π_2 denote "success" probabilities associated with row 1 and row 2 respectively. Then the odds ratio is given by

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}.$$

Several authors (Mehta, 1998; Agresti, 1992, 2001) showed that

$$P(n_{11} = k \mid n, n_{1+}, n_{+1}; \theta) = \frac{\binom{n_{1+}}{k}\binom{n-n_{1+}}{n_{+1}-k}\theta^k}{\sum_u \binom{n_{1+}}{u}\binom{n-n_{1+}}{n_{+1}-u}\theta^u}. \quad (1)$$

Under the null hypothesis of independence ($\theta = 1$) the conditional distribution given in (1) is hypergeometric. To test $H_0 : \theta = 1$ versus $H_1 : \theta > 1$ the p-value is given by $P = \sum_S P(t \mid n, n_{1+}, n_{+1}; \theta = 1)$ where $S = \{t; t \geq n_{11}\}$. This test is called the ▶*Fisher's Exact test* (Fisher, 1935).

$I \times J$ Tables

For $I \times J$ tables, statistical independence of X and Y (H_0) corresponds to the log-linear model (Bishop et al., 1975)

$$\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y. \quad (2)$$

To test H_0, model (2) has to be compared with the saturated model

$$\log(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}. \quad (3)$$

Cornfield (1956) showed that the distribution of $\{n_{ij}\}$ given $\{n_{i+}\}$ and $\{n_{+j}\}$ depends only on the odds ratios (α_{ij}) such that

$$P(n_{ij} \mid \{n_{i+}\}, \{n_{+j}\}; \alpha_{ij}) = \frac{\prod_{i=1}^{I-1}\prod_{j=1}^{J-1}\alpha_{ij}^{n_{ij}}}{\prod_{i=1}^{I-1}\prod_{j=1}^{J-1} n_{ij}!} \ldots\ldots \quad (4)$$

where

$$\alpha_{ij} = \frac{\pi_{ij}/\pi_{IJ}}{\pi_{iJ}/\pi_{Ij}}.$$

Under H_0, $\alpha_{ij} = 1$ for all i and j thus the conditional distribution of n_{ij} given $\{n_{i+}\}$ and $\{n_{+j}\}$ under H_0 is multiple hypergeometric and is given by

$$P(n_{ij} \mid \{n_{i+}\}, \{n_{+j}\}) = \frac{\left(\prod_i n_{i+}!\right)\left(\prod_j n_{+j}!\right)}{n!\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!}.$$

The methods for $I \times J$ tables can be extended to multiway tables (Agresti, 1992).

Exact Inference for Logistic Models

Consider the case of a single binary response variable (Y) and several explanatory variables $X_1, \ldots, X_k$. In this situation a logistic model (Hosmer and Lemeshow (2000)) is preferred to a log-linear model. This model is expressed as $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{k}\beta_j x_{ij}$ where π_i is the probability of belonging to the research category of interest for the ith subject, $x_{i1}, \ldots, x_{ik}$ are the values of the k explanatory variables observed for the ith subject and $x_{i0} = 1$ for all i. When $\{y_i\}$ are independent Bernoulli outcomes, Cox (1970) showed that the sufficient statistics for β_j are $T_j = \sum_i y_i x_{ij}$; $j = 0, \ldots, k$ and illustrated how to conduct inference for β_j using the conditional distribution of T_j given $\{T_i, i \neq j\}$.

Other Topics and Available Software

Agresti (1992) discusses less common topics such as exact confidence intervals, exact goodness of fit tests, controversies regarding the exact conditional approach, exact unconditional approach and Bayesian approach. Software support for exact inference for categorical data are available in SAS, Stata, R, SPSS, StatXact and LogXact.

About the Author

Dr Sooriyarachchi is Professor of Statistics at the University of Colombo Sri Lanka. She has written several papers both in international and local journals on methodological developments for categorical data. She has received several awards (Presidential award and University of Colombo Award). She has been a visiting Research Fellow at the University of Reading UK for three years.

Cross References

►Categorical Data Analysis
►Fisher Exact Test
►Statistical Inference
►Variation for Categorical Variables

References and Further Reading

Agresti A (1992) A survey of exact inference for contingency tables. Stat Sci 7(1):131–153
Agresti A (2001) Exact inference for categorical data: recent advances and continuing controversies. Stat Med 20:2709–2722
Bishop YMM, Feinberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT press, Cambridge, MA
Cornfield J (1956) A statistical problem arising from retrospective studies. Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, vol 4. University of California, Berkeley, 135–148
Cox DR (1970) Analysis of binary data. Chapman and Hall, Boca Raton, FL
Fisher RA (1935) The logic of inductive inference. J R Stat Soc Ser A Stat Soc 98:39–54
Good P (1993) Permutation tests. Springer Verlag, New York
Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York
Mehta CR (1998) Exact inference for categorical data. Encycl Biostat 2:1411–1422

Exchangeability

Sander Greenland[1], David Draper[2]
[1]Professor
University of California-Los Angeles, Los Angeles, CA, USA
[2]Professor
University of California-Santa Cruz, Santa Cruz, CA, USA

In probability theory, the random variables $Y_1, \ldots, Y_N$ are said to be *exchangeable* (or *permutable* or *symmetric*) if their joint distribution $F(y_1, \ldots, y_N)$ is a symmetric function; that is, if F is invariant under permutation of its arguments, so that $F(z_1, \ldots, z_N) = F(y_1, \ldots, y_N)$ whenever $z_1, \ldots, z_N$ is a permutation of $y_1, \ldots, y_N$. There is a related epidemiologic usage described in the article on ►*confounding*.

Exchangeable random variables are identically distributed, and iid variables are exchangeable. In many ways, sequences of exchangeable random variables play a role in subjective Bayesian theory analogous to that played by independent identically distributed (iid) sequences in frequentist theory. In particular, the assumption that a sequence of random variables is exchangeable allows the development of inductive statistical procedures for inference from observed to unobserved members of the sequence (Bernardo and Smith 1994; De Finetti 1937, 1974; Draper 1995; Draper et al. 1993; Lindley and Novick 1981).

Now suppose that $Y_1, \ldots, Y_N$ are iid given an unknown parameter θ that indexes their joint distribution. Such variables will not be unconditionally independent when θ is a random variable, but will be exchangeable. Suppose, for example, $Y_1, \ldots, Y_N$ have a joint density. The

unconditional density of $Y_1, \ldots, Y_N$ will be

$$f(y_1, \ldots, y_N) = \int_\theta f(y_1, \ldots, y_N|\theta) dF(\theta)$$
$$= \int \prod_i f(y_i|\theta) dF(\theta).$$

Exchangeability of $Y_1, \ldots, Y_N$ follows from the identity of the marginal densities in the product.

However, given that these densities depend on θ, the integral and product cannot be interchanged, so that $f(y_1, \ldots, y_N) \neq \prod_i f(y_i)$. We thus have that a mixture of iid sequences is an exchangeable sequence, but not iid except in trivial cases.

One consequence of this result is that the usual procedures for generating a sequence $Y_1, \ldots, Y_N$ of iid random variables for inference on an unknown parameter (such as Bernoulli trials of binary data with unknown success probability) generate an exchangeable sequence when the parameter is generated randomly and the sequence is considered unconditionally. From a Bayesian perspective, this means that, when your uncertainty about the parameter is integrated with your uncertainty about the realizations of $Y_1, \ldots, Y_N$, the latter are (for you) exchangeable but dependent even if the realizations are physically independent. This subjective dependence is immediately clear if you consider (say) a sequence of 99 iid (but possibly biased) coin tosses, with Y_i the indicator of heads on toss i. Then, starting from a uniform prior distribution for the chance of heads θ, you should have $\Pr(Y_{99} = 1) = {}^1/_2$ before seeing any toss, but $\Pr(Y_{99} = 1 | Y_i = 1 \text{ for } i = 1, \ldots, 98) = 0.99$ after seeing the first 98 tosses come up heads (Good 1983).

A generalization important for statistical modeling is *partial* or *conditional* exchangeability (De Finetti, 1937, 1974). For example, suppose that the sequence $Y_1, \ldots, Y_N$ is partitioned into disjoint subsequences. Then the sequence is said to be partially exchangeable given the partition if each subsequence can be permuted without changing the joint distribution. For example, if the Y_i represent survival times within a cohort of male stroke patients, then a judgment of unconditional exchangeability of the Y_i would be unreasonable if the patient ages were known, because age is a known predictor of survival time. Nonetheless, one might regard the survival times as partially exchangeable, given age, if no further prognostically relevant partitioning was possible based on the available data.

While exchangeability is weaker than iid, finite subsequences of an infinite exchangeable sequence of Bernoulli (binary) variates have representations as mixtures of iid Bernoulli sequences – a partial converse of the fact that any mixture of iid sequences is an exchangeable sequence. More precisely, suppose that $Y_1, Y_2, \ldots$ is an infinite sequence of exchangeable Bernoulli variates (that is, every finite subsequence of the sequence is exchangeable), and that θ is the limit of $(Y_1 + \cdots + Y_n)/n$ as n goes to infinity. De Finetti [1974, Chapter 11] showed that there exists a distribution function $P(\theta)$ for θ such that, for all n,

$$\Pr(Y_1 = y_1, \ldots, Y_n = y_n) \equiv \Pr(y_1, \ldots, y_n)$$
$$= \int_\theta \theta^s (1-\theta)^{n-s} dP(\theta),$$

where $s = y_1 + \cdots + y_n$. Many Bayesian statisticians find this theorem helpful, because it partially specifies the form of the predictive probability $\Pr(y_1, \ldots, y_n)$ when $Y_1, \ldots, Y_n$ can be modeled as a subsequence of an infinite exchangeable sequence. In this representation, $P(\theta)$ is recognizable as a *prior distribution* for θ, a distribution that may be developed from what is known about θ before the Y_i are observed. As noted in (Freedman 1995), however, the strength of the theorem's conclusion is easy to overstate: it does *not* imply that all binary data must be analyzed using the representation; it merely says that if you judge $Y_1, Y_2, \ldots$ to be an exchangeable sequence, then there is a prior $P(\theta)$ that allows you to use to specify $\Pr(y_1, \ldots, y_n)$ as a mixture of iid variables over that prior.

Finite versions of De Finetti's theorem exist (Diaconis and Freedman 1980). If $Y_1, \ldots, Y_n$ is the start of an exchangeable Bernoulli sequence $Y_1, \ldots, Y_N$ and n/N is small enough, then $\Pr(y_1, \ldots, y_n)$ may be approximately represented by the mixture in the theorem, with the approximation improving as n/N approaches zero. There are further generalizations to exchangeable sequences of polytomous variates, as well as to exchangeable sequences of continuous variates (Diaconis and Freedman 1980). The latter generalization requires a prior distribution on the space of continuous distributions, however, which can be much harder to specify than a prior for a vector of multinomial parameters, and which may lead to intractable computational problems (Draper 1995).

About the Authors

For Dr. Greenland's biography *see* the entry ▶Confounding and Confounder Control.

David Draper is a Professor of Statistics in the Department of Applied Mathematics and Statistics (in the Baskin School of Engineering) at the University of California, Santa Cruz (UCSC). He was President of the International

Society for Bayesian Analysis in 2001. From 2001 to 2007 he served as the Founding Chair of the Applied Mathematics and Statistics Department at UCSC. He is a Fellow of the American Association for the Advancement of Science, the American Statistical Association, the Institute of Mathematical Statistics and the Royal Statistical Society. He is the author or co-author of 92 contributions to the research literature. Since 1993 he has given 63 invited, special invited or plenary talks at major research conferences and leading statistics departments in 21 countries worldwide. He has been an Associate Editor for 6 leading journals (including *Journal of the American Statistical Association* and the *Journal of the Royal Statistical Society*), and has organized or co-organized 6 international research conferences.

Cross References

►Bayesian Statistics
►Confounding and Confounder Control

References and Further Reading

Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, New York
De Finetti B (1937) Foresight: its logical laws, its subjective sources. In: Kyburg HE, Smokler HE (eds) Studies in subjective probability (reprinted in 1964). Wiley, New York
De Finetti B (1974) Theory of probability (two vols). Wiley, New York
Diaconis P, Freedman DA (1980) Finite exchangeable sequences. Ann Probab 8:745–764
Draper D (1995) Assessment and propagation of model uncertainty (with discussion). J R Stat Soc Ser B 57:45–97
Draper D, Hodges J, Mallows C, Pregibon D (1993). Exchangeability and data analysis (with discussion). J R Stat Soc Ser A 196: 9–37
Freedman DA (1995) Some issues in the foundations of statistics (with discussion). Found Sci 1:19–83
Good IJ (1983) Good thinking. University of Minnesota Press, Minneapolis
Lindley DV, Novick MR (1981) The role of exchangeability in inference. Annal Stat 9:45–58

Expected Value

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

Synonyms

Expectation, First usual moment, Mean value

Genesis

The origin of this notion is closely related with so called *problem of points*. It may be explained by the simplest two-person game based on tossing a coin according the following rules. In each toss, if the coin comes up heads then the player A gets a point; otherwise the point goes to the player B. The first to get 10 points wins 100 Francs. Suppose the game is interrupted in the moment when A reached 8 while B 7 points. How to divide the winnings?

This problem was posed in 17th century by a French man Antoine Gombaud (who named himself Chevalier de Méré, although he was not a nobleman) and was undertaken by Blaise Pascal and Luis de Fermat in a series of letters. Fermat noted that the game would over after four more tosses with possible $2^4 = 16$ equally likely outcomes and the fewer ratio is 11:5 for player A. Thus the player A should receive 68.75 Francs and B 31.25 Francs.

Formal Definition and Equivalent Expressions

Let $X = X(\omega)$ be a random variable defined on a set Ω of outcomes endowed with a family $A=\{A : A \subseteq \Omega\}$ of random events and let $P = P(A)$ be a probability measure on (Ω, A). *Expected value* of the r.v. X [notation: EX] is defined as the Lebesgue's integral of the function $X = X(\omega)$ w.r.t. probability measure P, i.e.

$$EX = \int_{\Omega} X(\omega) dP(\omega).$$

(if such integral exists). The expected value EX may also be expressed in terms of the distribution of X, i.e. by the measure P_X on (R, B), where B is the family of Borel sets on the real line R, defined by $P_X(B) = P(\{\omega : X(\omega) \in B\})$ for $B \in B$. Namely, EX is the Lebesgue's integral of the form

$$EX = \int_{R} x dP_X(x). \tag{1}$$

If $E|X| < \infty$, then Eq. (1) may be expressed as the Riemann–Stielties integral $EX = \int_{-\infty}^{+\infty} x dF(x)$ w.r.t. the cumulative distribution function $F(\alpha) = P(X \le \alpha)$.

Computing Rules

If X is a discrete r.v. taking values x_i with probabilities p_i, $i = 1, 2, \ldots$, then (1) reduces to

$$EX = \sum_i x_i p_i,$$

and if X is a continuous r.v. with density function $f(x)$ then (1) reduces to the Riemann integral

$$EX = \int_{-\infty}^{+\infty} x f(x) dx.$$

Moreover, if X is nonnegative, then

$$EX = \int_0^\infty [1 - F(x)]dx.$$

Caution Expected value may do not exist. Example: Cauchy random variable with density $f(x) = \frac{\lambda}{\pi[\lambda^2+(x-\theta)^2]}$ for $x \in R$.

An attribute of the expected value If $EX = \mu$ then $E(X-\mu)^2 \le E(X-c)^2$ for any $c \in R$ with the strict inequality if $c \neq \mu$.

Moments of a Random Variable It is worth to add that if $X = X(\omega)$ is a random variable and f is a Borel function, i.e. a real function of a real variable such that $\{x : f(x) \le c\} \in \mathcal{B}$ for all $c \in R$ then the composition $f[X(\omega)]$ is also random variable. Special attention is given to random variables of the form $Y(\omega) = f[X(\omega)]$, where f is a polynomial of type $f(x) = x^k$ or $(x-EX)^k$. The expectation of the form EX^k and $E(X-EX)^k$ is called, respectively, the k-th usual and central moment of the initial random variable X. If EX^k (or $E(X-EX)^k$) exists then there exist also all i-th (both usual and central) moments for $i \le k$.

Basic Properties

1. If $X(\omega) = c$ for all $\omega \in \Omega$ then $EX = c$.
2. If X is bounded then EX is finite.
3. If $X_1(\omega) \le X_2(\omega)$ for all $\omega \in \Omega$ then $EX_1 \le EX_2$.
4. $E(\alpha X_1 + \beta X_2) = \alpha EX_1 + \beta EX_2$ for all $\alpha, \beta \in R$.
5. $E\left(\sum_{n=1}^\infty X_n\right) = \sum_{n=1}^\infty EX_n$, providing $\sum E|X_n| < \infty$.
6. $E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n EX_i$, providing $X_1, \ldots, X_n$ are independent
7. $g(EX) \le E(g(X))$ for any convex function g.
8. *Markov's inequality*: $P[f(X) \ge c] \le \frac{E[f(X)]}{c}$ for any nonnegative function g and positive scalar c.
9. *Weak law of large numbers*: For any sequence $\{X_n\}$ of independent identically distributed random variables X_n with finite expectation $EX_n = \mu$ and for any $c > 0$, $\lim_{n\to\infty} P\left(\left|\frac{X_1+\ldots+X_n}{n} - \mu\right| < c\right) = 1$
10. *Strong law of large numbers:* For any sequence $\{X_n\}$ of independent identically distributed random variables X_n with finite expectation $EX_n = \mu$, $P\left(\lim_{n\to\infty} \frac{X_1+\ldots+X_n}{n} = \mu\right) = 1$.

Expected Values of Some Well Known Discrete and Continuous Distributions

1. Zero-one distribution. It corresponds to a random variable X taking only two values: 0 and 1 with corresponding probabilities $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $0 < p < 1$. In this case $EX = p$.
2. Binomial (or Bernoulli) distribution with parameters n and p, where n is a positive integer while $0 < p < 1$. It corresponds to the sum $X = X_1 + \ldots + X_n$ of n independent zero-one random variables with $P(X = 1) = p$. In this case $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$ and $EX = np$.
3. Geometric distribution with parameter p. It corresponds to the time $X = n$ of the first success ($X_n = 1$ while $X_i = 0$ for all $i < n$) in the sequence $\{X_i\}$ of independent zero-one random variables X_i with the same probability $P(X_i = 1)$. In this case $P(X = n) = p(1-p)^{n-1}$ for $n = 1, 2, \ldots$ and $EX = \frac{1}{p}$.
4. Poisson distribution with a positive parameter λ. It corresponds to a discrete random variable X being the limit of the sequence $\{X_i\}$ of Bernoulli distributions with parameters (n_i, p_i) where $\lim_{i\to\infty} n_i = \infty$ and $\lim_{i\to\infty} n_i p_i = \lambda$. Formally $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, \ldots$ In this case $EX = \lambda$.
5. Uniform distribution on the interval $< a, b >$. It corresponds to a continuous random variable X with the density function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if} \quad a \le x \le b \\ 0 & \text{if} \quad \text{not.} \end{cases}.$$

   In this case $EX = \frac{a+b}{2}$.
6. Exponential distribution with a parameter $\lambda > 0$. It corresponds to a continuous random variable X with density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if} \quad x \ge 0 \\ 0 & \text{if} \quad x < 0 \end{cases}.$$

   In this case $EX = \frac{1}{\lambda}$.
7. Normal distribution with parameters $\mu \in R$ and $\sigma^2 > 0$. It corresponds to a continuous random variable X with density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for all } x \in R.$$

   In this case $EX = \mu$.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Laws of Large Numbers
►Random Variable

References and Further Reading

Brémaud P (1994) An introduction to probabilistic modeling. Springer-Verlag, New York

Devlin K (2008) The unfinished game: Pascal, Fermat, and the seventeenth-century letter that made the world modern. Basic Books, New York
Feller W (1971) An introduction to probability theory and its applications, vol 2. Wiley, New York
Kolmogorov AN (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. (Foundations of the theory of probability, 1956 2nd edn. Chelsea, New York) Springer, Berlin
Prokhorov AW (1982) Expected value. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 3. Soviet Encyclopedia, Moscow, pp 600–601 (in Russian)
Stirzaker D (1995) Elementary probability. Cambridge University Press, Cambridge

Experimental Design: An Introduction

Klaus Hinkelmann
Professor Emeritus
Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Experimental sciences and industrial research depend on data to draw inferences and make recommendations. Data are obtained in essentially two ways: from observational studies or from experimental; i.e., interventional, studies. The distinction between these two types of studies is important, because only experimental studies can lead to causal inferences. In order to ensure that proper inferences can be drawn, any such experiment has to be planned carefully subject to certain principles of *design of experiments*.

Steps of Designed Investigations

Any investigation begins with the formulation of a question or research hypothesis in the context of a particular subject matter area, such as agriculture, medicine, industry, etc. For a given situation the researcher has to identify and characterize the experimental units to be used in the experiment. The experimental units are then subjected to different treatments, which are the objective of the study and about which statistical and scientific inferences should be drawn. These treatments are employed in a suitable error-control design adhering to the three important principles of experimental design as exposited by R.A. Fisher (1926, 1935): ▶Randomization, replication, and local control (blocking). Randomization of the treatment assignment is performed over the entire set or suitable subsets (blocks) of the experimental units in order to eliminate any bias. Replication of the treatments is essential to assess the experimental error that is needed to formally draw any statistical inference, such as testing of hypotheses or estimating confidence intervals in order to assess the efficacy of the treatments. The randomization procedure is also used to derive linear models on which the statistical analysis of the observations is based, usually some form of ▶analysis of variance (see Hinkelmann and Kempthorne 2008). Based on the analysis of the data, the results need to be interpreted in the context of the originally posed research hypothesis. This may lead to further investigations of new hypotheses in subsequent experiments.

Components of Experimental Design

Each experimental design consists of three components: treatment design, error-control design, and sampling or observational design. The *treatment design* determines the number and types of treatments to be included in the experiment. The treatments may be quantitative or qualitative. They may have a factorial structure in that each treatment is determined by a combination of different levels of two or more treatment factors. The most common factorial structure has each treatment occurring at two levels, especially for screening experiments. Even that may lead to a number of treatments too large for practical purposes. In that case a fractional factorial may have to be considered, where only a suitable subset of all level combinations is included. Once the appropriate treatment design has been determined, one must consider how to assign the treatments to the available experimental units; i.e., one must choose a suitable *error-control design*. For example, if the experimental units are essentially uniform, then a completely randomized design is appropriate, which means that each treatment is randomly assigned to, say, r experimental units, where r is the number of replications for each treatment. If, however, the experimental units are characterized by different extraneous factors, such as locations in an agricultural experiment or age and gender in a psychological experiment or type of raw material and type of machine in an industrial setting, then some form of block design has to be used. Each combination of different "levels" of the extraneous factors forms a set of essentially uniform experimental units to which the treatments are then randomly assigned. The most common block design is the randomized complete block design, where each block has as many experimental units as there are treatments, and each treatment is randomly assigned to one unit in each block. Other types of block designs are ▶incomplete block designs, where not every treatment can occur in each block. This poses a number of combinatorial problems. Still other error-control designs

employ blocking in two directions, such as subjects and periods in the form of a Latin square in a psychological experiment. The aim of blocking is to generate sets with uniform experimental units which will result in reduction of error and, hence, in a more efficient design. The third component of an experimental design is the *sampling* or *observational design*. Its function is to determine what the observational units are. In most experiments only one observation is obtained from each experimental unit, in which case experimental and observational units are identical. On other occasions several observations are obtained from each experimental unit, for example several items from an industrial production lot. This is referred to as subsampling. In this case experimental and observational units are not identical. In connection with the error-control design this then allows the separate estimation of experimental and observational (measurement) errors.

Modeling Observations

In most cases it is clear what the observations should be, but in other cases this may need careful consideration in order to arrive at a useful type of measurement which will be most meaningful in the context of the formulated research hypothesis. Following the choice of measurement, a model for the observations can, in general terms, be expressed as follows:

$$\text{Response} = f(\text{explanatory variables}) + \text{error},$$

where f is an unknown function and the explanatory variables refer to treatment and blocking factors as employed in the treatment and error-control designs, respectively. More specifically, however, using the notion of experimental unit-treatment additivity and randomization theory one can derive a linear model of the form

$$\text{Response} = \text{overall mean} + \text{block effect}(s) + \text{treatment effect} + \text{error},$$

where the distributional properties of the error term are determined essentially by the randomization process (see Hinkelmann and Kempthorne 2008). This model may be enlarged, where appropriate, by including certain block-treatment interaction terms depending on the blocking structure (see Cox 1984).

Analysis of Data

The analysis of the observations from a designed experiment is based on the model given above using analysis of variance techniques. The analysis of variance table is used to estimate the error variance component (including experimental and observational error) or, in the case of subsampling, the experimental and observational error variance components. The F-test based on the ratio (treatment mean square/error mean square) or, in case of subsampling, (treatment mean square/experimental error mean square) serve as approximation to the randomization test that there are no differences among the treatment effects. Other tests may have to be performed to more fully investigate differences among treatments by considering specific comparisons among the treatments or treatment effect trends or interactions between treatment factors or block × treatment interactions. The possibilities are being dictated essentially by the treatment and error-control designs employed and the research hypothesis postulated at the beginning of the experiment.

About the Author

Dr. Klaus Hinkelmann is Professor Emeritus of Statistics, Virginia Polytechnic Institute and State University (Virginia Tech), where he has served as Director of Graduate Studies (1976–1982) and Head of Department (1982–1993). He served as Associate Editor of *Biometrics* (1975–1984) and the *Biometrical Journal* (1995–1999), Editor of *Biometrics* (1990–1993) and the *Current Index to Statistics* (1995–1999), Book Review Editor of *Journal of Statistical Planning and Inference* (1995–1997), Member of Editorial Board *Selected Tables in Mathematical Statistics* (1974–1990). He was President of the Virginia Chapter of the American Statistical Association (1983–1984), Vice-President of Mu Sigma Rho (American Statistical Honor Society) (1994–1996), Member, Council of the International Biometric Society (1982–1985, 1988–1989). He was elected Fellow of the American Statistical Association (1983), Member of the International Statistical Institute (1985), Fellow of the American Association for the Advancement of Science (1987). Dr. Hinkelmann is the author/co-author of more than 50 articles in statistical and subject-matter journals. He is the author/editor of three books. With Oscar Kempthorne as co-author he wrote *Design and Analysis of Experiments, Vol.1: Introduction to Experimental Design* (1994, 2nd edition 2008) and *Vol.2: Advanced Experimental Design* (John Wiley and Sons 2005).

Cross References

►Clinical Trials: An Overview
►Design of Experiments: A Pattern of Progress
►Factorial Experiments
►Optimum Experimental Design
►Randomization
►Statistical Design of Experiments (DOE)
►Uniform Experimental Design

References and Further Reading

Cox DR (1984) Interaction (with discussion). Int Stat Rev 52:1–32

Fisher RA (1926) The arrangement of field experiments. J Min Agr Engl 33:503–513

Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh

Hinkelmann K, Kempthorne O (2008) Design and analysis of experiments, vol 1: introduction to experimental design, 2nd edn. Wiley, Hoboken

Expert Systems

Ali S. Hadi
Professor and Vice Provost
The American University in Cairo, Cairo, Egypt
Emeritus Professor
Cornell University, Ithaca, NY, USA

The area of artificial intelligence (AI) is concerned with the design and implementation of computer systems that exhibit the abilities associated with the human intelligence (e.g., the ability to memorize, learn, think, etc.). The AI area has seen a great surge of research and rapid development during the past three or so decades; see, e.g., Luger (2009) and the references therein. As a result many branches of AI (e.g., automatic game playing, automated reasoning and theorem proving, robotics, artificial vision, natural language processing (see ►Statistical Natural Language Processing), pattern recognition (see ►Pattern Recognition, Aspects of and ►Statistical Pattern Recognition Principles), expert systems, etc.) have evolved. Expert systems are one of the successful branches of AI and most AI branches include an expert system component in them.

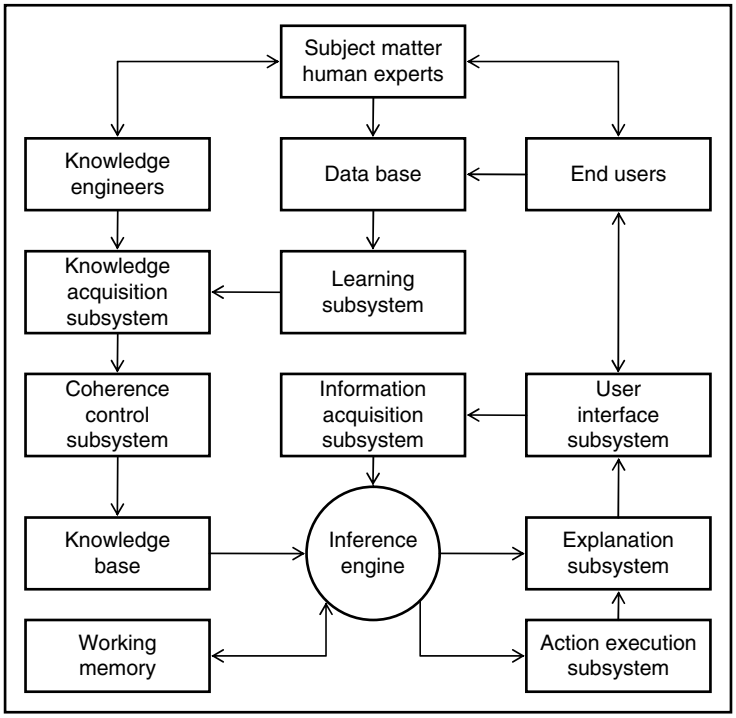
Several definitions of expert systems have evolved over time (see, e.g., Stevens 1984; Durkin, 1994; Castillo et al. 1997). These definitions, however, can be summarized as follows: An expert system is a computer system (hardware and software) that emulates human experts in a given area of specialization. As such, an expert system attempts to either completely replace the human experts (e.g., Automated Teller Machines), help the human experts become even better experts by providing them with fast answers to complex problems (e.g., Medical Expert Systems), or provide decision makers quick answers that enables them to make wise decisions in the face of uncertainty (e.g., Automating the Underwriting of Insurance Applications; Aggour and Cheetham, 2005).

Successful expert systems are usually domain specific. One of the earliest expert systems is DENDRAL, developed in the late 1960s (Shortliffe, 1976) to infer the structure of organic molecules from information about their chemical characteristics. MYCIN is another example of domain specific (medicine) expert systems. It was developed in the 1970s to diagnose and prescribe treatment for spinal meningitis and bacterial infections of the blood (Buchanan and Shortliffe, 1984). Today expert systems have varied applications in many fields such as medicine, education, business, design, and science (Waterman, 1985; Feigenbaum et al., 1988; Durkin, 1994). Dozens of commercial software packages with easy interfaces are now available which are used to develop expert systems in these fields of applications. Several books and papers have been written about the subject; see, e.g., Neapolitan (1990), Ignizio (1991), Jensen (1996), Schneider et al. (1996), Castillo et al. (1997), Jackson (1999), Pearl (1988, 2009), Russell and Norvig (2003), Giarratano and Riley (2005), Cowell et al. (2007), and Darwiche (2009).

The development of an effective and successful expert system requires team-work and collaboration of humans from different fields. First, domain experts (e.g., doctors, lawyers, business managers, etc.) provide knowledge about the specific domain. For example, medical doctors can provide information about the relationships among symptoms and diseases. Second, *knowledge engineers* translate the knowledge provided by the subject-matter specialists into a language that the expert system can understand. Third, during the design, development, and implementation of an expert system, it is important that the needs of the end-users are kept in mind and that their input be taken into consideration throughout the process.

The information provided by the domain experts is stored in one of the components of expert systems known as the *Knowledge Base*. A knowledge base can consist of two types of knowledge: deterministic and stochastic. Deterministic knowledge typically consists of a set of crisp *rules*. The simplest rule takes the form "*If A Then B*," where *A* is a logical expression, known as the *premise* or *antecedent* of the rule, which is related to input variable(s), and *B* is a logical expression, known as the *conclusion* or *consequent* of the rule, which is related to an output variable. An example of a rule is "*If* PassWord = Correct *Then* Access = Permitted," where the input variable PassWord can take one of two values (Correct or Incorrect) and Access is an output variable that can take one of two or more values (e.g., Permitted or Denied).

Stochastic knowledge usually involves some elements of uncertainty. For example, the relationships among symptoms and diseases are not one-to-one. A disease can cause many symptoms and the same symptom can be caused by a number of different diseases. It is therefore necessary to develop some means for dealing with the



Expert Systems. Fig. 1 A typical architecture of an expert system

uncertainty and imprecision in the knowledge we receive. Expert systems that involve only deterministic knowledge are referred to as *deterministic* or *rule-based* expert systems. Two types of expert systems that can deal with stochastic or imprecise knowledge are *fuzzy* expert systems and *probabilistic* expert systems.

Fuzzy systems, first introduced by Zadeh (1965), are based on fuzzy logic instead of Boolean logic. The knowledge base of a fuzzy expert system consists of a set of *fuzzy rules* and a set of *membership functions* defined on the input variables. Fuzzy rules are similar in form to the crisp rules except that the premise of the fuzzy rule describes the degree of truth for the rule and the membership functions are applied to their actual values to determine the degree of truth for each rules conclusion. The edited book by Siler and Buckley (2005) is a good source of many papers and references on fuzzy expert systems.

In probabilistic expert system, the uncertainty is measured in terms of probabilities. Thus, the knowledge base of a probabilistic expert system consists of the joint probability distribution function (pdf) of all variables. This joint pdf, however, can be specified by a set of conditional probability distributions, one for each variable. Each conditional pdf gives the probability for each value of the variable given the values of a set of associated variables known as its *parents*. The specifications of these conditional distributions can be either given by the domain experts and/or learned from the available data.

To emulate the performance of the human experts in a given domain of application, expert systems consist of several components and sub components. A typical architecture of an expert system is depicted in Fig. 1, which is adapted from Castillo et al. (1997), and explained below. Knowledge can be provided directly by the subject-matter experts, but it can also be learned from historical data. The data, when available, are stored in a *data base* and the knowledge learned from data is done by the *learning subsystem*. The flow of knowledge from either source is controlled by the *knowledge acquisition* subsystem. Before the knowledge is added to the *knowledge base*, it must be checked for consistency by the *coherence control* subsystem to prevent incoherent knowledge from reaching the knowledge base.

Through the *User Interface* subsystem, *Information Acquisition* subsystem collects information or inquiries from the end-users. This information together with the knowledge in the knowledge base are used by the *Inference Engine* to draw conclusions. The inference engine and knowledge base are then the most important components of an expert system.

The inference engine is the brain of an expert system. It contains the decision-making logic of the expert system which allows the expert system to draw conclusions and provide expert answers (e.g., diagnose a disease). In deterministic expert systems, the inference is made based on *Boolean* or classic logic. In advanced expert systems, the inference engine can also enhance the knowledge base by adding a set of *concluded* rules using various inference strategies such as *Modus Ponens*, *Modus Tollens*, *Rule Chaining*, etc.

In fuzzy expert systems, the inference is made based on *fuzzy* logic; see, e.g., Schneider et al. (1996). Given realizations of a specified set of variables, the inference engine of probabilistic expert systems computes the conditional probabilities for each values of the other variables given the realization of specified set of variables. This is known as the *propagation* of uncertainty.

In the process of drawing conclusions, the inference engine may need to store temporary information, which is stored in the *Working Memory*. Finally, if actions are needed, they are specified by the inference engine through the *Action Execution* subsystem and the actions and conclusions are explained to the users through the *Explanation* subsystem.

About the Author

For biography *see* the entry ►Ridge and Surrogate Ridge Regressions.

Cross References

►Forecasting Principles
►Fuzzy Logic in Statistical Data Analysis
►Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Fuzzy Sets: An Introduction
►Statistical Natural Language Processing

References and Further Reading

Aggour KS, Cheetham W (2005) Automating the underwriting of insurance applications. In: Neil Jacobstein and Bruce Porter (eds) Proceedings of the seventeenth innovative applications of artificial intelligence conference, Pittsburgh, PA. AAAI Press, Menlo Park, CA pp 1451–1458

Buchanan BG, Shortliffe EH (eds) (1984) Rule-based expert systems: the MYCIN experiments of the Stanford heuristic programming project. Addison-Wesley, Reading, MA

Castillo E, Gutiérrez JM, Hadi AS (1997) Expert systems and probabilistic network models. Springer-Verlag, New York

Cowell RG, Dawid AP, Lauritzen SL, Spiegelhalter DJ (2007) Probabilistic networks and expert systems exact computational methods for Bayesian Networks. Springer-Verlag, New York

Darwiche A (2009) Modeling and reasoning with Bayesian networks. Cambridge University Press, Cambridge, UK

Durkin J (1994) Expert systems: design and development. Maxwell Macmillan, New York

Feigenbaum E, McCorduck P, Nii HP (1988) The rise of the expert company. Times Books, New York

Giarratano JC, Riley G (2005) Expert systems, principles and programming, 4th edn. PWS Publishing, Boston

Ignizio JP (1991) Introduction to expert systems: the development and implementation of rule-based expert systems. McGraw-Hill, New York

Jackson P (1999) Introduction to expert systems, 3rd edn. Addison-Wesley, Reading, MA

Jensen FV (1996) An introduction to Bayesian networks, UCL Press, London

Luger GF (2009) Artificial intelligence: structures and strategies for complex problem solving, 6th edn. Addison-Wesley, Reading, MA

Neapolitan R (1990) Probabilistic reasoning in expert systems: theory and Algorithms. Wiley, New York

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo, CA

Pearl J (2009) Causality: models, reasoning and inference, 2nd edn. Cambridge University Press, Cambridge, UK

Russell SJ, Norvig P (2003) Artificial Intelligence: a modern approach, 2nd edn. Prentice Hall/Pearson Education, NJ

Schneider M, Langholz G, Kandel A, Chew G (1996) Fuzzy expert system tools. Wiley, New York

Shortliffe EH (1976) Computer-based medical consultations: MYCIN. Elsevier/North Holland, New York

Siler W, Buckley JJ (2005) Fuzzy expert systems and fuzzy reasoning. Wiley, New York

Stevens L (1984) Artificial intelligence. The search for the perfect machine. Hayden Book Company, Hasbrouck Heights, NJ

Waterman DA (1985) A guide to expert systems. Addison-Wesley, Reading, MA

Zadeh LA (1965) Fuzzy sets. Inform Control 8:338–353

Explaining Paradoxes in Nonparametric Statistics

Anna E. Bargagliotti[1], Donald G. Saari[2]
[1]University of Memphis, Memphis, TN, USA
[2]Distinguished Professor, Director
University of California-Irvine, Irvine, CA, USA

Introduction

Nonparametric statistical methods based on ranks are commonly used to test for differences among alternatives and to extract overall rankings. But depending on the method used to analyze the ranked data, inconsistencies can occur and different conclusions can be reached. Recent advances (Bargagliotti and Orrison 2009; Bargagliotti and Saari 2009; Haunsperger and Saari 1991; Haunsperger 1992; Saari 2008) prove that all possible inconsistencies among

outcomes for these methods are caused by hidden symmetries within the data.

To analyze how methods utilize and interpret different data configurations, decompose a nonparametric method into a two step process. The first step is where the method uses a uniquely defined procedure to convert ranked data into an overall ranking. The second step uses a test statistic (based on the procedure ranked outcome) to determine whether there are significant differences among alternatives. Differences and even conflicting results can occur at each step when analyzing the same ranked data set. These peculiarities can complicate the choice of an appropriate method in a given situation.

To demonstrate, consider the general setting in which ranked performance data are collected on a set of three alternatives where larger values indicate a better performance:

A	B	C
15	14	13
10	12	11
8	7	9
6	5	4
1	2	3

(1)

With the Kruskal–Wallis procedure (Kruskal and Wallis 1952), which computes the rank-sum for each alternative, these data define the tied ranking $A \sim B \sim C$. In contrast, Bhapkar's V procedure (Bhapkar 1961) analyzes ranked data by considering all possible 3-tuples of the form (a_i, b_j, c_m) for $i, j, m = 1, \ldots, 5$. For each 3-tuple, the V procedure assigns a point to the alternative with the highest rank; the tally for each alternative is the sum of the assigned points and the V procedure ranking is based on these tallies. With the Eq. 1 data, the V procedure yields the $A > B > C$ ranking, which conflicts with the Krusal–Wallis ranking.

A still different outcome arises with pairwise comparison procedures such as the Mann–Whitney (Whitney 1947) or Wilcoxon (Wilcoxon 1945) (see ▶Wilcoxon–Mann–Whitney Test). These rules also are based on all possible tuples, but now, for a specified pair, an alternative receives a point for each tuple where it is ranked above its opponent. The Eq. 1 data lead to the $A > B, B > C, C > A$ cyclic rankings, which conflict with those from the earlier approaches.

In this article, we show that subtle symmetry structures hidden within the data explain why different outcomes can be obtained by different nonparametric ranking procedures for the same data set. Namely, we show how and why different nonparametric methods interpret data structures differently and why this causes the inconsistencies experienced by different methods.

Same Data, Different Methods, Different Outcomes

Two forms of ranked data can be constructed: full ranked data and block ranked data. An entry in a full ranked data set describes the datum's ranking relative to all others in the set; an entry in a block ranked data set describes the datum's ranking relative to all others within the block. For example, Eq. 2 illustrates a full ranked data set with the overall ordering of all the entries:

A	B	C
1	2	3
5	6	4
9	8	7

(2)

while Eq. 3 ranks the entries within each row, so a row is considered a block:

A	B	C
1	2	3
2	3	1
3	2	1

(3)

Standard methods of analysis for the full ranked data set include, but are not limited to, the Kruskal–Wallis, Mann–Whitney, and Bhapkar's V tests. When analyzing the block ranked data, the Friedman (Friedman 1937) or Anderson tests (Anderson 1959) are commonly employed. Each of these nonparametric methods is susceptible to ranking and statistical paradoxes. We explain how certain data configurations cause either ranking or statistical significance paradoxes.

For the purpose of clarity, our examples and analysis emphasize the three-sample setting with the three alternatives A, B, and C. All symmetries for the three samples thus come from S_3 – the space of all ways to permute these alternatives. Two natural symmetries, the $\mathbb{Z}_3$ orbit and the $\mathbb{Z}_2$ orbit of a triplet, are what cause the differences among nonparametric procedures.

The set of permuted rankings defined by the $\mathbb{Z}_3$ orbit of the ranking $A > B > C$ is:

$$\{A > B > C, \quad B > C > A, \quad C > A > B\}. \tag{4}$$

while the other *"rotational symmetry"* set is generated with the ranking $A \succ C \succ B$:

$$\{A \succ C \succ B, \quad C \succ B \succ A, \quad B \succ A \succ C\}. \tag{5}$$

Either configuration is constructed by moving the top ranked alternative in one triplet to be bottom ranked in the next triplet. Thus each alternative is in first, second, and third place precisely once over the triplet of rankings.

Data sets that feature one of these $\mathbb{Z}_3$ structures (but not the $\mathbb{Z}_2$ structure described below) require combining three rotational parts:

A	B	C
27	26	25
22	24	23
20	19	21

,

A	B	C
16	18	17
14	13	15
12	11	10

,

A	B	C
8	7	9
6	5	4
1	3	2

(6)

The first array arranges the row data in the expected $A \succ B \succ C$, $B \succ C \succ A$, $C \succ A \succ B$ order of a $\mathbb{Z}_3$ orbit. Notice how the top defining row of the next two sets also manifest this cycle; e.g., the top rows also define the $A \succ B \succ C$, $B \succ C \succ A$, $C \succ A \succ B$ order. The rest of each block array continues the $\mathbb{Z}_3$ cyclic structure. Thus each set of three rows reflects the $\mathbb{Z}_3$ symmetry, and the three sets are connected with a $\mathbb{Z}_3$ symmetry.

The importance of the rotational symmetry is captured by Theorem 1.

Theorem 1 *Bargagliotti and Saari (2009) Let A, B, C represent the three alternatives. For a data component that introduces a rotational symmetry (Eq. 4 or 5), all three-sample procedures yield a completely tied outcome. The pairwise procedure outcomes for such a component, however, form a cycle where, for each pair, the tally of tuples is the same. Such a component is strictly responsible for all non-transitive paired comparison behavior.*

To illustrate with the Eq. 6 pure rotational data, the Kruskal–Wallis and V procedures have the ranking $A \sim B \sim C$, but the paired comparisons yield the $A \succ B$, $B \succ C$, $C \succ A$ cycle. Components within a data set yielding these rotational configurations are totally responsible for all possible paired comparison paradoxes (for methods such as the Mann–Whitney and the Wilcoxon) including cyclic effects and all possible disagreements with the three-sample Kruskal–Wallis test. All three-sample procedures essentially ignore rotation configurations present in data and consider the alternatives as tied.

The *"inversion symmetry"* of the $\mathbb{Z}_2$ orbit consists of the ranking of a triplet and its inverted version. As an illustration, the $\mathbb{Z}_2$ orbit of the ranking $A \succ B \succ C$ is the set:

$$\{A \succ B \succ C, \; C \succ B \succ A\} \tag{7}$$

The two other sets generated by the $\mathbb{Z}_2$ orbit are

$$\{A \succ C \succ B, \; B \succ C \succ A\}, \quad \{B \succ A \succ C, \; C \succ A \succ B\} \tag{8}$$

An example capturing the inversion structure with no rotational symmetry is:

A	B	C
12	11	10
7	9	8
6	4	5
1	2	3

(9)

In this example, the rows have opposing rankings, while the first and fourth rows and the second and third rows reverse each other. (See Bargagliotti and Saari (2009) for examples that do not have the same number of rows with one ranking as with its reversal.)

The following theorem captures the importance of the inversion symmetry for ranking procedures.

Theorem 2 *Bargagliotti and Saari (2009) Consider the three alternatives A, B, C. For a data component that creates a inversion symmetry (e.g., Eqs. 7, 8), the Kruskal–Wallis and the pairwise procedures yield a tied outcome, but the ranking for all other three-sample procedures is not a tie. All possible differences among three-sample procedures are caused by this inversion symmetry.*

To illustrate with the pure inversion data in Eq. (9), the Kruskal–Wallis and pairwise procedures have the $A \sim B \sim C$ ranking, while the V procedure outputs $A \sim B \succ C$. For any data set, components causing an inversion symmetry are treated as a tie by the Kruskal–Wallis and paired comparison methods, but they influence, and can even change the rankings, of all other three-sample procedures. Indeed, the rotational and inversion symmetries are responsible for all possible differences in this class of nonparametric procedures. Thus components causing rotational symmetries create differences between paired comparison and three-sample approaches, while inversion symmetries cause all possible differences among three-sample procedures; only the Kruskal–Wallis test is immune to these symmetries Bargagliotti and Saari (2009).

These symmetries also cause significant differences when testing among alternatives with block ranked data; we illustrate this behavior with the Friedman and the Anderson tests. The Friedman test statistics is given by

$$Q = \frac{n\sum_{j=1}^{k}(\bar{r}_{\cdot j} - \bar{r})^2}{\frac{1}{n(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}(r_{ij} - \bar{r})^2} \tag{10}$$

where n is the number of blocks, k is the number of alternatives, $\bar{r}_{\cdot j}$ is the mean of the rankings for alternative j, r_{ij} is the *ith* ranking for alternative j, and $\bar{r}$ is the grand mean of the rankings. The Anderson statistic is defined as

$$A = \frac{k-1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(n_{ij} - \frac{n}{k}\right)^2 \tag{11}$$

where n_{ij} is the number of times alternative j was assigned rank i by the respondents, n is the number of blocks, and k is the number of alternatives. Both statistics follow a χ^2 distribution with $k-1$ and $(k-1)^2$ degrees of freedom respectively.

As both tests are over three-samples, the above theorems assert that all differences are displayed by the inversion, $\mathbb{Z}_2$ symmetries that are embedded within data sets. As this suggests examining difference over a pure inversion block set, consider the following data set where the number of rows with a given ranking equals the number of rows with the inverted ranking.

Ranking	*Number of Voters*
$A \succ B \succ C$	12
$C \succ B \succ A$	12
$B \succ A \succ C$	10
$C \succ A \succ B$	10
$A \succ C \succ B$	3
$B \succ C \succ A$	3

(12)

The effects of this inversion symmetry on tests is reflected by the following statement:

Example 1 *For the data set in Eq. 12, the Anderson test statistic, A, equals* 10.72 *resulting in a p-value of* .029 *while the Friedman statistic, Q, equals* 0 *with associated p-value of* 1. *At the* .05 *significance level, the Anderson test considers the three alternatives significantly different but the Friedman test does not.*

It is the inversion symmetry structures embedded within data that can cause these tests to yield conflicting outcomes. This is because the Friedman test is based on the average ranks of each alternative while the Anderson statistic is more sensitive to inversion symmetry structures as its outcome is based on how often each alternative is in first, second, and third place. The tests disagree when the block data reports similar means but vastly different marginal distributions for each alternative caused by the inversion symmetries. In Eq. 3, each alternative receives an average weighted score of 2 overall but the data has the following distribution:

Alternative	First Place	Second Place	Third Place
A	15	13	22
B	20	24	6
C	15	13	22

(13)

As this table displays, although the average score for each alternative is the same, the distribution of rankings (caused by the inversion symmetry) is quite different. With this relatively small data set of 50 observations, the Anderson test "picks up" on this difference in distribution while the Friedman test does not (see Bargagliotti and Orrison (2009) for a mathematical characterization of these types of paradoxes).

These examples illustrate how different nonparametric methods interpret the specific data configurations. Different tests and procedures emphasize or deemphasize certain symmetry configurations in the data, which can lead to conflicting outcomes. But by decomposing the data space in terms of the relevant symmetry configurations, it becomes possible to determine how different nonparametric methods interpret specific sets of data (see Bargagliotti and Saari (2009) for the full data decomposition and Marden (1995) and Bargagliotti and Orrison (2009) for the block data decomposition). Depending on the situation, certain methods may be more desirable to use than others. For instance, if a certain data structure that is not viewed as being important influences the outcomes of a specified method, then the method may not be an appropriate choice. Conversely, if a method ignores a type of data structure that is accepted as being valuable while another method does not, then this information provides support for adopting the second method.

Conclusion

Although nonparametric statistical methods based on ranks are commonly used to test for differences among alternatives and to extract overall rankings, different methods can lead to different conclusions. These paradoxes and inconsistencies can occur at both the ranking procedure

step or the test statistic step; they are based on how the methods ignore, or react, to certain data structures. For three alternatives, these structures are completely captured by the rotational and inversion symmetries (Bargagliotti and Saari 2009; Saari 2008).

By understanding what types of data cause statistical methods to have different outcomes, we have better understanding of how different methods work. This provides insight about which nonparametric method should be applied in a given situation and offers guidelines for the construction of new methods that utilize desired data structures.

About the Authors

Dr. Bargagliotti is an Assistant Professor of Mathematics and Statistics at the University of Memphis. Since finishing her doctorate in 2007, she has published several articles related to nonparametric statistics as well as mathematics and statistics education. In addition, she has received several education related grants as PI, co-PI, or Senior Personnel totaling in over 6 million in external funding from the U.S. Department of Education, Tennessee Department of Education, and the National Science Foundation.

Professor Saari was elected to the United States National Academy of Sciences (2001). He was named a Fellow of the American Academy of Arts and Sciences (2004). He holds four honorary doctorates. Professor Saari was awarded the Lester R. Ford Award (1985), Chauvenet Prize (1995) and Allendoerfer Award (1999). He holds the Pacific Institute for the Mathematical Sciences Distinguished Chair at the University of Victoria in British Columbia, Canada. He has published 11 books and about 180 papers on topics ranging from dynamical systems, celestial mechanics, mathematical economics, decision analysis (in engineering and elsewhere) and voting theory. In 2009, Professor Saari was elected to the Finnish Academy of Science and Letters.

Cross References

►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests

References and Further Reading

Anderson RL (1959) Use of contingency tables in the analysis of consumer preference studies. Biometrics 15:582–590
Bargagliotti AE, Orrison M (2009) Statistical inconsistencies of ranked data (manuscript under review)
Bargagliotti AE, Saari DG (2009) Symmetry of nonparametric statistical tests on three samples (manuscript under review)
Bhapkar VP (1961) A nonparametric test for the problem of several samples. Ann Math Stat 32:1108–1117
Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32:675–701
Haunsperger DB, Saari DG (1991) The lack of consistency for statistical decision procedures. Am Stat 45:252–255
Haunsperger DB (1992) Dictionary of paradoxes for statistical tests on k samples. J Am Stat Assoc 87:249–255
Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47:583–612
Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18:50–60
Marden JI (1995) Analyzing ranked data, Chapman & Hall, London
Saari DG (2008) Disposing dictators, demystifying voting paradoxes, Cambridge University Press, New York
Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bull 1:80–83

Exploratory Data Analysis

Kate Smith-Miles
Professor, Head
Monash University, Melbourne, VIC, Australia

Exploratory data analysis (EDA) is a term first utilized by John Tukey (1977), and is intended to contrast with the more traditional statistical approach to data analysis that starts with hypothesis testing and model building. Instead of using confirmatory data analysis (CDA) methods to verify or refute suspected hypotheses, Tukey (1977) advocated the use of basic descriptive statistics and visualization methods to generate information that would lead to the development of hypotheses to test. The objectives of EDA (see Velleman and Hoaglin 1981) are therefore to provide statistical summaries of the data as a pre-processing step before building a model or testing hypotheses. The EDA phase enables new hypotheses to be suggested about the causes of the relationships in the data, and provides an opportunity to determine whether the assumption that various models rely upon are valid for the particular dataset. It provides us with a very rapid feel for the data: the shape of its distributions, the presence of ►outliers (which may represent errors in the data which need to be corrected prior to hypothesis testing and model building, or may be accurate exceptions), and the measures of central tendency and spread that are so critical for understanding the character of the data.

The most common methods employed in EDA are the five-number summaries of the data, consisting of the minimum and maximum value of each variable, the median,

Exploratory Data Analysis. Table 1 Basic graphical methods of Exploratory Data Analysis

Box Plot (Box-and-Whiskers plot)	Graphical display of five number summaries of data, showing quartiles, as well as minimum and maximum values. The graphical representation shows the distribution as well as existence of outliers.
Stem Plot (Stem-and-leaf display)	Graphical display showing the distribution of the data created by sorting the data, and then using the leading digits from each data value as the "stems" and the subsequent digits as the "leaves." A vertical line separates the leaves from each stem. The digits to the right of the vertical line each represent one observation. The stems should be selected so that they create evenly distributed "bins" for the resulting histogram.
Bubble Charts	Graphical display of showing colored bubbles of different sizes in a 2-dimensional space. A bubble chart is therefore able to represent four variables simultaneously, with two variables represented by the *x* and *y* axes, one (nominal) variable represented by the color of a bubble, and one (ordinal) variable represented by the size of the bubble.
Radar/Spider Plots	Graphical display of multivariate data with each variable represented as an axis stemming from a central origin at an angle given by $360/n$ for n variables ($n \geq 3$). A multivariate data point is usually represented in a unique color and shown as a "spider web" woven through the relevant points of each axis.
Scatterplot matrices	Graphical display of multivariate data with the cell in row *i* and column *j* of the matrix containing a scatterplot of variable *i* versus variable *j*.

and the first and third quartiles, often represented as a box-plot (see ►Summarizing Data with Boxplots). Other plots of the data are also commonly employed, such as histograms, scatter plots, stem-and-leaf plots, all designed to explore the relationships that might exist in the data, and the distributions of the variables, including the presence of outliers. A number of graphical EDA methods are summarized in Table 1.

In addition to graphing of variables, simple statistical measures such as the correlations between variables, as well as multi-way frequency tables are often constructed to explore the dependence between variables and explore possible causal effects. Tukey (1977) proposed a number of other useful measures to characterize a data sample, including the median polish (see Emerson and Hoaglin, 1983) which sweeps out row and column medians from a two-way table to expose each datum composed of a common median plus individual row and column effects. Tukey (1977) also proposed methods for estimating the linear relationships in the data that are robust in the presence of ►outliers by fitting regression lines through the median points of three bins for example, as well as methods for smoothing noisy time series to expose the underlying trends. These methods aim to reveal the true nature of the underlying data, counteracting the effect of outliers, and therefore providing a more robust and accurate picture of the data prior to model building and hypothesis testing.

Beyond the basic EDA methods advocated by Tukey (1977), which pre-dated the current era of powerful computational methods, in recent years a greater arsenal of EDA methods have become available to assist with the exploration of a dataset and to reveal the relationships that may lead to a more successful model development phase. These methods include cluster analysis (Hair 2006), ►multidimensional scaling (Borg and Groenen 2005), and methods such as self-organizing feature maps (Kohonen 1988; Deboeck and Kohonen 1998) that reduce high dimensional data to low dimensional maps of the most prominent features.

The approach that Tukey proposed in 1977 seems commonsense nowadays – before building a model or testing hypotheses, take the time to understand the data, to verify its accuracy, and to explore it with a view to generating the right questions to ask a model. At the time though, it was a significant departure from the prevailing statistical practices of confirmatory data analysis. These days, it is a philosophy that is underscored in more recent developments in ►data mining (Maimon and Rokach 2005), where the importance of EDA as a pre-processing steps is seen as critical to avoid the "garbage in – garbage out" consequences of blindly developing models (see Pyle 1999).

About the Author

Kate Smith-Miles is a Professor and Head of the School of Mathematical Sciences at Monash University in Australia.

Prior to commencing this role in January 2009, she held a Chair in Engineering at Deakin University (where she was Head of the School of Engineering and Information Technology from 2006–2008, and a Chair in Information Technology at Monash University, where she worked from 1996–2006). She has published 2 books on neural networks and data mining applications, and over 200 refereed journal and international conference papers in the areas of neural networks, combinatorial optimization, intelligent systems and data mining. She has been a member of the organizing committee for over 50 international data mining and neural network conferences, including several as chair. Professor Smith-Miles was Chair of the IEEE Technical Committee on Data Mining (2007–2008).

Cross References

►Data Analysis
►Interactive and Dynamic Statistical Graphics
►Stem-and-Leaf Plot
►Summarizing Data with Boxplots

References and Further Reading

Borg I, Groenen PJF (2005) Modern multidimensional scaling: theory and applications. Springer-Verlag, New York
Deboeck GJ, Kohonen TK (1998) Visual explorations in finance. Springer-Verlag, New York
Emerson JD, Hoaglin DC (1983) Analysis of two-way tables by medians. In: Hoaglin DC, Mosteller F, Tukey JW (eds) Understanding robust and exploratory data analysis Wiley, New York, pp 165–210
Hair JF (2006) Multivariate data analysis, Prentice-Hall, New Jersey
Kohonen T (1989) Self-organization and associative memory. Springer-Verlag, Berlin
Maimon OZ, Rokash L (2005) Data mining and knowledge discovery handbook, Springer, New York
Pyle D (1999) Data preparation for data mining, Morgan Kaufmann, San Francisco
Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading, MA
Velleman PF, Hoaglin DC (1981) The ABC's of EDA. Duxbury Press, Boston, MA

Exponential and Holt-Winters Smoothing

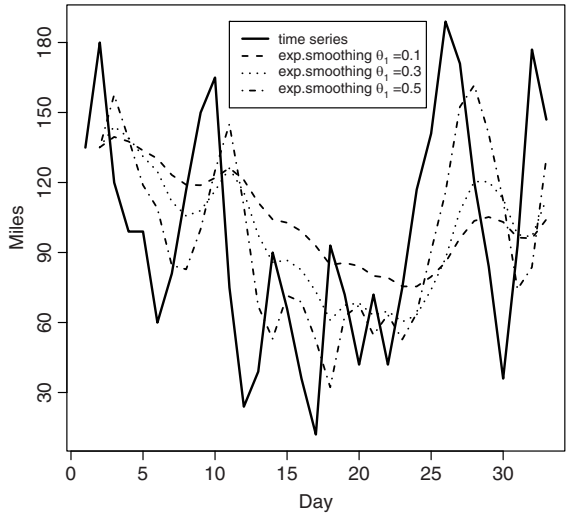
Roland Fried[1], Ann Cathrice George[2]
[1]Professor
TU Dortmund University, Dortmund, Germany
[2]TU Dortmund University, Dortmund, Germany

Exponential smoothing techniques are simple tools for smoothing and forecasting a time series (that is, a sequence of measurements of a variable observed at equidistant points in time). Smoothing a time series aims at eliminating the irrelevant noise and extracting the general path followed by the series. Forecasting means prediction of future values of the time series. Exponential smoothing techniques apply recursive computing schemes, which update the previous forecasts with each new, incoming observation. They can be applied online since they only use past observations which are already available at the corresponding point in time. Although exponential smoothing techniques are sometimes regarded as naive prediction methods, they are often used in practice because of their good performance, as illustrated e.g. in Chapter 4 of Makridakis et al. (1998). We speak of exponential smoothing techniques in plural because different variants exist which have been designed for different scenarios.

Suppose we observe a variable y at equidistant points in time, which are denoted by $t \in \mathbb{N}$. The sequence of measurements of y is called a time series and denoted by $(y_t : t \in \mathbb{N})$. *Simple exponential smoothing* computes the smoothed value $\tilde{y}_t$ at time t according to the following recursive scheme:

$$\tilde{y}_t = \alpha y_t + (1-\alpha)\tilde{y}_{t-1}, \qquad (1)$$

where $\alpha \in (0,1)$ is a *smoothing parameter*. The smaller we choose the value of α, the less weight is given to the most recent observation and the more weight is given to the smoothed value $\tilde{y}_{t-1}$ at the previous time $t-1$. As a



Exponential and Holt-Winters Smoothing. Fig. 1 Daily sea miles covered by Columbus on his first passage to America in 1492 and outputs of simple exponential smoothing with different smoothing parameters

consequence, the series $(\tilde{y}_t : t \in \mathbb{N})$ will be smoother. This is illustrated in Fig. 1, which depicts the number of sea miles covered by Columbus on each day of his first passage to America, as well as the results of exponential smoothing for several values of the smoothing parameter $\alpha \in \{0.1, 0.3, 0.5\}$. Obviously, the smaller we choose the weight α of the most recent observation, the less variable (smoother) is the resulting sequence of smoothed values.

Recursion (1) needs a starting value $\tilde{y}_k$, which can be obtained for instance as the average of the first k observations $y_1, \ldots, y_k$, where $k \geq 1$ is the length of a period used for initialization. Recursion (1) is run for the times $t = k+1, \ldots$ thereafter.

Successive replacement of $\tilde{y}_{t-1}$ by $\alpha y_{t-1} + (1-\alpha)\tilde{y}_{t-2}$, of $\tilde{y}_{t-2}$ by $\alpha y_{t-2} + (1-\alpha)\tilde{y}_{t-3}$, etc., in (1) allows to express $\tilde{y}_t$ as a weighted average of the previous observations $y_{t-1}, \ldots, y_{k+1}$ and the starting value $\tilde{y}_k$:

$$\tilde{y}_t = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \ldots + \alpha(1-\alpha)^{t-k-1} y_{k+1} + (1-\alpha)^{t-k} \tilde{y}_k \,. \tag{2}$$

Note that the weights $\alpha, \alpha(1-\alpha), \ldots, \alpha(1-\alpha)^{t-k-1}, (1-\alpha)^{t-k}$ sum to 1. Equation 2 shows that the weight of the preceding observation y_{t-h} is $\alpha(1-\alpha)^h$ and thus decreases geometrically in the time lag h, which stands in the exponent. This explains the name exponential smoothing.

Simple exponential smoothing can be justified by the assumption that the mean level of the time series is locally almost constant at neighboring points in time. Applying weighted least squares for estimation of the level at time t from the observations y_{t-h}, $h = 0, 1, \ldots$, available at time t (and the starting value) with geometrically decreasing weights $\alpha(1-\alpha)^h$ leads to the above Eqns. 1 and 2. The assumption of an almost constant level also explains how to use recursion (1) to predict future values $y_{t+1}, \ldots$. The h-step ahead forecast $\hat{y}_{t+h|t}$ of y_{t+h} from the observations up to time t, which can be derived from simple exponential smoothing, equals the smoothed value at time t,

$$\hat{y}_{t+h|t} = \tilde{y}_t, \qquad h = 1, 2, \ldots \,. \tag{3}$$

This simple prediction does not depend on the forecast horizon h. This is reasonable for a time series with a constant level, but it will result in bad forecasts for series with a trend. *Double exponential smoothing* incorporates a trend variable b_t representing the local slope of the trend at time t into the recursions to overcome this deficiency:

$$\begin{aligned} \tilde{y}_t &= \alpha\, y_t + (1-\alpha)\,(\tilde{y}_{t-1} + b_{t-1}) \\ b_t &= \beta\,(\tilde{y}_t - \tilde{y}_{t-1}) + (1-\beta)\, b_{t-1}. \end{aligned} \tag{4}$$

b_t can be regarded as an estimate of the local slope of the trend at time t. Now this slope is assumed to be locally almost constant at neighboring points in time. The increment $\tilde{y}_t - \tilde{y}_{t-1}$ of the estimated level is a simple and highly variable estimate of the slope at time t. The parameters $\alpha \in (0,1)$ and $\beta \in (0,1)$ are smoothing parameters regulating the amount of smoothing applied for the calculation of the smoothed slope b_t and the smoothed level $\tilde{y}_t$ at time t.

Starting values $\tilde{y}_k$ and b_k for the recursions (4) can be obtained by fitting an ordinary least squares regression line to an initial period, as described by Bowerman et al. (2005). Regressing y_t versus the time t, for $t = 1, \ldots, k$, yields an intercept $\hat{a}_0$ and a slope $\hat{b}_0$ resulting in

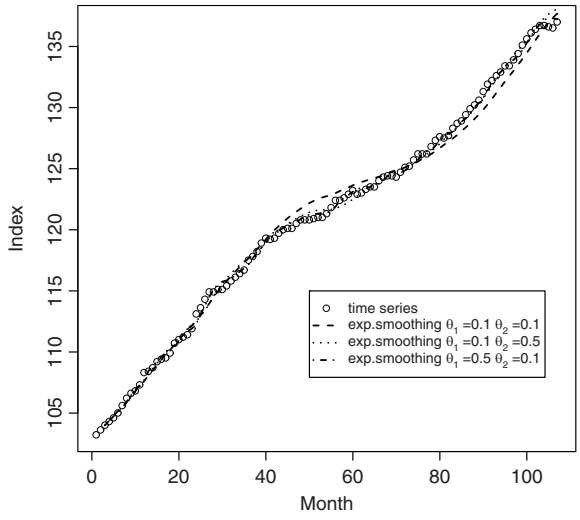
$$\begin{aligned} \tilde{y}_k &= \hat{a}_0 + \hat{b}_0 k\,, \\ b_k &= \hat{b}_0. \end{aligned}$$

An h-step-ahead forecast of y_{t+h} given the data up to time t can be obtained from recursion (4) as

$$\hat{y}_{t+h|t} = \tilde{y}_t + h b_t, \qquad h = 1, 2, \ldots \,. \tag{5}$$

These forecasts form a straight line, starting at $\tilde{y}_t$ and with slope equal to the most recent estimate b_t.

Figure 2 illustrates the results of double exponential smoothing applied to a time series representing the monthly consumer price index in Spain from January 1993 till December 2001, as well as the results of double exponential smoothing using several combinations of smoothing parameters. If values of α and β about 0.5 are chosen for these data, we get quite good 1-step ahead predictions,



Exponential and Holt-Winters Smoothing. Fig. 2 Monthly consumer price index in Spain from 1/1993 till 12/2001 and outputs of double exponential smoothing with different smoothing parameters

whereas the sequence of predictions is too stiff and does not adapt quickly to the bends of this time series if both α and β are small.

The *seasonal Holt-Winters method* is a further extension of the above procedure to incorporate periodic seasonality. For this we decompose the smoothed values $\tilde{y}_t$ into the sum of a trend component ℓ_t and a seasonal component s_t,

$$\tilde{y}_t = \ell_t + s_t.$$

The trend - consisting of the local level ℓ_t and the local slope b_t - and the seasonal component s_t are smoothed separately by the recursions

$$\begin{aligned} \ell_t &= \alpha\,(y_t - s_{t-m}) + (1-\alpha)\,(\ell_{t-1} + b_{t-1}) \\ b_t &= \beta\,(\ell_t - \ell_{t-1}) + (1-\beta)\,b_{t-1} \\ s_t &= \gamma\,(y_t - \ell_t) + (1-\gamma)\,s_{t-m}, \end{aligned} \tag{6}$$

with m being the period of the seasonality. Note that ℓ_t is calculated from the observations corrected for seasonality. h-step ahead forecasts are obtained by means of

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t-m+i}, \qquad h = jm + i,$$
$$i = 1, 2, \ldots, m, \quad j = 0, 1, \ldots.$$

The performance of all these exponential smoothing techniques depends on the smoothing parameters α, β and γ, which are sometimes set arbitrarily to values between 0.05 and 0.3. Alternatively, they can be chosen by minimizing the sum of the squared 1-step ahead prediction errors $\sum_{t=k+1}^{n}(y_t - \hat{y}_{t-1|1})^2$ or another error criterion.

Many modifications and improvements of exponential smoothing schemes have been suggested in the literature. Robustness against aberrant values (►outliers) is an important issue because these have large effects on the outcome of exponential smoothing. Gelper et al. (2010) modify the recursions and replace the incoming observations by a cleaned value whenever it lies outside some prediction bounds, using the Huber function.

There is a direct relation between exponential smoothing techniques and certain ARIMA models and state space models, like local constant or local linear models, for which an appropriately designed exponential smoothing technique leads to optimum forecasts in the mean square error sense. This connection can be used for selecting an adequate exponential smoothing scheme, see Ord et al. (1997) or Hyndman et al. (2002). Model selection procedures are important in the context of exponential smoothing since besides the schemes for additive seasonal and trend components outlined above there are also schemes for multiplicative components and for combinations of both, see Hyndman and Khandakar (2008). Gardner (2008) provides an overview on recent work in the field of exponential smoothing, see also Hyndman et al. (2008).

Cross References

►Business Forecasting Methods
►Forecasting: An Overview
►Moving Averages
►Trend Estimation

References and Further Reading

Bowerman B, O'Connell R, Koehler A (2005) Forecasting, time series, and regression, Thomson Books Cole, Belmont, CA
Gardner ES (2008) Exponential smoothing: the state of the art. Int J Forecast 22:637–666
Gelper S, Fried R, Croux C (2010) Robust forecasting with exponential and holt-winters smoothing. J Forecast 29(3):285–300
Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: the forecast package for R. J Stat Softw 27:1–22
Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008) Forecasting with exponential smoothing: the state space approach. Springer-Verlag, Berlin
Hyndman RJ, Koehler AB, Snyder RD, Grose S (2002) A state-space framework for automatic forecasting using exponential smoothing methods. Int J Forecast 18:439–454
Makridakis S, Wheelwright S, Hyndman R (1998) Forecasting, methods and applications, 3rd edn. Wiley, Hoboken, NJ
Ord JK, Koehler AB, Snyder RD (1997) Estimation and prediction for a class of dynamic nonlinear statistical models. J Am Stat Assoc 92:1621–1629

Exponential Family Models

Rolf Sundberg
Professor of Mathematical Statistics
Stockholm University, Stockholm, Sweden

An exponential family is a parametric model for a data set y whose distribution can be expressed by a density (or for discrete data: probability function) of type

$$f(y;\,\theta) = e^{\theta'\,t(y)}\,h(y)\,/C(\theta)$$

where $\theta'\,t$ is the scalar product of a k-dimensional *canonical parameter* vector θ and a k-dimensional *canonical statistic* $t = t(y)$, that is

$$\theta'\,t = \sum_{j=1}^{k} \theta_j\,t_j(y),$$

and the two factors C and h are two functions, the former interpretable as a norming constant (integral or sum). As a first simple example, consider a sample $y = (y_1, \ldots, y_n)$

from an exponential distribution with intensity parameter $\theta > 0$, whose density is

$$f(y;\theta) = \theta^n \, e^{-\theta \sum y_i}, \qquad \text{all } y_i > 0$$

and which is essentially already in exponential family form, with $t(y) = -\sum y_i$, $h(y) = 1$ if all $y_i \geq 0$, else $h(y) = 0$, and $C(\theta) = \theta^{-n}$. As a second simple example, let y be the number of successes in n Bernoulli trials with a probability $0 < p < 1$ for success, The probability for $y = 0, 1, \ldots, n$ is

$$f(y;p) = \binom{n}{y} p^y \, (1-p)^{n-y} = e^{\theta y} \binom{n}{y} / (1+e^{\theta})^n,$$

when parametrized by $\theta = \log \frac{p}{1-p}$ (the logit transform of p) instead of p. A sample from the Gaussian (Normal) distribution, with its two parameters, is an example with $k = 2$. Its canonical statistic has the components $\sum y_i$ and $\sum y_i^2$ (except for constant factors).

Very many statistical models have these features in common. Other distributions providing basic examples of exponential families are the Poisson, the Geometric, the Multinomial, the Multivariate normal, the Gamma and the Beta families. Of more statistical importance, however, are all those statistical models which can be constructed with these building blocks without leaving the exponential families: the Gaussian linear models (i.e. Linear regression models and ANOVA type models), Logistic regression models (see ►Logistic Regression), Multinomial or Poisson-based log-linear models for contingency tables, Graphical models (discrete or Gaussian, specified by conditional independencies), and the basic setting for ►Generalized linear models. Somewhat more specialized examples are Rasch's models for item analysis (in education testing), some basic models for spatial ►Poisson processes, without or with spatial interaction, and exponential random graph models (ERGM) for social networks.

We think of the vector t and the parameter θ as both in effect k-dimensional (not less). This implies that t is minimal sufficient for θ. To some extent, the importance of these models is explained by a general characterization of a distribution family with a minimal sufficient statistic of dimension independent of the sample size, when the latter increases (and satisfying a couple of other regularity conditions), as necessarily being an exponential family.

We say that the family is *full* if the canonical parameter space is maximal, that is if it contains all possible θ-values. A full exponential family in its canonical parametrization has several nice properties. Some families with pathological properties at the boundary of the parameter space are excluded by the (not quite necessary) requirement that the canonical parameter space be open. Such families are called *regular*, and most families in practical use satisfy this demand. Here is a selection of properties of full or regular exponential families:

- $\log C$ is strictly convex and infinitely differentiable, and its first and second derivatives are the mean value vector $\mu_t(\theta)(= E_\theta(t))$ and the variance-covariance matrix $V_t(\theta)$ for t.
- The parameter space for θ is a convex subset of R^k.
- The log-likelihood function $\log L(\theta)$ is a strictly concave function.
- The maximum likelihood estimate is the unique root of the likelihood equation system (when a maximum exists – a concave function can have an infinite supremum).
- The likelihood equation has the simple form $t = \mu_t(\theta)$.
- The observed and expected (Fisher) informations in θ are the same, namely $V_t(\theta)(= \mathrm{Var}_\theta(t))$, and its inverse is the large sample variance of the MLE of θ.
- An equivalent alternative parametrization of the model is by the mean value vector μ_t.
- If u and v are subvectors of t, which together form t, then the conditional model for v given u is also an exponential family, for each u, and with the canonical parameter θ_v (the subvector of θ corresponding to v).
- An equivalent alternative parametrization of the original family is the *mixed* parametrization, formed by $\mu_u(\theta)$ and θ_v together. This parametrization is information orthogonal, that is the information matrix is block diagonal.
- A parametric model having an unbiased efficient estimator, in the sense of equality in the ►Cramér–Rao inequality, is necessarily an exponential family in mean value parametrization, with its mean value parameter $\mu_t(\theta)$ estimated by the corresponding canonical statistic t.
- Exponential families are well suited for the saddle point approximation of the density of the maximum likelihood estimator (also called the p* formula or the magical formula). This approximation is typically quite efficient even for small samples. For the MLE $\hat{\theta}$ of the canonical parameter, with true value θ_0, the approximation is

$$f(\hat{\theta};\theta_0) \approx \frac{\sqrt{\det\{V_t(\hat{\theta})\}}}{\sqrt{2\pi}} \frac{L(\theta_0)}{L(\hat{\theta})},$$

  where $\det\{\}$ denotes the determinant of the information matrix (that was V_t for θ).

The properties listed above indicate that full exponential families form nice statistical models. In some applications, however, we need to go a bit outside this model class. One type of such situations goes under the natural name Curved exponential families. In this case, the family

is not full, but the canonical parameter is specified to lie in a curved subspace of the canonical parameter space. Important examples are the Behrens–Fisher model (two normal samples with the mean values prescribed to be the same, but not their variances), and the model called Seemingly unrelated regressions (the SUR model, popular in econometrics). To some extent the nice properties of the full family remain, but in important respects they do not. The minimum sufficient statistic is of a higher dimension than the parameter of the curved model, and this typically causes problems. For example, the likelihood equations do not always have a unique root in the models mentioned above.

The number of parameters in a multivariate Gaussian distribution model may be reduced by setting to zero a correlation or covariance between two of the variables. However, this would only seemingly be a simplification, because the model would turn into a curved family, and the minimal sufficient statistic would not be reduced. Instead, for a real simplification the zero should be inserted among the canonical parameters, that is in the inverse of the covariance matrix, corresponding to the specification of a conditional independence (so-called covariance selection modeling).

Another type of situation is when an exponential family is thought of as a model for some data x, but these data are not observable in the form desired. Instead, only some function $y = y(x)$ is observed, implying a loss of information. Such situations are known as incomplete data. Examples are grouped and censored data, data from a mixture of two or more distributions, non-Gaussian distributed x observed with Gaussian noise, missing data in multivariate analysis. Some specific distributions can be described in this way, for example the Negative binomial, the Cauchy and the t distributions. The theory starts from the observation that the conditional distribution of x given y is an exponential family that only differs from that for x in its norming constant, say $C_y(\theta)$ instead of $C(\theta)$. It follows that the density for y can be written $f(y, \theta) = C_y(\theta)/C(\theta)$, and next that the likelihood equations are $E_\theta(t|y) = \mu_t(\theta)$, which means that the complete data statistic t has been replaced by its conditional expected value, given the observed data y. Similarly, the Fisher information is now $\mathrm{Var}_\theta(E_\theta(t|y))$, instead of the complete data formula $\mathrm{Var}_\theta(t)$.

The unique root property does not hold for the incomplete data likelihood equations. Typically an iterative method of solution is needed. The EM algorithm is often well suited for this, and is particularly simple for incomplete data from exponential families: Given $\theta^{(i)}$ after i steps, calculate $E_{\theta^{(i)}}(t|y)$, insert its value in the likelihood equation (left hand side above, as a replacement for t), and solve for θ, to get $\theta^{(i+1)}$. The rate of convergence is determined by the loss of information relative to the complete data, as calculated from the information matrices.

Exponential families with $t(y_i) = y_i$ for an individual observation i are called linear. Examples are the Binomial, the Poisson, the Exponential, and the Gaussian with known variance. Such families play a basic role in the class of *Generalized Linear Models*. In such a model, the observations y_i follow a linear exponential family, with varying parameter θ_i. The form of the family relates θ_i with the mean value μ_{y_i}. Furthermore, μ_{y_i} is described by some predictor or regressor variables, via a link function.

- Some jointly observed explanatory (continuous or class) variables, x_i say, are tried to explain (linearly) the systematic part of the variability in y, via a linear predictor function $\eta(x) = \beta^T x$ with unknown coefficients β.
- The mean value μ_y need not be the linear predictor (as in a linear Gaussian model), but the mean value is linked to the predictor η by a specified *link function* g, $\eta = g(\mu)$.
- The observations y_i follow a linear exponential family, with varying parameter θ_i. The form of the family relates θ_i with the mean value μ_{y_i}.

When the link is the *canonical link* the model is a full exponential family, otherwise it is a curved family. Sometimes an additional variance parameter must be introduced (as with the Gaussian). In that case, the model will at least be a (full or curved) exponential family for each fixed value of the variance parameter.

About the Author

For biography *see* the entry ►Chemometrics.

Cross References

►Beta Distribution
►Gamma Distribution
►Generalized Linear Models
►Logistic Regression
►Multinomial Distribution
►Multivariate Normal Distributions
►Poisson Distribution and Its Application in Statistics
►Random Permutations and Partition Models
►Sequential Probability Ratio Test
►Statistical Distributions: An Overview
►Sufficient Statistics

References and Further Reading

Barndorff-Nielsen OE (1978) Information and exponential families. Wiley, Chichester

Barndorff-Nielsen OE, Cox DR (1994) Inference and asymptotics. Chapman & Hall, London
Brazzale AR, Davison AC, Reid N (2007) Applied asymptotics: case studies in small-sample statistics. Cambridge University Press, Cambridge
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Roy Stat Soc B 39:1–38
Efron B (1978) The geometry of exponential families. Ann Stat 6:362–376
Sundberg R (1974) Maximum likelihood theory for incomplete data from an exponential family. Scand J Stat 1:49–58

Extreme Value Distributions

Isabel Fraga Alves[1], Cláudia Neves[2]
[1]Associate Professor, CEAUL, DEIO Faculty of Sciences University of Lisbon, Lisbon, Portugal
[2]Assistant Professor, UIMA University of Aveiro, Aveiro, Portugal

Introduction

Extreme Value distributions arise as limiting distributions for maximum or minimum (*extreme values*) of a sample of independent and identically distributed random variables, as the sample size increases. Extreme Value Theory (EVT) is the theory of modelling and measuring events which occur with very small probability. This implies its usefulness in risk modelling as risky events per definition happen with low probability. Thus, these distributions are important in statistics. These models, along with the Generalized Extreme Value distribution, are widely used in risk management, finance, insurance, economics, hydrology, material sciences, telecommunications, and many other industries dealing with extreme events. The class of Extreme Value Distributions (EVD's) essentially involves three types of extreme value distributions, types I, II and III, defined below.

Definition 1 (Extreme Value Distributions *for maxima*). *The following are the standard Extreme Value distribution functions:*

(i) Gumbel (type I): $\Lambda(x) = \exp\{-\exp(-x)\}$, $x \in \mathbb{R}$;

(ii) Fréchet (type II): $\Phi_\alpha(x) = \begin{cases} 0, & x \le 0; \\ \exp\{-x^{-\alpha}\}, & x > 0,\ \alpha > 0; \end{cases}$

(iii) Weibull (type III):

$$\Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{\alpha}\}, & x \le 0,\ \alpha > 0; \\ 1, & x > 0. \end{cases}$$

The EVD families can be generalized with the incorporation of location (λ) and scale (δ) parameters, leading to

$$\Lambda(x;\lambda,\delta) = \Lambda((x-\lambda)/\delta),$$
$$\Phi_\alpha(x;\lambda,\delta) = \Phi_\alpha((x-\lambda)/\delta),$$
$$\Psi_\alpha(x;\lambda,\delta) = \Psi_\alpha((x-\lambda)/\delta), \quad \lambda \in \mathbb{R}, \delta > 0.$$

Among these three families of distribution functions, the type I is the most commonly referred in discussions of extreme values. Indeed, the Gumbel distribution $\{\Lambda(x;\lambda,\delta) = \Lambda((x-\lambda)/\delta); \lambda \in \mathbb{R}, \delta > 0\}$, is often coined "*the*" extreme value distribution.

Proposition 1 (Moments and Mode of EVD). *The mean, variance and mode of the EVD as in definition 1 are, respectively:*

(i) Gumbel – Λ: $E[X] = \gamma = 0.5772\ldots =$ *Euler's constant;* $Var[X] = \pi^2/6$; *Mode* $= 0$;

(ii) Fréchet – Φ_α: $E[X] = \Gamma(1-1/\alpha)$, *for* $\alpha > 1$; $Var[X] = \Gamma(1-2/\alpha) - \Gamma^2(1-1/\alpha)$, *for* $\alpha > 2$; *Mode* $= (1+1/\alpha)^{-1/\alpha}$;

(iii) Weibull – Ψ_α: $E[X] = -\Gamma(1+1/\alpha)$; $Var[X] = \Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)$; *Mode* $= -(1-1/\alpha)^{-1/\alpha}$, *for* $\alpha > 1$, *and Mode* $= 0$, *for* $0 < \alpha \le 1$;

here Γ *denotes the gamma function* $\Gamma(s) := \int_0^\infty x^{s-1}e^{-x}dx$, $s > 0$.

Definition 2 (Extreme Value Distributions *for minima*). *The standard converse EVD's for minima are defined as:* $\Lambda^*(x) = 1 - \Lambda(-x)$, $\Phi^*_\alpha(x) = 1 - \Phi_\alpha(-x)$ *and* $\Psi^*_\alpha(x) = 1 - \Psi_\alpha(-x)$.

Emil Gumbel

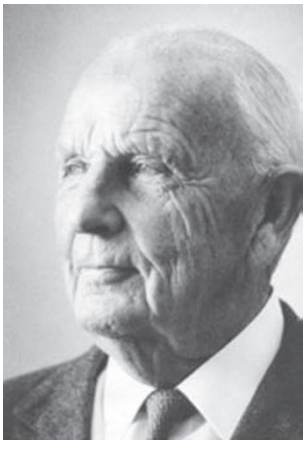
The Gumbel distribution, named after one of the pioneer scientists in practical applications of the Extreme

Value Theory (EVT), the German mathematician Emil Gumbel (1891–1966), has been extensively used in various fields including hydrology for modeling extreme events. Gumbel applied EVT on real world problems in engineering and in meteorological phenomena such as annual flood flows (Gumbel 1958):

> *"It seems that the rivers know the theory. It only remains to convince the engineers of the validity of this analysis."*

Maurice Fréchet

The EVD of type II was named after Maurice Fréchet (1878–1973), a French mathematician who devised one possible limiting distribution for a sequence of maxima, provided convenient scale normalization (Fréchet 1927). In applications to finance, the Fréchet distribution has been of great use apropos to the adequate modeling of market-returns which are often heavy-tailed.



Waloddi Weibull

The EVD of type III was named after Waloddi Weibull (1887–1979), a Swedish engineer and scientist well-known for his work on strength of materials and fatigue analysis (Weibull 1939). Even though the ▶Weibull distribution was originally developed to address the problems for minima arising in material sciences, it is widely used in many other areas thanks to its flexibility. If $\alpha = 1$, the Weibull distribution function for *minima*, Ψ_α^*, in Definition 2, reduces to the Exponential model, whereas for $\alpha = 2$ it mimics the Rayleigh distribution which is mainly used in the telecommunications field. Furthermore, Ψ_α^* resembles the Normal distribution when $\alpha = 3.5$.

Owing to the equality for a random sample $(X_1, \ldots, X_n)$

$$\min(X_1, \ldots, X_n) = -\max(-X_1, \ldots, -X_n)$$

it suffices to consider henceforth only the EVD's for maxima featuring in Definition 1. In probability theory and statistics, the Generalized Extreme Value (GEV) distribution is a family of continuous probability distributions developed under the extreme value theory in order to combine the Gumbel, Fréchet and Weibull families. The GEV distribution arises from the extreme value theorem (Fisher-Tippett 1928 and Gnedenko 1943) as the limiting distribution of properly normalized maxima of a sequence of independent and identically distributed (i.i.d.) random variables. Because of this, the GEV distribution is fairly used as an approximation to model the maxima of long (finite) sequences of random variables. In some fields of application the GEV distribution is in fact known as the Fisher-Tippett distribution, named after Sir Ronald Aylmer Fisher (1890–1962) and Leonard Henry Caleb Tippett (1902–1985) who recognized the only three possible limiting functions outlined above in Definition 1.

Extreme Value Theory and Max-Stability

Richard von Mises (1883–1953) studied the EVT in 1936, giving in particular the von Mises sufficient conditions on the hazard rate (assuming the density exists) in order to give a situation in which EVT behavior occurs, leading to one of the above three types of limit laws that is, giving an extremal domain of attraction $\mathcal{D}(G)$ for the extreme-value distribution G. Later on, and motivated by a storm surge in the North Sea (31 January–1 February 1953) which caused extensive flooding and many deaths, the Netherlands Government gave top priority to understanding the causes of such tragedies with a view to risk mitigation. Since it is the maximum sea level which is the danger, EVT became a Netherlands scientific priority. A relevant work in the field is the doctoral thesis of Laurens de Haan in 1970.

The fundamental extreme value theorem (Fisher-Tippett 1928; Gnedenko 1943) ascertains the Generalized

Extreme Value distribution in the von Mises-Jenkinson parametrization (von Mises 1936; Jenkinson 1955) as an unified version of all possible non-degenerate weak limits of partial maxima of sequences comprising i.i.d. random variables $X_1, X_2, \ldots$. That is:

Theorem 1 (Fisher-Tippett 1928; Gnedenko 1943). *If there exist normalizing constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that*

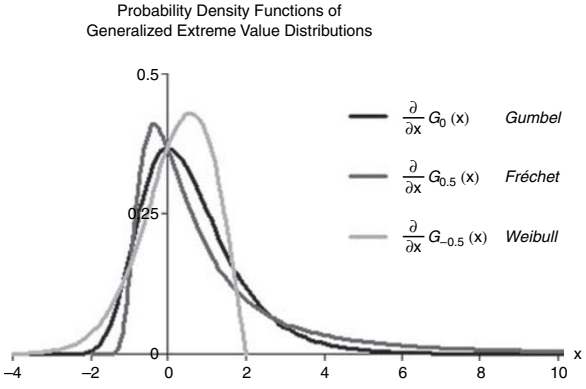
$$\lim_{n\to\infty} P\left\{a_n^{-1}\left(\max(X_1,\ldots,X_n) - b_n\right) \le x\right\} = G(x),$$

for some non-degenerate distribution function G, then it is possible to redefine the normalizing constants in such a way that

$$G(x) = G_\xi(x) := \exp\left(-(1+\xi x)^{-1/\xi}\right),$$

for all x such that $1+\xi x > 0$, with extreme value index $\xi \in \mathbb{R}$. Taking $\xi \to 0$, then $G_\xi(x)$ reduces to $\Lambda(x)$ for all $x \in \mathbb{R}$ (cf. Definition 1). Thus the distribution function F belongs to the domain of attraction of G_ξ, which is denoted by $F \in \mathcal{D}(G_\xi)$.

Remark 1 Note that, as $n \to \infty$, the $\max(X_1,\ldots,X_n)$ detached of any normalization converges in distribution to a degenerate law assigning probability one to the right endpoint of F, $x_F := \sup\{x : F(x) < 1\}$.



For $\xi < 0$, $\xi = 0$ and $\xi > 0$, the G_ξ distribution function reduces to Weibull, Gumbel and Fréchet distributions, respectively. More precisely,

$$\Lambda(x) \equiv G_0(x),$$
$$\Phi_\alpha(x) \equiv G_{1/\alpha}(\alpha(x-1)),$$

and

$$\Psi_\alpha(x) \equiv G_{-1/\alpha}(\alpha(1+x)).$$

For exhaustive details on EVD see Chapter 22 of Johnson et al. (1995).

Proposition 2 (Moments and Mode of GEV). *The mean, variance and mode of the GEV as in Theorem 1 are, respectively:*

$$E[X] = -\frac{1}{\xi}[1-\xi(1-\xi)], \text{for } \xi < 1; Var[X] = \frac{1}{\xi^2}[\Gamma(1-2\xi) - \Gamma^2(1-\xi)], \text{for } \xi < 1/2;$$
$$\text{Mode} = \frac{1}{\xi}[(1+\xi)^{-\xi} - 1], \text{for } \xi \neq 0.$$

Proposition 3 (Skewness of GEV). *The skewness coefficient of GEV distribution, defined as $skew_{G_\xi} := E[\{X - E[X]\}^3]/\{Var[X]\}^{3/2}$, is equal to zero at $\xi_0 \simeq -2.8$. Moreover, $skew_{G_\xi} > 0$, for $\xi > \xi_0$, and $skew_{G_\xi} < 0$, for $\xi < \xi_0$. Furthermore, for the Gumbel distribution, $skew_{G_0} \simeq 1.14$.*

The Fréchet domain of attraction contains distributions with polynomially decay tails. All distribution functions belonging to Weibull domain of attraction are light-tailed with finite right endpoint. The intermediate case $\xi = 0$ is of particular interest in many applied sciences where extremes are relevant, not only because of the simplicity of inference within the Gumbel domain $D(G_0)$ but also for the great variety of distributions possessing an exponential tail whether having finite right endpoint or not. In fact, separating statistical inference procedures according to the most suitable domain of attraction for the sampled distribution has become an usual practice. In this respect we refer to Neves and Fraga Alves (2008) and references therein.

Definition 3 (Univariate Max-Stable Distributions). *A random variable X with distribution function F is max-stable if there are normalizing sequences $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ such that the independent copies $X_1, X_2, \cdots, X_n$ satisfy the equality in distribution $\max(X_1, \cdots, X_n) \stackrel{d}{=} a_n X + b_n$. Equivalently, F is a max-stable distribution function if $[F(x)]^n = F((x-b_n)/a_n)$, all $n \in \mathbb{N}$.*

The class GEV, up to location and scale parameters, $\{G_\xi(x;\lambda,\delta) = G_\xi((x-\lambda)/\delta), \lambda \in \mathbb{R}, \delta > 0\}$, represents the only possible max-stable distributions.

Additional information can be found in Kotz and Nadarajah (2000), a monograph which describes in an organized manner the central ideas and results of probabilistic extreme-value theory and related extreme-value distributions – both univariate and multivariate – and their applications, and it is aimed mainly at a novice in the field. De Haan and Ferreira (2006) constitutes an excellent introduction to EVT at the graduate level, however requiring some mathematical maturity in regular variation, ▶point processes, empirical distribution functions, and Brownian motion. Reference Books in Extreme Value Theory

and in real world applications of EVD's and Extremal Domains of Attraction are: Embrechts et al. (2001), Beirlant et al. (2004), David and Nagaraja (2003), Gumbel (1958), Castillo et al. (2005) and Reiss and Thomas (2007).

Acknowledgments

This work has been partially supported by FCT/POCI 2010 project.

About the Authors

Isabel Fraga Alves is Associate Professor, PhD Thesis "Statistical Inference in Extreme Value Models", past-Coordinator of Center of Statistics and Applications of University of Lisbon (2006–2009), Elected Member of International Statistical Institute, Member of Bernoulli Society for Mathematical Statistics and Probability, Portuguese Statistical Society and Portuguese Mathematical Society.

Cláudia Neves is Assistant Professor at the Department of Mathematics of the University of Aveiro, member of the Institute of Mathematical Statistics and member of the Portuguese Statistical Society.

Cross References

►Generalized Extreme Value Family of Probability Distributions
►Generalized Weibull Distributions
►Location-Scale Distributions
►Statistical Distributions: An Overview
►Statistics of Extremes
►Weibull Distribution

References and Further Reading

Beirlant J, Goegebeur Y, Segers J, Teugels J (2004) Statistics of extremes: theory and applications. Wiley, England

Castillo E, Hadi AS, Balakrishnan N, Sarabia JM (2005) Extreme value and related models with applications in engineering and science. Wiley, Hoboken, NJ

David HA, Nagaraja HN (2003) Order statistics, 3rd edn. Wiley, Hoboken, NJ

de Haan L (1970) On regular variation and its application to the weak convergence of sample extremes. Math. Centre Tracts vol. 32, Mathematisch Centrum, Amsterdam

de Haan L, Ferreira A (2006) Extreme value theory: an introduction. Springer series in operations research and financial engineering, Boston

Embrechts P, Klüppelberg C, Mikosch T (2001) Modelling extremal events for insurance and finance, 3rd edn. Springer, Berlin

Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distribution of the largest and smallest member of a sample. Proceedings of the Cambridge Philosophical Society, vol 24. pp 180–190

Fréchet M (1927) Sur la loi de probabilit de l'écart maximum. Ann Soc Polon Math (Cracovie) 6:93–116

Gnedenko BV (1943) Sur la distribution limite du terme maximum d'une série aléatoire. Ann Math 44:423–453

Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York

Jenkinson AF (1955) The frequency distribution of the annual maximum (or minimum) values of meteorological elements. Quart J Roy Meteo Soc 81:158–171

Johnson NL, Balakrishnan N, Kotz S (1995) Continuous univariate distributions, vol 2., 2nd edn. Wiley, New York

Kotz S, Nadarajah S (2000). Extreme value distributions: theory and applications. Imperial College Press, London

Neves C, Fraga MI, Alves MI (2008) Testing extreme value conditions – an overview and recent approaches. In: Statistics of extremes and related fields, Jan Beirlant, Isabel Frag Alves, Ross Leadbetter (eds INE), REVSTAT - Statistical Journal, special issue, vol. 6, 1, 83–100

Reiss R-D, Thomas M (2001, 2007) Statistical analysis of extreme values, with application to insurance, finance, hydrology and other fields, 2nd and 3rd edn. Birkhuser Verlag, Basel

von Mises R, (1936) La distribution da la plus grande de n valeurs, Revue de l'Union Interbalkanique vol. 1 pp. 1–20

Weibull W (1939) A Statistical theory of the strength of materials. Ingeniors Vetenskaps Akadenien, Handlingar, vol. 151–3, pp 45–55

Extremes of Gaussian Processes

Sinisa Stamatovic
Professor
University of Montenegro, Podgorica, Montenegro

One of the oldest, most difficult, and most important problems in the theory of random processes has been the precise calculation of the probability

$$\mathbb{P}\left(\sup_{t\in[0,T]} X(t) > u\right) \tag{1}$$

where $X(t)$ is a random process. This problem is especially attractive for a Gaussian process. Even today, there is no explicit formula of the probability (1) in the general Gaussian situation, despite the fact that the set of finite dimensional distributions of a Gaussian process has a simple form. Because of this, there are multitudes of approximations and techniques of deriving approximations to (1), particularly when u is large. ►Gaussian processes appear in various fields (finance, hydrology, climatology, etc.) and finding the probability of attaining high level is particularly significant.

Extremes of Gaussian processes are well established in probability theory and mathematical statistics. Furthermore, besides the already mentioned problem, this theory addresses the following problems, among others: prediction of extremes, the moments of the number of crossings, and limit theorems for high excursions. The ruin probability with Gaussian process as input and the extremes of Gaussian process in random environment are two contemporary directions in the theory of extremes. To acquaint the reader with extremes of Gaussian processes we recommend monographs (Leadbetter et al., 1983; Lifshits 1995; Piterbarg 1996) and review articles (Alder 2000; Albeverio and Piterbarg 2006; Piterbarg and Fatalov 1995).

We will focus on one important part of the theory. J. Pickands III suggested a natural and elegant way of computing the asymptotic behavior of the probability (1) when $u \to \infty$, see Pickands (1969), Piterbarg and Fatalov (1995), Piterbarg (1996). Today his method is generalized for a wide class of Gaussian processes and fields. Approximation is based on the Bonferroni inequality, so Pickands' method is known as a double sum method. We will extract two central theorems. First, we will consider a stationary, centered Gaussian process $X(t),\ t \geq 0$. Assume also that its covariance function $r(t)$ satisfies the conditions

$$r(t) = 1 - |t|^{\alpha} + o(|t|^{\alpha}),\ t \to 0$$

for some $0 < \alpha \leq 2$ and

$$r(t) < 1, t > 0.$$

Let $\chi(t)$ be a fractional Brownian motion with a shift,

$$\mathbb{E}\chi(t) = -|t|^{\alpha}$$

and

$$cov(\chi(t), \chi(x)) = |t|^{\alpha} + |s|^{\alpha} - |t-s|^{\alpha}.$$

Theorem 1

$$\mathbb{P}\left(\sup_{t\in[0,p]} X(t) > u\right) = H_{\alpha} p u^{\frac{2}{\alpha}} \Psi(u)(1+o(1)),$$

as $u \to \infty$, *where*

$$\Psi(u) = \frac{1}{\sqrt{2\pi}} \int_u^{\infty} \exp\left(-\frac{z^2}{2}\right) dz,$$

$$H_{\alpha} = \lim_{T\to\infty} \frac{H_{\alpha}(T)}{T}, H_{\alpha}(T) = \mathbb{E}\exp\left(\sup_{t\in[0,T]} \chi(t)\right).$$

Let $X(t),\ t \in [0,T]$ be a centered Gaussian process with continuous trajectories. Denote by $\sigma^2(t)$ and $r(t,s)$ corresponding variance and correlation functions. Suppose that variance function attains its maximum at a unique point $t_0,\ t_0 \in (0,T)$. We introduce the following assumptions:

$\mathbb{D}1$ For some positive numbers a and β,

$$\sigma(t) = 1 - a|t-t_0|^{\beta}(1+o(1)), t - t_0 \to 0.$$

$\mathbb{D}2$ For the correlation function $r(t,s)$

$$r(t,s) = 1 - |t-s|^{\alpha}(1+o(1)), t \to t_0,\ s \to t_0,$$

where $0 < \alpha \leq 2$.

$\mathbb{D}3$ For some G and $\gamma > 0$

$$\mathbb{E}(X(t) - X(s))^2 \leq G|t-s|^{\gamma}.$$

Theorem 2 Assume that the conditions $\mathbb{D}1, \mathbb{D}2, \mathbb{D}3$ hold. Under these conditions:

$\mathbb{A}$. if $\beta > \alpha$, then

$$\mathbb{P}\left(\sup_{t\in[0,T]} X(t) > u\right) = \frac{2H_{\alpha}\Gamma(1/\beta)}{\beta a^{\frac{1}{\beta}}} u^{\frac{2}{\alpha}-\frac{2}{\beta}} \Psi(u)(1+o(1)),$$
$$u \to \infty,$$

$\mathbb{B}$. if $\beta = \alpha$, then

$$\mathbb{P}\left(\sup_{t\in[0,T]} X(t) > u\right) = 2H_{\alpha}^{a} \Psi(u)(1+o(1)),\ u \to \infty,$$

where

$$H_{\alpha}^{a}(S) = \mathbb{E}\exp\left(\sup_{t\in[-S,S]} (\chi(t) - a|t|^{\alpha})\right), H_{\alpha}^{a} = \lim_{S\to\infty} H_{\alpha}^{a}(S),$$

$\mathbb{C}$. if $\beta < \alpha$, then

$$\mathbb{P}\left(\sup_{t\in[0,T]} X(t) > u\right) = \Psi(u)(1+o(1)),\ u \to \infty.$$

Theorem 2 gives us asymptotic formulas for distributions of Kolmogorov-Smirnov type statistics that appear in goodness-of-fit tests for parameter families of distributions $F(x;\theta)$. Let θ_N be the maximum likelihood estimate for θ. It is well known that the empirical process

$$\zeta_N(x) = \sqrt{N}(F_N(x) - F_N(x,\theta_N))$$

weakly converges (under some regularity conditions) to some Gaussian centered process $\zeta(t)$ with a covariance function

$$cov(\zeta(x),\zeta(y)) = F^0(x) \wedge F^0(y) - F^0(x)F^0(y)$$
$$- \left(\int_{-\infty}^{x} G(z)dz\right)^T J^{-1} - \left(\int_{-\infty}^{y} G(z)dz\right)^T.$$

Here $F_0(x) = F(x,\theta_0),\ \theta_0$ is the true value of the parameter, vector $G(z)$ is of the form

$$G(z) = \left(\frac{\partial}{\partial\theta_i} \ln f(z,\theta_0), i = 1,2,\ldots,q\right)^T,$$

q is the dimension of the parameter θ, f is the density of F with respect to the Lebesgue measure,

$$J = \left(\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta_i} \ln f(z, \theta_0) \frac{\partial}{\partial \theta_j} \ln f(z, \theta_0) dz, \; i, j = 1, 2, \ldots, q \right)$$

is the Fisher's information matrix. By Theorem 2 we can find the asymptotic behavior of $\mathbb{P}\left(\sup_{x \in \mathbb{R}} \zeta(x) > u \right)$ and $\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\zeta(x)| > u \right)$ as $u \to \infty$ if we know the behavior of variance and covariance of the process $\zeta(x)$ near the maximum point of the variance. We will present an asymptotic formula obtained by V. Fatalov, for testing hypothesis

$$H_1 : F(x) \in \left\{ \Phi\left(\frac{x - a}{\sigma} \right), |a| < \infty, 0 < \sigma < \infty \right\},$$

both parameters are unknown. Suppose H_1 is true. Then the covariance function of the process $\zeta(x)$ is equal to

$$r(t, s) = t \wedge s - ts - \varphi(\Phi^{-1}(t))(\varphi(\Phi^{-1}(s)) \\ (1 + 0,5\Phi^{-1}(t)\Phi^{-1}(s)),$$

where $t = F(x), s = F(y)$, $\Phi(t)$ is the standard normal distribution function and $\varphi(t)$ is the standard normal density function. Thanks to the Theorem 2,

$$\mathbb{P}\left(\sup_{t \in [0,1]} |\zeta(t)| > u \right) = 2\sqrt{\frac{2\pi}{\pi - 2}} \exp\left(-\frac{2\pi}{\pi - 2} u^2 \right) \\ (1 + o(1)), \; u \to \infty.$$

Cross References

►Fisher-Tippett Theorem
►Gaussian Processes
►Stochastic Processes: Classification

References and Further Reading

Adler R (2000) On excursion sets, tube formulas and maxima of random fields. Ann Appl Probab 20:1–74

Albeverio S, Piterbarg V (2006) Mathematical methods and concepts for the analysis of extreme events, In: Albeverio S, Jentsch V, Kantz H (eds) Extreme events in nature and society, Springer-Verlag, Berlin, pp 47–68

Leadbetter MR, Lindgren G, Rootzen H (1983) Extremes and related properties of random sequences and processes. Springer-Verlag, Berlin

Lifshits M (1995) Gaussian random functions, Kluwer, Dordrecht, The Netherlands

Pickands J III (1969) Upcrossing probabilities for stationary Gaussian processes. Trans Amer Math Soc 145:51–73

Piterbarg V (1996) Asymptotic methods in the theory of Gaussian processes and fields. (Translations of Mathematical Monographs), vol 148. AMS, Providence, RI

Piterbarg V, Fatalov V (1995) The Laplace method for probability measures in Banach Spaces, Russian Math Surv 50:1151–1239

F

F Distribution

Enrique M. Cabaña
Professor
Universidad de la República, Monterideo, Uruguay

George W. Snedecor (1882–1974) promoted the development of statistics in the USA by contributing to the foundation of a department of statistics at Iowa State University, reputed to be the first one in the country, and helping with his writings the diffusion and application of Sir Ronald A. Fisher's (1890–1962) work on the ►analysis of variance and covariance (Fisher 1950, 1971).

Snedecor named "F" the distribution of the ratio of independent estimates of the variance in a normal setting as a tribute to Fisher, and now that distribution is known as the *Snedecor F*. It is a continuous skew probability distribution with range $[0, +\infty)$, depending on two parameters denoted ν_1, ν_2 in the sequel. In statistical applications, ν_1 and ν_2 are positive integers.

Definition of the F Distribution

Let Y_1 and Y_2 be two independent random variables distributed as chi-square, with ν_1 and ν_2 degrees of freedom, respectively (abbreviated $Y_i \sim \chi^2_{\nu_i}$). The distribution of the ratio $Z = \dfrac{Y_1/\nu_1}{Y_2/\nu_2}$ is called the F distribution with ν_1 and ν_2 degrees of freedom.

The notation $Z \sim \mathrm{F}_{\nu_1,\nu_2}$ expresses that Z has the F distribution with ν_1 and ν_2 degrees of freedom. The role of ν_1 and ν_2 in this definition is often emphasized by saying that ν_1 are the degrees of freedom of the numerator, and ν_2 are the degrees of freedom of the denominator.

Remark 1 Let $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ denote the usual estimator of the variance σ^2 obtained from $X_1, X_2, \ldots, X_n$ i.i.d. Normal(μ, σ^2), $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Since s^2 is distributed as $\sigma^2\chi^2_{n-1}/(n-1)$ (this means that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$), then the ratio of two such independent estimators of the same variance has the F distribution that, for this reason, is often referred to as *the distribution of the variance ratio*.

This leads to an immediate application of the F distribution: Assume that s_1^2 and s_2^2 are the estimators of the variances of two normal populations with variances σ_1^2 and σ_2^2 respectively, computed from independent samples of sizes n_1 and n_2 respectively. Then the ratio $F = s_1^2/s_2^2$ is distributed as $\frac{\sigma_1^2}{\sigma_2^2} F_{n_1-1,n_2-1}$.

When $\sigma_1^2 = \sigma_2^2$, $F \sim F_{n_1-1,n_2-1}$. On the other hand, when $\sigma_1^2 > \sigma_2^2$ or $\sigma_1^2 < \sigma_2^2$, F is expected to be respectively larger or smaller than a random variable with distribution F_{n_1-1,n_2-1}, and this suggests the use of F to test the null hypothesis $\sigma_1^2 = \sigma_2^2$: the null hypothesis is rejected when F is significantly large or small.

Remark 2 The joint probability density of the i.i.d. Normal$(0,1)$ variables $X_1, X_2, \ldots, X_n$ in $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is $\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|\boldsymbol{x}\|^2/2}$, with $\|\boldsymbol{x}\|^2 = \sum_{i=1}^n x_i^2$ equal to the Euclidean norm of $\boldsymbol{x}$. Since it depends on $\boldsymbol{x}$ only through its norm, it follows that the new coordinates $X_1^*, X_2^*, \ldots, X_n^*$ of $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ in any orthonormal basis of $\mathbf{R}^n$ have the same joint density and therefore are i.i.d. Normal$(0,1)$.

If $\mathcal{R}$ denotes the subspace generated by the first p vectors of the new basis, and $\mathcal{R}^\perp$ is its orthogonal complement, then the angle Ψ of $\boldsymbol{X}$ with $\mathcal{R}^\perp$ has tangent

$$\tan\Psi = \sqrt{\frac{Y_1^*}{Y_2^*}}, \text{ with } Y_1^* = \sum_{i=1}^p \left(X_i^*\right)^2 \sim \chi^2_p, Y_2^* = \sum_{i=p+1}^n \left(X_i^*\right)^2 \sim \chi^2_{n-p},$$

and hence

$$Z := \frac{n-p}{p}\tan^2\Psi \sim \mathrm{F}_{p,n-p}.$$

This geometrical interpretation of the F distribution is closely related to the F test in the analysis of variance (Scheffe 1959).

Important applications of F distribution include: F test for testing equality of two population variances, F test for fit of regression models, and Scheffe's method of multiple comparison.

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

The Density and Distribution Function of F_{ν_1,ν_2}

Let $Y_i \sim \chi^2_{\nu_i}, i = 1,2$. In order to compute the probability density of $Z = \frac{Y_1/\nu_1}{Y_2/\nu_2} \sim F_{\nu_1,\nu_2}$, introduce the random angle Ψ by $Z = \frac{\nu_2}{\nu_1}\tan^2\Psi$, and start by computing the distribution of $C = \cos^2\Psi = (1+\tan^2\Psi)^{-1} = \frac{\nu_2}{\nu_1 Z+\nu_2} = \frac{Y_2}{Y_1+Y_2}$:

$$\begin{aligned}\mathbf{P}\{C \le c\} &= \mathbf{P}\left\{Y_1 \ge \left(\tfrac{1}{c}-1\right)Y_2\right\}\\ &= \int_0^\infty f_2(y_2)\int_{\left(\frac{1}{c}-1\right)y_2}^\infty f_1(y_1)dy_1dy_2,\end{aligned}$$

where $f_i(t) = \frac{e^{-t/2}t^{\nu_i/2-1}}{2^{\nu_i/2}\Gamma(\nu_i/2)}$ is the density of the χ^2 distribution with ν_i degrees of freedom.

By replacing the analytical expressions of the χ^2 densities in

$$f_C(c) := \frac{d}{dc}\mathbf{P}\{C \le c\} = \int_0^\infty f_2(t)\frac{t}{c^2}f_1\left(\left(\tfrac{1}{c}-1\right)t\right)dt$$

one gets

$$\begin{aligned}f_C(c) &= \int_0^\infty \frac{e^{-t/2}t^{\nu_2/2-1}}{2^{\nu_2/2}\Gamma(\nu_2/2)}\\ &\quad\times\frac{t}{c^2}\frac{e^{-(1/c-1)t/2}(1/c-1)^{\nu_1/2-1}t^{\nu_1/2-1}}{2^{\nu_1/2}\Gamma(\nu_1/2)}dt\\ &= \frac{(1/c-1)^{\nu_1/2-1}}{c^2 2^{(\nu_1+\nu_2)/2}\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\\ &\quad\times\int_0^\infty e^{-t/2c}t^{(\nu_1+\nu_2)/2-1}dt\\ &= c^{\nu_2/2-1}(1-c)^{\nu_1/2-1}\frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}\\ &= \frac{c^{\nu_2/2-1}(1-c)^{\nu_1/2-1}}{B(\nu_1/2,\nu_2/2)}.\end{aligned}$$

This last expression, obtained by using the well-known relation $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x,y)}$ between Euler's Beta and Gamma functions, shows that $\cos^2\Psi$ has the ▶Beta distribution with parameters $(\nu_2/2, \nu_1/2)$ and consequently $\sin^2\Psi = 1-\cos^2\Psi$ has the Beta distribution with parameters $(\nu_1/2, \nu_2/2)$ and density $f_S(s) = \frac{s^{\nu_1/2-1}(1-s)^{\nu_2/2-1}}{B(\nu_1/2,\nu_2/2)}$.

The distribution function of $Z \sim F_{\nu_1,\nu_2}$ is

$$\begin{aligned}F_{\nu_1,\nu_2}(z) &= \mathbf{P}\{Z \le z\} = \mathbf{P}\left\{\tan^2\Psi \le \frac{\nu_1}{\nu_2}z\right\}\\ &= \mathbf{P}\left\{\cos^2\Psi \ge \frac{\nu_2}{\nu_1 z+\nu_2}\right\}\\ &= \mathbf{P}\left\{\sin^2\Psi \le \frac{\nu_1 z}{\nu_1 z+\nu_2}\right\} = \int_0^{\frac{\nu_1 z}{\nu_1 z+\nu_2}} f_S(s)ds\\ &= \frac{B\left(\frac{\nu_1 z}{\nu_1 z+\nu_2};\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)},\end{aligned}$$

where $B(t;a,b) = \int_0^t s^{a-1}(1-s)^{b-1}ds$ denotes the incomplete Beta function with parameters a, b evaluated in t. It may be noticed that the distribution function of F_{ν_1,ν_2} evaluated at z is the same as the distribution function of a $\text{Beta}(\nu_1/2, \nu_2/2)$ random variable evaluated at $\frac{\nu_1 z}{\nu_1 z+\nu_2}$.

By differentiating the c.d.f. the density of the F distribution is obtained:

$$\begin{aligned}f_{\nu_1,\nu_2}(z) &= \frac{\nu_1\nu_2}{(\nu_1 z+\nu_2)^2}f_S\left(\frac{\nu_1 z}{\nu_1 z+\nu_2}\right)\\ &= \frac{\sqrt{\nu_1^{\nu_1}\nu_2^{\nu_2}}}{zB(\nu_1/2,\nu_2/2)}\sqrt{\frac{z^{\nu_1}}{(\nu_1 z+\nu_2)^{\nu_1+\nu_2}}}.\end{aligned}$$

Figures 1 and 2 show graphs of f_{ν_1,ν_2} for several values of the parameters.

Some Properties of F Distribution

The moments of $Y \sim \chi^2_\nu$ are $\mathbf{E}Y^k = 2^k\frac{\Gamma(\frac{\nu}{2}+k)}{\Gamma(\frac{\nu}{2})}$ for $k > -\frac{\nu}{2}$, and infinite otherwise. Therefore, from the expression of $Z = \frac{\nu_2 Y_1}{\nu_1 Y_2}$ as the ratio of independent random variables $Y_i \sim \chi^2_{\nu_i}$ we get $\mathbf{E}Z^k = \left(\frac{\nu_2}{\nu_1}\right)^k \mathbf{E}Y_1^k\mathbf{E}Y_2^{-k} = \left(\frac{\nu_2^k}{\nu_1^k}\right)2^k\frac{\Gamma(\frac{\nu_1}{2}+k)}{\Gamma(\frac{\nu_1}{2})} \times 2^{-k}\frac{\Gamma(\frac{\nu_2}{2}-k)}{\Gamma(\frac{\nu_2}{2})} = \left(\frac{\nu_2}{\nu_1}\right)^k\frac{\Gamma(\frac{\nu_1}{2}+k)\Gamma(\frac{\nu_2}{2}-k)}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}$, provided $k < \nu_2/2$. If this last restriction does not hold, the moment is infinite. In particular,

$$\mathbf{E}Z = \frac{\nu_2}{\nu_2-2} \text{ for } \nu_2 > 2,$$

$$\mathbf{Var}Z = \frac{2(\nu_1+\nu_2-2)\nu_2^2}{\nu_1(\nu_2-2)^2(\nu_2-4)} \text{ for } \nu_2 > 4.$$

Other descriptive parameters are the mode $\frac{(\nu_1-2)\nu_2}{\nu_1(\nu_2+2)}$ for $\nu_1 > 2$, the skewness coefficient

$$\frac{2\sqrt{2}(2\nu_1+\nu_2-2)\sqrt{\nu_2-4}}{\sqrt{\nu_1(\nu_1+\nu_2-2)}(\nu_2-6)}$$

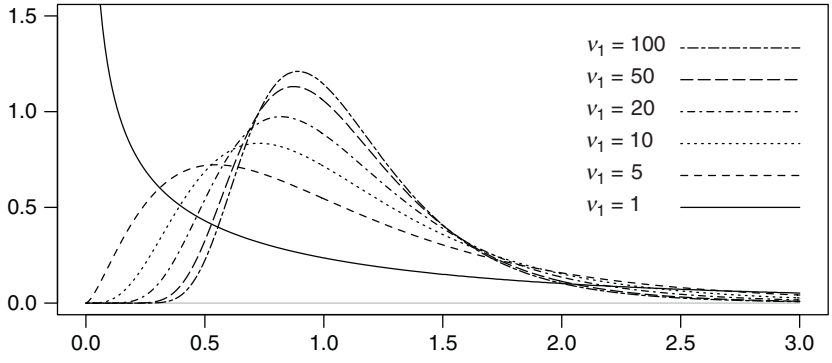
for $\nu_2 > 6$ and the kurtosis

$$\frac{3(\nu_2-4)\left(4(\nu_2-2)^2+\nu_1^2(\nu_2+10)+\nu_1(\nu_2-2)(\nu_2+10)\right)}{\nu_1(\nu_2-8)(\nu_2-6)(\nu_1+\nu_2-2)} - 3$$

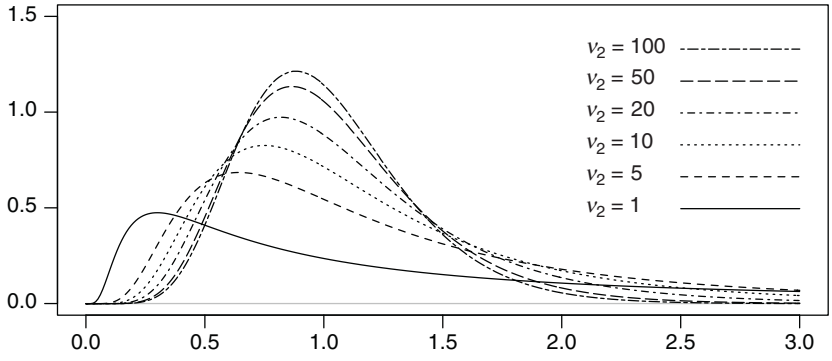
for $\nu_2 > 8$.

On Numerical Computations

There exist many tables of the F distribution, but the simpler way to obtain the numerical values of the density, the distribution function or its inverse, is to use the facilities provided by statistical software. The pictures here included and the numerical computations required by them were made by using the free software "R" (R Development Core Team 2008).



F Distribution. Fig. 1 Densities of F distribution with $\nu_1 = 1, 5, 10, 20, 50, 100$ and $\nu_2 = 20$



F Distribution. Fig. 2 Densities of F distribution with $\nu_1 = 20$ and $\nu_2 = 1, 5, 10, 20, 50, 100$

About the Author

Professor Enrique Cabaña is a member of the International Statistical Institute (elected in 1994) and founding President of the Latin American Regional Committee of The Bernoulli Society (1981–1983). He was Head of the Mathematical Centre of the Universidad de la República (1987–1990), Pro-Vice-Chancellor for Research of the same University (1999–2006) and Director of the Programme for the Development of Basic Sciences (PEDECIBA) of Uruguay (1997–2000). He has been teaching probability and statistics since 1958, mainly in Uruguay but also in Chile (1975–1977) and Venezuela (1978–1986) and his recent papers coauthored with Alejandra Cabaña develop several applications of L^2-techniques for the assessment of models based on transformations of stochastic processes.

Cross References

►Analysis of Variance
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Tests for Homogeneity of Variance

References and Further Reading

Fisher RA (1950) Contributions to mathematical statistics. Wiley, New York
Fisher RA (1971) Collected Papers of Fisher RA. In Bennet JH (ed) The University of Adelaide
R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
Scheffe H (1959) The analysis of variance. Wiley, New York

Factor Analysis and Latent Variable Modelling

David J. Bartholomew
Professor Emeritus of Statistics
London School of Economics and Political Science, London, UK

Background

Factor analysis was invented in 1904 by Professor Charles Spearman at University College London. Spearman was a psychologist and, for half century, factor analysis largely

remained the preserve of psychologists. Latent class analysis was developed by Paul Lazarsfeld at Columbia University in New York in the nineteen fifties and was designed for use by sociologists. Yet both of these techniques, and others like them, are essentially statistical and are now recognized as sharing a common conceptual framework which can be used in a wide variety of fields. It was not until the late nineteen thirties that statisticians, such as M. S. Bartlett, made serious contributions to the field.

Both factor analysis and latent class analysis are examples of the application of what would now be called latent variable models. Statistics deals with things that vary and in statistical theory such quantities are represented by random variables. In most fields these variables are observable and statistical analysis works with the observed values of such variables. But there are some important applications where we are interested in variables which cannot be observed. Such variables are called *latent* variables. In practice these often arise in the social sciences and include such things as human intelligence and political attitudes.

A latent variable model provides the link between the latent variables, which cannot be observed, and the manifest variables which can be observed. The purpose of the analysis is to determine how many latent variables are needed to explain the correlations between the manifest variable, to interpret them and, sometimes, to predict the values of the latent variables which have given rise to the manifest variables.

The Linear Factor Model

The basic idea behind factor analysis and other latent variable models is that of regression, or conditional expectation. We may regress each of the manifest (observed) variables on the set of latent variables (or factors). Thus, if we have p manifest variables, denoted by $x_1, x_2, \ldots x_p$ and q factors, denoted by $f_1, f_2, \ldots f_q$, the model may be written

$$x_i = \alpha_0 + \alpha_1 f_1 + \alpha_2 f_2 + \ldots \alpha_q f_q + e_i \qquad (i = 1, 2, \ldots, p) \quad (1)$$

where, without loss of generality, the fs are assumed to have zero means and unit standard deviations. The error term e_i is also assumed to have zero mean and standard deviation, σ_i. We often assume that all distributions are normal, in which case we refer to this as the *normal linear factor model*.

There are p linear equations here but the model cannot be fitted like the standard linear regression model (see ►Linear Regression Models) because the number of factors is unknown and the values of the f_s are not known, by definition. We, therefore, have to use indirect methods which depend on the fact that the correlation coefficients between the xs depend only on the αs and the σs. In practice, efficient computer programs are available which take care of the fitting.

The Latent Class Model

In a latent class model the manifest and latent variables are both categorical, often binary, instead of continuous. Thus the xs may consist of binary answers to a series of questions of the YES/NO variety. These are often coded 0 and 1 so that the manifest variable, x_i takes one of the two values 0 and 1. On the basis of these data we may wish to place individuals into one of several categories. In such cases the model is usually expressed in terms of probabilities. For example, for the ith manifest variable we may specify that

$$\Pr[x_i = 1] = \frac{\exp - \alpha_{i0} - \alpha_{i1} f}{1 + \exp - \alpha_{i0} - \alpha_{i1} f} \quad (2)$$

Because x_i is binary, the left hand side of the equation may also be written, $E(x_i)$ The reason for this somewhat strange expression is that probabilities necessarily lie between 0 and 1. The link with the linear expression of the previous section is made clearer if we write it in terms of the logit function. In that case we have

$$\text{logitE}[x_i] = \alpha_{i0} + \alpha_{i1} f. \quad (3)$$

This becomes a latent class model if we let f be a binary variable; this is a way of letting the probability on the left hand side of the equation take just two values.

Other Latent Variable Models

Prior to the last step, we actually had a latent profile model with one continuous factor, f. Further latent variables could have been added to the right hand side in exactly the same way as with the general linear factor model. Similarly, we could have had a continuous variable on the left hand side with discrete variables on the right hand side. Beyond this, in principle, there could be mixtures of continuous and/or categorical variables on both sides of the equation.

Much recent work is on what are called linear structural relations models where the interest is in the assumed (linear) relationships among the latent variables.

The Literature

There is an enormous literature on factor analysis and latent variable models, much of it very old and difficult to follow. This is not helped by the fact that much of the work has been published in books or journals appropriate to the disciplinary origins of the material and the level of mathematical expertise expected of the readers. One of the very few broad treatments from a statistical angle is given in:

Bartholomew, D.J. and Knott, M. (2011) *Latent Variable Models and Factor Analysis*, 2nd edition, Kendall's Library of Statistics 7, Arnold.

The references given there will lead on to many other aspects of the field, some of which have been touched on above.

About the Author

Past President of the Royal Statistical Society (1993–1995), David Bartholomew, was born in England in 1931. After undergraduate and postgraduate study at University College London, specializing in statistics, he worked for two years in the operational research branch of the National Coal Board. In 1957 he began his academic career at the University of Keele and then moved to the University College of Wales, Aberystwyth as lecturer, then senior lecturer in statistics. This was followed by appointment to a chair in statistics at the University of Kent in 1967. Six years later he moved to the London School of Economics as Professor of Statistics where he stayed until his retirement in 1996. During this time at the LSE he also served as Pro-Director for three years. He is a Fellow of the British Academy, a Member of the International Statistical Institute, and a Fellow of the Institute of Mathematical Statistics, He has authored, co-authored or edited about 20 books and about 120 research papers and articles, including the text *Latent Variable Models and Factor Analysis* (1987, Griffin; 2nd edition (with Martin Knott), Edward Arnold, 1999).

Cross References

►Correspondence Analysis
►Data Analysis
►Mixed Membership Models
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Principal Component Analysis
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Psychiatry, Statistics in
►Psychology, Statistics in
►Statistical Inference in Ecology
►Statistics: An Overview
►Structural Equation Models

References and Further Reading

Bartholomew DJ, Knott M (2011) Latent variable models and factor ananlysis: A unified approach, 3rd edn. Wiley Black well (in press)

Factorial Experiments

Kalina Trenevska Blagoeva
Associate Professor, Faculty of Economics
University "Ss. Cyril and Methodius", Skopje, Macedonia

Statistically designed experiments are an important tool in data analysis. The objective of such experimentation is to estimate the effect of each experimental factor on a response variable and to determine how the effect of one factor varies over the levels of other factors. Each measurement or observation is made on an item denoted as an *experimental unit*. Although some ideas of the several varying factors simultaneously appeared in England in the nineteenth century, the first major systematic discussion on factorial designs was given by Sir Ronald Fisher in his seminal book The Design of Experiments (Chap. 6) in 1935.

A *factorial experiment* is an experiment in which several factors (such as fertilizers or antibiotics) are applied to each experimental unit and each factor is applied at two, or more, levels. The levels may be quantitative (as with amounts of some ingredient) or qualitative (where the level refers to different varieties of wheat) but in either case are represented by elements of a finite set, usually by $0, 1, 2, \ldots, k_i - 1$ where the ith factor occurs at k_i levels. A factorial experiment in which t independent factors are tested, and in which the ith factor has k_i levels is labeled a $k_1 \times k_2 \times \cdots \times k_t$ factorial experiment. If $k_1 = k_2 = \cdots = k_t = k$, then the experiment is designated as a k^t *symmetrical factorial experiment*. An important feature of a complete factorial experiment is that all possible factor-level combinations are included in the design of the experiment.

Each controllable experimental variable, such as temperature or diet, in a factorial experiment is termed a *factor*. The *effect* of a factor on the response variable is the change in the average response between two experimental conditions. When the effect is computed as the difference between the average response at a given level of one factor and the overall average based on all of its levels after averaging over the levels of all the other factors, it is labeled the *main effect* of that factor. The difference in the effects of factors at different levels of other factors represents the *interaction* between factors. We can estimate the effect of each factor, independently of the others (the main effect), and the effect of the interaction of two (or more) factors (the interaction effect).

A factorial experiment allows for estimation of experimental error in two ways. The experiment can be replicated, or the sparsity-of-effects principle can often be

exploited. Replication is more common for small experiments and is a very reliable way of assessing experimental error. When the number of factors is large, replication of the design can become operationally difficult. In these cases, it is common to only run a single replicate of the design and to assume that factor interactions of more than a certain order (say, between three or more factors) are negligible. A *single replicate factorial design* is a factorial experiment in which every treatment combination appears precisely once in the design. As with any statistical experiment, the experimental runs in a factorial experiment should be randomized to reduce the impact that bias could have on the experimental result. In practice, this can be a large operational challenge.

Factorial experiments also can be run in block designs, where blocks refer to groups of experimental units or test runs (such as batches of raw material) that are more homogeneous within a block than between blocks. Combinations of the levels of two or more factors are defined as *treatment combinations*. If the number of treatment combinations is not too large, it is often possible to run the experiment in block designs in which some information is available within blocks on all factorial effects (i.e., main effects and interactions). Such effects are said to be partially confounded with blocks. However, factorial experiments with many factors, or with factors at many levels, involve large numbers of treatment combinations. The use of designs that require a number of replicates of each treatment combination then becomes impractical. To overcome this problem, designs using a single replicate are frequently used. Information on all or part of some of the factorial effects will consequently no longer be available from comparisons within blocks; these effects, or some components of them, will be said to be totally confounded with blocks.

Any number of factor levels can be used in a factorial experiment provided there is an adequate number of experimental units. However, the number of experimental runs required for three-level (or more) factorial designs will be considerably greater than for their two-level counterparts. Factorial designs are therefore less attractive if a researcher wishes to consider more than two levels. When the number of test runs required by a complete factorial experiment cannot be run due to time or cost constraints, a good alternative is to use fractional factorial experiments. These types of designs reduce the number of test runs.

Cross References

►Design of Experiments: A Pattern of Progress
►Interaction
►Research Designs
►Statistical Design of Experiments (DOE)

References and Further Reading

Box GEP, Hunter WG, Hunter JS (2005) Statistics for experimenters: design, innovation, and discovery, 2nd edn. Wiley, New York
Cox DR, Reid N (2000) The theory of the design of experiments. Chapman & Hall/CRC, London
Fisher R (1935) The design of experiments. Collier Macmillan, London
Mukherjee R, Wu CFJ (2006) A modern theory of factorial design. Springer Series in Statistics, Springer, New York
John A, Williams ER (1995) Cyclic and computer generated designs. Chapman & Hall, New York
Raktoe BL (1992) Factorial designs. Krieger Pub Co, Reprinted edition, Malabar, Florida

False Discovery Rate

John D. Storey
Associate Professor
Princeton University, Princeton, NJ, USA

Multiple Hypothesis Testing

In hypothesis testing, *statistical significance* is typically based on calculations involving ►p-values and Type I error rates. A p-value calculated from a single statistical hypothesis test can be used to determine whether there is statistically significant evidence against the null hypothesis. The upper threshold applied to the p-value in making this determination (often 5% in the scientific literature) determines the Type I error rate; i.e., the probability of making a Type I error when the null hypothesis is true. *Multiple hypothesis testing* is concerned with testing several statistical hypotheses simultaneously. Defining statistical significance is a more complex problem in this setting.

A longstanding definition of statistical significance for multiple hypothesis tests involves the probability of making one or more Type I errors among the family of hypothesis tests, called the *family-wise error rate*. However, there exist other well established formulations of statistical significance for multiple hypothesis tests. The Bayesian framework for classification naturally allows one to calculate the probability that each null hypothesis is true given the observed data (Efron et al. 2001; Storey 2003), and several frequentist definitions of multiple hypothesis testing significance are also well established (Shaffer 1995).

Soric (1989) proposed a framework for quantifying the statistical significance of multiple hypothesis tests based on

the proportion of Type I errors among all hypothesis tests called statistically significant. He called statistically significant hypothesis tests *discoveries* and proposed that one be concerned about the rate of false discoveries when testing multiple hypotheses. (A false discovery, Type I error, and false positive are all equivalent. Whereas the false positive rate and Type I error rate are equal, the false discovery rate is an entirely different quantity.) This false discovery rate is robust to the false positive paradox and is particularly useful in exploratory analyses, where one is more concerned with having mostly true findings among a set of statistically significant discoveries rather than guarding against one or more false positives. Benjamini and Hochberg (1995) provided the first implementation of false discovery rates with known operating characteristics. The idea of quantifying the rate of false discoveries is directly related to several pre-existing ideas, such as Bayesian misclassification rates and the positive predictive value (Storey 2003).

Applications

In recent years, there has been a substantial increase in the size of data sets collected in a number of scientific fields, including genomics, astrophysics, neurobiology, and epidemiology. This has been due in part to an increase in computational abilities and the invention of various technologies, such as high-throughput biological devices. The analysis of high-dimensional data sets often involves performing simultaneous hypothesis tests on each of thousands or millions of measured variables. Classical multiple hypothesis testing methods utilizing the family-wise error rate were developed for performing just a few tests, where the goal is to guard against any single false positive occurring. However, in the high-dimensional setting, a more common goal is to identify as many true positive findings as possible, while incurring a relatively low number of false positives. The false discovery rate is designed to quantify this type of trade-off, making it particularly useful for performing many hypothesis tests on high-dimensional data sets.

Hypothesis testing in high-dimensional genomics data sets has been particularly influential in increasing the popularity of false discovery rates (Storey and Tibshirani 2003). For example, DNA microarrays measure the expression levels of thousands of genes from a single biological sample. It is often the case that microarrays are applied to samples collected from two or more biological conditions, such as from multiple treatments or over a time course. A common goal in these studies is to identify genes that are differentially expressed among the biological conditions, which involves performing a hypothesis tests on each gene. In addition to incurring false positives, failing to identify truly differentially expressed genes is a major concern, leading to the false discovery rate being in widespread use in this area. In a notably different area of application, the false discovery rate was utilized in an astrophysics study to detect acoustic oscillations on the distribution of matter in present time, which had implications towards confirming the Big Bang theory of the creation of the universe (Lindsay et al. 2004). The body of scientific problems to which the false discovery rate is applied continues to grow.

Mathematical Definitions

Although multiple hypothesis testing with false discovery rates can be formulated in a very general sense (Storey 2007; Storey et al. 2007), it is useful to consider the simplified case where m hypothesis tests are performed with corresponding p-values $p_1, p_2, \ldots, p_m$. The typical procedure is to call hypotheses statistically significant whenever their corresponding p-values are less than or equal to some threshold t, where $0 < t \leq 1$. This threshold can be fixed or data-dependent, and the procedure for determining the threshold involves quantifying a desired error rate.

Table 1 describes the various outcomes that occur when applying this approach to determining which of the m hypothesis tests are statistically significant. Specifically, V is the number of Type I errors (equivalently false positives or false discoveries) and R is the total number of hypothesis tests called significant (equivalently total discoveries). The *family-wise error rate* (FWER) is defined to be

$$\text{FWER} = \mathbf{Pr}(V \geq 1),$$

and the *false discovery rate* (FDR) is usually defined to be (Benjamini and Hochberg 1995):

$$\text{FDR} = \mathbf{E}\left[\frac{V}{R \vee 1}\right] = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right]\mathbf{Pr}(R > 0).$$

The effect of "$R \vee 1$" in the denominator of the first expectation is to set $V/R = 0$ when $R = 0$. As demonstrated by Benjamini and Hochberg (1995), the FDR offers a less strict

False Discovery Rate. Table 1 Possible outcomes from m hypothesis tests based on applying a significance threshold $t \in (0,1]$ to their corresponding p-values

	Not significant (p-value $> t$)	Significant (p-value $\leq t$)	Total
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

multiple testing criterion than the FWER, allowing it to be more appropriate for some applications.

Two other false discovery rate definitions have been proposed in the literature, where the main difference is in how the $R = 0$ event is handled. These quantities are called the *positive false discovery rate* (pFDR) and the *marginal false discovery rate* (mFDR), and they are defined as follows (Storey 2003, 2007):

$$\text{pFDR} = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right],$$
$$\text{mFDR} = \frac{\mathbf{E}[V]}{\mathbf{E}[R]}.$$

Note that pFDR = mFDR = 1 whenever all null hypotheses are true, whereas FDR can always be made arbitrarily small because of the extra term $\mathbf{Pr}(R > 0)$. Some have pointed out that this extra term in the FDR definition may lead to misinterpreted results, and pFDR or mFDR offer more scientifically relevant values (Storey 2003; Zaykin et al. 1998), while others have argued that FDR is preferable because it allows for the traditional "strong control" criterion to be met (Benjamini and Hochberg 1995). All three quantities can be utilized in practice, and they are all similar when the number of hypothesis tests is particularly large.

Control and Estimation

There are two approaches to utilizing false discovery rates in a conservative manner when determining multiple testing significance. One approach is to fix the acceptable FDR level beforehand, and find a data-dependent thresholding rule so that the expected FDR of this rule over repeated studies is less than or equal to the pre-chosen level. This property is called *FDR control* (Benjamini and Hochberg 1995; Shaffer 1995). Another approach is to fix the p-value threshold at a particular value and then form a point estimate of the FDR whose expectation is greater than or equal to the true FDR at that particular threshold (Storey 2002). The latter approach has been useful in that it places multiple testing in the more standard context of point estimation, whereas the derivation of algorithms in the former approach may be less tractable. Indeed, it has been shown that the point estimation approach provides a comprehensive and unified framework (Storey et al. 2004).

For the first approach, (Benjamini and Hochberg 1995) proved that the algorithm below for determining a data based p-value threshold controls the FDR at level α when the p-values corresponding to true null hypotheses are independent and identically distributed (i.i.d.) Uniform(0,1). Other p-value threshold determining algorithms for FDR control have been subsequently studied (e.g., Benjamini and Liu 1999). This algorithm was originally introduced by Simes (1986) to control the FWER when all p-values are independent and all null hypotheses are true, although it also provides control of the FDR for any configuration of true and false null hypotheses.

FDR Controlling Algorithm (Simes, 1986; Benjamini and Hochberg, 1995)
1. Let $p_{(1)} \le \ldots \le p_{(m)}$ be the ordered, observed p-values.
2. Calculate $\widehat{k} = \max\{1 \le k \le m : p_{(k)} \le \alpha \cdot k/m\}$.
3. If $\widehat{k}$ exists, then reject null hypotheses corresponding to $p_{(1)} \le \ldots \le p_{(\widehat{k})}$. Otherwise, reject nothing.

To formulate the point estimation approach, let $\text{FDR}(t)$ denote the FDR when calling null hypotheses significant whenever $p_i \le t$, for $i = 1, 2, \ldots, m$. For $t \in (0,1]$, we define the following ▶stochastic processes based on the notation in Table 1:

$$V(t) = \#\{\text{true null } p_i : p_i \le t\},$$
$$R(t) = \#\{p_i : p_i \le t\}.$$

In terms of these, we have

$$\text{FDR}(t) = \mathbf{E}\left[\frac{V(t)}{R(t) \vee 1}\right].$$

For fixed t, Storey (2002) provided a family of conservatively biased point estimates of $\text{FDR}(t)$:

$$\widehat{\text{FDR}}(t) = \frac{\widehat{m}_0(\lambda) \cdot t}{[R(t) \vee 1]}.$$

The term $\widehat{m}_0(\lambda)$ is an estimate of m_0, the number of true null hypotheses. This estimate depends on the tuning parameter λ, and it is defined as

$$\widehat{m}_0(\lambda) = \frac{m - R(\lambda)}{(1 - \lambda)}.$$

It can be shown that $\mathbf{E}[\widehat{m}_0(\lambda)] \ge m_0$ when the p-values corresponding to the true null hypotheses are Uniform(0,1) distributed (or stochastically greater). There is an inherent bias/variance trade-off in the choice of λ. In most cases, when λ gets smaller, the bias of $\widehat{m}_0(\lambda)$ gets larger, but the variance gets smaller. Therefore, λ can be chosen to try to balance this trade-off. Storey and Tibshirani (2003) provide an intuitive motivation for the $\widehat{m}_0(\lambda)$ estimator, as well as a method for smoothing over the $\widehat{m}_0(\lambda)$ to obtain a tuning parameter free $\widehat{m}_0$ estimator. Sometimes instead

of m_0, the quantity $\pi_0 = m_0/m$ is estimated, where simply $\widehat{\pi}_0(\lambda) = \widehat{m}_0(\lambda)/m$.

To motivate the overall estimator $\widehat{\text{FDR}}(t) = \widehat{m}_0(\lambda) \cdot t/[R(t) \vee 1]$, it may be noted that $\widehat{m}_0(\lambda) \cdot t \approx V(t)$ and $[R(t) \vee 1] \approx R(t)$. It has been shown under a variety of assumptions, including those of Benjamini and Hochberg (1995), that the desired property $\mathbf{E}\left[\widehat{\text{FDR}}(t)\right] \geq \text{FDR}(t)$ holds.

Storey et al. (2004) have shown that the two major approaches to false discovery rates can be unified through the estimator $\widehat{\text{FDR}}(t)$. Essentially, the original FDR controlling algorithm can be obtained by setting $\widehat{m}_0 = m$ and utilizing the p-value threshold $t^*_\alpha = \max\left\{t : \widehat{\text{FDR}}(t) \leq \alpha\right\}$. By allowing for the different estimators $\widehat{m}_0(\lambda)$, a family of FDR controlling procedures can be derived in this manner. In the asymptotic setting where the number of hypothesis tests m is large, it has also been shown that the two approaches are essentially equivalent.

Q-Values

In single hypothesis testing, it is common to report the p-value as a measure of significance. The "q-value" is the FDR based measure of significance that can be calculated simultaneously for multiple hypothesis tests. Initially it seems that the q-value should capture the FDR incurred when the significance threshold is set at the p-value itself, $\text{FDR}(p_i)$. However, unlike Type I error rates, the FDR is not necessarily strictly increasing with an increasing significance threshold. To accommodate this property, the q-value is defined to be the minimum FDR (or pFDR) at which the test is called significant (Storey 2002, 2003):

$$\text{q-}value(p_i) = \min_{t \geq p_i} \text{FDR}(t) \quad \text{or}$$
$$\text{q-}value(p_i) = \min_{t \geq p_i} \text{pFDR}(t).$$

To estimate this in practice, a simple plug-in estimate is formed, for example:

$$\widehat{\text{q}}\text{-}value(p_i) = \min_{t \geq p_i} \widehat{\text{FDR}}(t).$$

Various theoretical properties have been shown for these estimates under certain conditions, notably that the estimated q-values of the entire set of tests are simultaneously conservative as the number of hypothesis tests grows large (Storey et al. 2004).

Bayesian Derivation

The pFDR has been shown to be exactly equal to a Bayesian derived quantity measuring the probability that a significant test is a true null hypothesis. Suppose that (a) $H_i = 0$ or 1 according to whether the ith null hypothesis is true or not, (b) $H_i \overset{i.i.d.}{\sim} \text{Bernoulli}(1-\pi_0)$ so that $\mathbf{Pr}(H_i = 0) = \pi_0$ and $\mathbf{Pr}(H_i = 1) = 1-\pi_0$, and (c) $P_i|H_i \overset{i.i.d.}{\sim} (1-H_i)\cdot G_0 + H_i \cdot G_1$, where G_0 is the null distribution and G_1 is the alternative distribution. Storey (2001, 2003) showed that in this scenario

$$\text{pFDR}(t) = \mathbf{E}\left[\frac{V(t)}{R(t)} \middle| R(t) > 0\right] = \mathbf{Pr}(H_i = 0 | P_i \leq t),$$

where $\mathbf{Pr}(H_i = 0|P_i \leq t)$ is the same for each i because of the i.i.d. assumptions. Under these modeling assumptions, it follows that $\text{q-value}(p_i) = \min_{t \geq p_i} \mathbf{Pr}(H_i = 0|P_i \leq t)$, which is a Bayesian analogue of the p-value – or rather a "Bayesian posterior Type I error rate." Related concepts were suggested as early as 1955 (Morton 1955). In this scenario, it also follows that $\text{pFDR}(t) = \int \mathbf{Pr}(H_i = 0|P_i = p_i) dG(p_i|p_i \leq t)$, where $G = \pi_0 G_0 + (1-\pi_0)G_1$. This connects the pFDR to the posterior error probability $\mathbf{Pr}(H_i = 0|P_i = p_i)$, making this latter quantity sometimes interpreted as a *local false discovery rate* (Efron et al. 2001; Storey 2001).

Dependence

Most of the existing procedures for utilizing false discovery rates in practice involve assumptions about the p-values being independent or weakly dependent. An area of current research is aimed at performing multiple hypothesis tests when there is dependence among the hypothesis tests, specifically at the level of the data collected for each test or the p-values calculated for each test. Recent proposals suggest modifying FDR controlling algorithms or extending their theoretical characterizations (Benjamini and Yekutieli 2001), modifying the null distribution utilized in calculating p-values (Devlin and Roeder 1999; Efron 2004), or accounting for dependence at the level of the originally observed data in the model fitting (Leek and Storey 2007, 2008).

About the Author

Dr. John D. Storey is Associate Professor of Genomics and Molecular Biology at Princeton University, with associated appointments in the Program in Applied and Computational Mathematics and the Department of Operations Research and Financial Engineering. He is currently an Associate editor of the *Annals of Applied Statistics* and *PLoS Genetics*, and has previously served on the editorial boards of *Biometrics*, *Biostatistics*, and *PLoS ONE*. He has published over 40 articles, including several highly cited articles on multiple hypothesis testing. He was recently recognized by Thomson Reuters as one of the top ten most cited mathematicians in the last decade.

Cross References

►Multiple Comparison
►Simes' Test in Multiple Testing

References and Further Reading

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 85:289–300
Benjamini Y, Liu W (1999) A step-down multiple hypothesis procedure that controls the false discovery rate under independence. J Stat Plann Infer 82:163–170
Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29: 1165–1188
Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc 99:96–104
Efron B, Tibshirani R, Storey JD, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96: 1151–1160
Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3:e161
Leek JT, Storey JD (2008) A general framework for multiple testing dependence. Proc Natl Acad Sci 105:18718–18723
Lindsay BG, Kettenring J, Siegmund DO (2004) A report on the future of statistics. Stat Sci 19:387–407
Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318
Shaffer J (1995) Multiple hypothesis testing. Ann Rev Psychol 46:561–584
Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754
Soric B (1989) Statistical discoveries and effect-size estimation. J Am Stat Assoc 84:608–610
Storey JD (2001) The positive false discovery rate: a Bayesian interpretation and the q-value. Technical Report 2001–2012, Department of Statistics, Stanford University
Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B 64:479–498
Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 31:2013–2035
Storey JD (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. J R Stat Soc Ser B 69: 347–368
Storey JD, Dai JY, Leek JT (2007) The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. Biostatistics 8:414–432
Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. J R Stat Soc Ser B 66:187–205
Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proc Natl Acad Sci 100:9440–9445
Zaykin DV, Young SS, Westfall PH (1998) Using the false discovery approach in the genetic dissection of complex traits: a response to weller et al. Genetics 150:1917–1918

Farmer Participatory Research Designs

Kaku Sagary Nokoe
Professor
University for Development Studies, Navrongo, Ghana

Introduction

Multilocation trials often follow classical on-station agronomic and breeding trials to test developed varieties under varying local conditions. Often these trials are imposed on farmer fields or set up as demonstration trials only to be viewed and ultimately to be adopted by rural poor farmers. On-farm trials involving the participation or use of farmers' fields have been applied in various studies, including: taungya and intercropping trials; mother–baby breeding trials aimed at selecting for specific traits in breed; augmented block designs (ABD) with emphasis on technology or selection of best or adaptable variety of crop under conditions of urgency and insufficient quantities of planting materials; crop livestock systems involving farmer management practices and animal preferences; and in the evaluation of adaptation and adoption of technologies. These trials are characterized by a high degree of variability within and between farmer fields (Mutsaers et al. 1997; Nokoe 1999; Odong 2002). Statistical issues of primary concern embrace the need for trial locations under farmer conditions, and why and how farmers may be involved to ensure acceptability and analyzability of selected designs.

From an intuitive but nonstatistical point of view, the involvement of all stakeholders (end user, researcher, community, donor) in the design of a trial, and the testing of trials under real-farm conditions utilizing options including maximum farmer management, is a sure way of enhancing adaptability and adoption. For breeders, there is a considerable advantage in time as duration from on-station to on-farm and then release is considerably shortened and results made more certain and output acceptable. On-farm research (OFR) has been variously classified, but generally could be grouped according to the level of farmer involvement. The class of interest in this entry is that involving the active participation of the farmer right from the design to the execution phases.

As an example, a participatory on-farm trial involving a crop–livestock system involved the following steps:

- Farmers and research institutions established formal collaborative linkages.
- Farmers discussed needs and prevailing cultural practices with researchers.

- Researchers and farmers evaluated intervention strategies.
- Statisticians guided selection of farmers and treatment allocations.
- Farmers randomly assigned inferior treatments allowed to change over time.

Arising from the above that is relevant from the point of view of designs is the fact that farmers are involved in the choice of treatments, blocks, and consequently the sampling or experimental designs. The implication is that block sizes are rarely of the same size or homogeneous, while considerable variability in some factors (such as variation in planting times) is common. In addition, it is common practice to have several standard controls (farmer practices), while in crop yield assessments the entire (not net or inner) plots are observed. Since, farmer differences are confounded in treatment, comparison of on-farm trials extend beyond differences in treatment effects. Mutsaers et al. (1997) point out that testing under farmer-field conditions and with their involvement provides a realistic assessment of the technologies or innovations under evaluation. Furthermore, the large number of farmers required is essential for capturing the expectedly high variation among farmer practices and sites. This large number should not be seen as a disadvantage, as the trade-off is the potentially high rate of adaptation and adoption of promising technologies (Nokoe 1999, 2000).

We shall consider general approaches aimed at the effective construction and analysis of farmer participatory designs.

The Design

Basic experimental principles (►randomization, replication, blocking, scope, and experimental error minimization) hold for participatory designs. The enforcement of these principles enables objectivity, estimation of standard errors, and effective comparison of treatment effects.

Identifying the Blocks

On-farm trials expectedly involve the use of block designs. The usual practice is to assume as blocks villages/communities (singly or cluster) and farm sites. This practice of assigning or identifying blocks is not very appropriate, though convenient. A preferred procedure would involve the use of statistical methods such as principal component and cluster analysis for constructing clusters. Classification variables must be relevant to the principal objectives of the study and would include socioeconomic, demographic, agronomic, and historical variables among others. It is emphasized that clusters are based on nontreatment characteristics that have the potential of influencing yields. As expected, the resulting blocks would not necessarily be contiguous, and that several farm sites from different villages, possibly distant-apart, may belong to the same block. Examples include classification of farm sites in an agroforestry and socioeconomic study and classification of farmers on the basis of soil type and cultural practices adopted. An alternative procedure, the post-model-based approach, would involve fitting a model (including discriminant functions) and then creating groups on the basis of limits of expected values, to which individuals may then be assigned.

Standard Block Structures

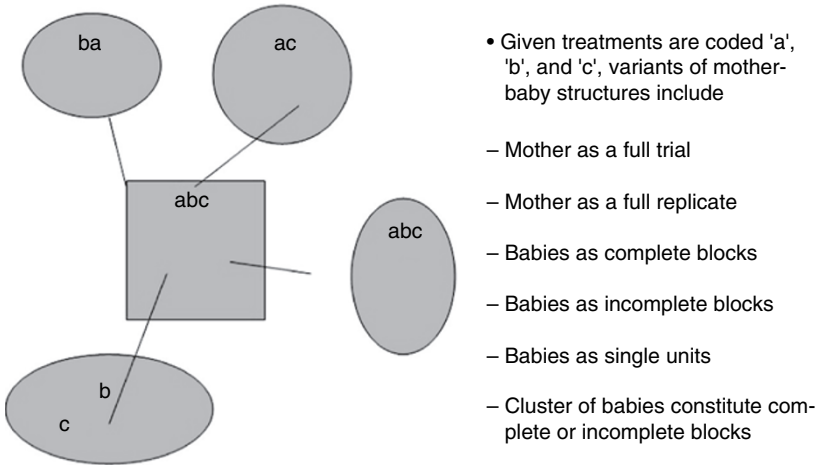
The block sizes may be equal (with each block receiving the same number of treatments) or balanced/unbalanced incomplete. Balanced complete block structure would generally imply treatments and pairs appear the same number of times in the experiment, and are present in all blocks. For incomplete block structures, several variants are available. These include alpha and cyclic incomplete block structures, and may be balanced with pairs of treatments appearing the same number of times in the experiment. Discussions on such designs are well documented (see, e.g., Cox and Reid 2000; and basic texts on designs). When block sizes are unequal, a fully balanced structure is not imaginable. Expectedly, in participatory designs, natural block structures are usually of the unequal and unbalanced type.

Augmented Block Structures

In recent times, block structures augmented with additional treatments, which are usually not replicated, are common in use. Augmented block designs (ABD) involve enlargement of blocks of a design (with treatments already assigned) to accommodate new treatments which appear usually once in the entire experiment (Federer 1955). The design allows for a wide range of technologies to be tested without necessarily straining resources or stifling farmers' interest [e.g., "1000 lines" in an on-farm situation is feasible]. Pinney (1991) provides an illustrated example for participatory agroforestry research.

Mother–Baby Structures

These block structures involve several blocks of varying sizes (Figure 1), with usually the larger sized one (the mother) having all or accommodating more treatments than the other blocks (baby/babies). The mother could also constitute a full trial with necessary replicates, and



Farmer Participatory Research Designs. Fig. 1 Treatment and Allocation to Blocks in Mother–Baby Trial

could represent an on-station or researcher-managed component of the trial. Babies may also be complete or incomplete blocks or as single experimental units with clusters of babies constituting blocks. The final block structure is arrived at after determining the number and sizes of the clusters.

Choice, Number of Treatments and Treatment Structures

The choice and number of treatments are made by consensus and on factors involved. The number need not be small as popularized in earlier works on OFR. However, the guiding principles are wide coverage (in the allocation of the treatments to the blocks), the willingness of the participating farmers, and availability of resources for the trial at the farmer/site level. The structure could be a single factor (say already packaged technologies or crop variety), factorial, or nested involving two or more factors. For factorial treatment structures, an alternative is the sequential or stepwise (step-up or step-down) arrangements. Stepwise allocation of treatments enable fewer number of treatments constituted from a number of factors, but the order of factor levels is crucial and needs to be well determined with all stakeholders. An example of factorial and corresponding stepwise is given in Table 1, where only four out of eight treatments are required for the stepwise (step-up and step-down options).

It is recommended that the decision to include a level of a factor, and at what step in the stepwise structure, must be determined jointly with all stakeholders. The sequence of factor levels affects and restricts the type of contrasts or comparisons that could reasonably be made (see, e.g., Mutsaers et al. 1991). It is also important the inability to estimate interactions is a major drawback of stepwise structures.

Observations and Measurements

Each farmer site represents different environments. Observations or measurements should therefore cover all other variables likely to account for the expectedly high variability. In crop trials with the response variable of interest being yield per hectare, there may be a need to include as many covariates (e.g., stand at establishment at plot level; soil depth, slope at field level; rainfall and labor cost at village level) and regressors (e.g., shade, labor size) as possible. In addition it is the gross and not the net plot that is observed. One basic advantage of such measurements from gross instead of usually uniform micro-net plot (as in on-station trials) is that yields are more likely to be realistic. Studies have shown that conversion of yield from micro to farmer level on the basis of small uniform on-station plot sizes considerably overestimates the real or farmer yield (Odulaja and Nokoe 1997).

Analytical Options

Several analytical options may be adapted from conventional ►analysis of variance and regression modeling. The particular option to use will be influenced by the desired objective, the hypotheses of interests, and the nature of the response variables. The response variables may be continous (normally or non-normally distributed) or discrete

Farmer Participatory Research Designs. Table 1 Breakdown of treatment structure using factorial and stepwise procedures

Treatment code	Maize variety	Fertilizer use	Planting density	Step
Factorial 2^3				
1	Local	Local	Farmers own	1
2	Local	Local	Recommended	1
3	Local	Recommended	Farmers own	1
4	Local	Recommended	Recommended	1
5	Recommended	Local	Farmers own	1
6	Recommended	Local	Recommended	1
7	Recommended	Recommended	Farmers own	1
8	Recommended	Recommended	Recommended	1
Stepwise (step-up)				
1	Local	Local	Farmers own	1
2	Recommended	Local	Farmers own	2
3	Recommended	Recommended	Farmers own	3
4	Recommended	Recommended	Recommended	4
Stepwise (step-down)				
1	Recommended	Recommended	Recommended	1
2	Recommended	Local	Recommended	2
3	Recommended	Local	Farmers own	3
4	Local	Local	Farmers own	4

(counts, ordinal, binary outcomes) or nominal outcomes. The options are briefly outlined.

Simple Analysis (Adjusted by Local Controls)

Treatments in farmer/village blocks are adjusted by farmer/village's own treatment (control) either directly (e.g., difference in response) or used as covariate (especially in situations where the farmer site is not a block).

Stability Analysis

Adaptability (stability) analysis made popular by Hildebrand in the 1980s for genotype by environment interaction can readily be adapted. This is a regression-based method used initially in genotype by environment interaction studies, where the treatment response is fitted to site mean and the estimated slope examined (see Example of output in Table 2). In the example, Variety differences were small; while high yields were associated with Fertilizer input. In particular, tropical zea streak resistance (TZSR) with Fertilizer input should be recommended – stable yields (with slope close to 1) imply same yield may be expected across all sites for this treatment combination. It may also be noted that Local/300 associated with certain sites (i.e., high yields of local with 300L fertilizer expected at some locations).

An alternative and enhanced procedure is through the use of *biplot and AMMI (additive main effect multiplicative analysis) models* (See, e.g., text of Milliken and Johnson 1987 or notes on Matmodel Software by Gauch). AMMI involves the partitioning of variance. For data presented in a contingency table format, biplot may also be obtained via the method of ►correspondence analysis which involves the partitioning of the chi-square.

Analysis of variance (ANOVA) with mixed models, where the error structure is adequately catered for and Contrasts, can be effectively used. In the mixed model scenario blocks, farmers, sites, etc., are usually treated as random effects being respectively a random sample from a large bulk. In particular, mixed modeling is most appropriate for augmented block designs (ABDs) and mother baby trials (MBTs) (Nokoe 1999), as it enables recovery of both inter-block and inter-treatment variation (Wolfinger et al. 1997). In ABDs, replicated lines are treated as fixed while non-replicated lines are considered as random.

Regression modeling is recommended for several situations where the design is not balanced, and/or when several auxiliary variables are available. These covariates or regressors, when included in the model, lead to considerable improvement in fit (Mutsaers et al. 1997; Carsky et al. 1998) and reduction of experimental error.

Categorical response variables, which constitute a substantial percentage of response variables, have not been modeled appropriately in several studies. These are better fitted by appropriate categorical modeling procedures, including the logistic and loglinear models (Bellon and Reeves 2002; Agresti 1990). An example of a trial with categorical responses is given in Table 3 (experimental setting and results) and Table 4 (partial data). In Table 4, the reader is to note the different types of response variables in the same data set – nominal (site history, indicating previous crop on farm site), ordinal (size of finger of plantains coded 1–3), and binary (size of plantain bunch and acceptability of product). It is important to indicate that a mixture of categorical and continuous variables is common in participatory design analysis.

Farmer Participatory Research Designs. Table 2 A simple regression based stability analysis for on-farm trial

Extracted output			
Fit Yield for variety, fertilizer, etc., as function of site mean (index). Comment on regression slopes.			
Variety	**Mean**	**slope, *b***	***P* > *t***
Local	2.464	1.02	0.0009
TZSR	2.831	0.97	0.0007
Fertilizer			
0	2.189	0.758	0.0003
300	3.107	1.240	< .0001
Treatment			
Local/ 0	1.984	0.697	0.0076
Local/300	2.945	1.356	0.0011
TZSR/0	2.393	0.818	0.0110
TZSR/300	3.269	1.124	0.0014
Conclusions: Variety differences low; high yields associated with Fertilizer input. TZSR with Fertilizer recommended			

Farmer Participatory Research Designs. Table 3 Partial data for plantain trial with categorical input and output variables

Zone	Farm Site	Treatment	Site Crop History	Weevil History	Bunch Size	Finger Size	Market Acceptability
East	1	A	Fallow	High	Small	1	Acceptable
	1	B	Fallow	High	Normal	1	Not
	1	C	Fallow	High	Normal	2	Not
	1	D	Plantain	Moderate	Normal	1	Acceptable
	2	A	Fallow	High	Small	2	Acceptable
	2	E	Fallow	Moderate	Small	1	Not
	2	D	Maize	High	Normal	2	Not
	2	C	Plantain	Low	Normal	3	Acceptable
West	8	C	Maize	Moderate	Small	2	Not
	8	B	Plantain	High	Normal	2	Not
	8	E	Plantain	Low	Small	3	Acceptable

Farmer Participatory Research Designs. Table 4 A summary of the plantain on-farm trial with categorical response variables

The experimental setting and variables	Models fitted and conclusions
Four Agro-Ecological zones across West/ Central/Eastern Africa involved Six plantain cultivars A, B, C, D, E, and F evaluated against local cultivar Only four cultivars allocated to farmers; each farmer provided four sites, one for each of the four allocated cultivars. Farmers local planted along with each cultivar or separately (as they wished) Previous field history recorded as: Fal – Fallow prior to trial Maz – Maize planted previous year Pla – Field already with plantain Weevil history (previous year when plantain had been cultivated): H – High Weevil population M – Medium Weevil population L – Low Medium population Response variables (assessed against local variety by farmers and chief) Bunch size: Binary - normal, small Finger size: Ordinal - 1 (least) to 4 (highest) Marketabilty: Binary - accept(able), not	*Logistic model* for bunch size, marketability as function of field history, weevil history, and other covariates, and regressors that may be available *Cumulative logit* model fitted to finger size Main conclusions: 1. Weevil history is the most important determinant of marketability 2. For bunch size, only finger size was significant at $p = 0.0434$. At 10% level of significance, study identified Weevil history and finger size as significant determinants of bunch size.

Conclusion

Participatory research designs can be planned and executed with scientifically verifiable results as outcome. It is emphasized that the understanding and active involvement of the farmer or end user are nontrivial considerations that need to be strictly adhered to for a successful research design aimed at addressing the needs of the usually poor rural farmer in developing countries. Adoption and adaptation of improved technologies are enhanced if all stakeholders are involved in the entire process. Blocking is a key ingredient in participatory designs, while the use of several regressors and covariates facilitate proper handling of the expectedly large variation between and within farm sites and farmer practices. The cocktail of analytical options requires adequate knowledge of statistical designs and the use of appropriate statistical software.

Acknowledgments

An earlier version of this entry had been presented at the regional sub-Saharan Africa Network of IBS meeting, and had benefited from inputs from workshop participants.

About the Author

Professor K S Nokoe, a Fellow of the Ghana Academy of Arts and Sciences, obtained his PhD from the University of British Columbia, and has over 30 years of teaching and research experience. He was Head of Department of Mathematical Sciences at the University of Agriculture in Nigeria (2002–2004), and the acting Vice-Chancellor of the University for Development Studies in Ghana (2007–2010). He had also served as biometrician in national and international research centres, published extensively in several peer-reviewed journals and supervised over 20 PhD and MSc students in Statistics, Statistical Ecology, Quantitative Entomology, Mensuration, Biometrics, and Statistical Computing among others. He is an Elected Member of the International Statistical Institute, Member of the International Association of Statistical Computing, Member of the International Biometric Society (IBS), and Founder of Sub-Saharan Africa Network (SUSAN) of IBS. He is the first recipient of the Rob Kempton Award of the International Biometric Society for "outstanding contribution to the development of biometry in developing countries."

Cross References

►Agriculture, Statistics in
►Research Designs
►Statistical Design of Experiments (DOE)

References and Further Reading

Agresti A (1990) Categorical data analysis. Wiley, New York
Carsky R, Nokoe S, Lagoke STO, Kim SK (1998) Maize yield determinants in farmer-managed trials in the Nigerian Northern Guinea Savanna. Experiment Agric 34:407–422

Cox DR, Reid N (2000) The theory of the design of experiments. Chapman & Hall/CRC, London
Federer WT (1955) Experimental design. MacMillan, New York
FCNS (2001) Food consumption and nutrition survey (Nigeria). Preliminary Report, IITA/USAID/UNICEF/FGN
Milliken GA, Dallas EJ (1987) Analysis of messy data vol. 2 nonreplicated experiments. Von Nostrand Reinhold, New York
Mutsaers HJW, Weber GK, Walker P (eds) (1991) On farm research in theory and practice. IITA Ibadan, Nigeria
Mutsaers HJW, Weber GK, Walker P, Fischer NM (1997) A field guide for on-farm experimentation. IITA/CTA/ISNAR, Ibadan
Nokoe S (1999) On farm trials: preventive and surgical approaches. J Trop For Resources (Special edition) 15(2):93–103
Nokoe S (2000) Biometric issues in agronomy: further on-station and on-farm designs and analyses. In: Akoroda MO (ed) Agronomy in Nigeria. Department of Agronomy, University of Ibadan, Nigeria, pp 35–42
Odong TL (2002) Assessment of variability in on-farm trial: a Uganda case. Unpublished dissertation, University of Natal MSc (Biometry), 103 pp
Odulaja A, Nokoe S (1997) Conversion of yield data from experimental plot to larger plot units. Discov Innovat 9: 137–141
Pinney A (1991) Farmer augmented designs for participatory agroforestry research. In: Patel MS, Nokoe S (eds) Biometry for development. ICIPE Science Press, Nairobi, pp 39–50
Wolfinger RD, Federer WT, Cordero-Brana O (1997) Recovering information in augmented designs using SAS PROC GLM and PROC MIXED. Agron J 89:856–859

Federal Statistics in the United States, Some Challenges

Edward J. Spar
Executive Director
Council of Professional Associations on Federal Statistics, Alexandria, VA, USA

The Current Status

Are the federal statistical agencies in the United States meeting their mandates? Overall, the answer is yes. Surveys that are required for policy purposes in health, education, labor, and other areas are being conducted with well-tested statistical designs that so far have reasonable margins of error. The decennial census, even with an under and over count meets the needs of the Constitution and thousands of federal, state and local data users. Measures, including labor force data, gross domestic product, the system of national accounts, health, education, and income estimates are excellently covered by the federal statistical agencies. Estimates of the population are reasonable even in situations where high immigration and/or internal migration, that have disproportionate influence, take place. The agencies are very sensitive of the need to maintain the confidentiality of respondents. Based on the above, it sounds as if the federal statistical system in the United States is healthy and on track; yet what about the future?

Ongoing and Upcoming Issues

Many new problems are facing the statistical agencies in the United States, and it will take enormous effort to solve them. Indeed, the agencies are fully aware of them and understand that there is a need for innovative thinking. An example of the type of innovation that has already taken place is the U.S. Census Bureau's American Community Survey. This is a replacement for the decennial census long form, and at the same time as an ongoing annual survey of about three million housing units, is unique. The ability to have data available every year for national, state, and local geographies is an important step for a dynamic country such as the United States. Another innovative set of data is the U.S. Census Bureau's Longitudinal Employer–Household Employer Dynamic. Using a mathematical model to insure non-disclosure, data are available for detailed employment statistics at very local geographic levels.

An issue that is becoming critical and is being looked at closely is the declining response rates in key federal surveys that measure employment, income, consumer expenditures, health and education, for example. Surveys that were achieving rates in the middle to high 90% range are now attaining response rates well below that. Clearly, the continuing decline in non-response will have serious effects on the usefulness of data collected. Either the statistical error will become so high so as to make the estimates of limited value, or, perhaps even worse, with biases due to non-response, the data may lose most of its value. Clearly the statistical agencies are aware of the problem and much research is being conducted to determine if address-listing techniques, for example, can be of use in conjunction with telephone interviewing. Some work has been accomplished in the areas of non-response and statistical and non-statistical biases but much more is required. The issue of conducting telephone surveys (see ►Telephone Sampling: Frames and Selection Techniques), given the elimination of land lines on the part of households and their turning to the increasing use of cell phones, must be addressed.

The data retrieval world has been transformed by the world-wide-web. The concept of charging for governmental data is no longer realistic given the assumption on the part of users that all data should be free on-line. Also, search engines such as Google have enabled users to retrieve diverse information as an integrated "package."

However, data integration across federal statistical agencies is for the most part limited. For example, there is no way to analyze and reconcile the many different measures of income between and sometimes even within an agency. Each agency creates its own web site and its own data dissemination system with little or no regard to the fact that the user has to go to a over a dozen sites and learn a dozen approaches to data retrieval to get a complete review of the socio-economic data of the United States. Indeed, if the user wants to integrate the data, it's much easier, but more expensive to go to a private sector vendor to do the work for you. At a time when the web is there for the specific purpose to retrieve information easily, freely, and comprehensively, this approach is outdated. The time has come for an integration of data processing and retrieval systems. This should be accomplished even though the structure of the federal statistical system in the United States is highly decentralized. The concept of a single system in the case of the United States, and probably most countries, is misleading. In reality what you have is a confederation of agencies for the most part reporting to different jurisdictions and quite independent of each other. In the United States, there is very limited administrative record data sharing and with separate Internet sites mentioned above, little integration of tabulated data sets. Each agency has its own budget and except for the purchasing of surveys from the U.S. Census Bureau, little in the way of financial interaction. Unfortunately, because of this lack of centralization, the agencies don't have great influence with the Congress. (This is not the case during the decennial census cycle where the apportionment of Congressional seats can impact a member of the House of Representatives. Other data series such as employment and inflation are also closely looked at.) This lack of influence can be a problem for an agency that each year must request funding for its programs. Would a centralized single agency help solve this? An agency large enough to be noticed by Congress and the Administration as being critical to the overall health of the nation would have a better opportunity of receiving the needed resources to implement innovative statistical techniques.

To perhaps overstate the case, the days of taking censuses and surveys may soon be coming to an end. We may be at the crossroads of having to rely, for the very most part, on administrative records. The use of administrative record data brings up issues of confidentiality on the part of agencies and the sensitivity to the privacy needs of the public. Yet these data may have to become the basis for measuring health, education, employment, expenditure, transportation, energy use and many more statistical needs on the part of the federal government. Using administrative data will call for public/private sector coordinated analyses and the allocation of talent and research dollars. If the use of administrative data becomes the norm, it is not too outré to see a time when no data will be real – put another way, they will be modeled estimates based on the original data series. As previously mentioned, we already see such a transformation in the U.S. Census Bureau's Longitudinal Employer-Household Employer Dynamic program produced at the local level. Indeed, once the block group level data from the American Community Survey are analyzed, we may also see some move in the same direction.

Over the next few years, much of the senior staffs of statistical agencies will be of retirement age. At the same time, it's difficult for agencies to hire new personnel and hold on to talented statisticians and economists that have entered the federal statistical system. The private sector offers both higher salaries and the opportunity to diversify. Indeed, the problem of "stove-piping" within statistical agencies, where talented people are expected to stay in one place for an overly extended period of time, is counter-productive. There is a need to develop a system whereby people can move not only within an agency, but also across agencies. Such a system of diverse training will be required so that personnel can develop the skills needed to address the concerns that have been mentioned in this review.

The challenges reviewed above are only the beginning. In order to properly measure the effects of the current and probably future economic crises in the United States, timely and relevant data are needed for those who have to make informed decisions affecting all Americans.

Cross References

►Census
►Nonresponse in Surveys
►Telephone Sampling: Frames and Selection Techniques

Fiducial Inference

Jan Hannig[1], Hari Iyer[2], Thomas C. M. Lee[3]
[1]Associate Professor
The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[2]Professor
Colorado State University, Fort Collins, CO, USA
[3]Professor
The University of California at Davis, Davis, CA, USA

Introduction

The origin of Generalized Fiducial Inference can be traced back to R. A. Fisher (Fisher 1930, 1933, 1935) who introduced the concept of a fiducial distribution for a

parameter, and proposed the use of this fiducial distribution, in place of the Bayesian posterior distribution, for interval estimation of this parameter. In the case of a one-parameter family of distributions, Fisher gave the following definition for a fiducial density $f(\theta|x)$ of the parameter based on a single observation x for the case where the cdf $F(x|\theta)$ is a monotonic decreasing function of θ:

$$f(\theta/x) = -\frac{\partial F(x|\theta)}{\partial \theta}. \tag{1}$$

In simple situations, especially in one parameter families of distributions, Fisher's fiducial intervals turned out to coincide with classical confidence intervals. For multiparameter families of distributions, the fiducial approach led to confidence sets whose frequentist coverage probabilities were close to the claimed confidence levels but they were not exact in the frequentist sense. Fisher's proposal led to major discussions among the prominent statisticians of the 1930's, 40's and 50's (e.g., Dempster 1966, 1968; Fraser 1961a, b, 1966, 1968; Jeffreys 1940; Lindley 1958; Stevens 1950). Many of these discussions focused on the nonexactness of the confidence sets and also nonuniqueness of fiducial distributions. The latter part of the 20th century has seen only a handful of publications Barnard (1995); Dawid and Stone (1982); Dawid et al. (1973); Salome (1998); Wilkinson (1977) as the fiducial approach fell into disfavor and became a topic of historical interest only.

Recently, the work of Tsui and Weerahandi (1989, 1991) and Weerahandi (1993, 1994, 1995) on generalized confidence intervals and the work of Chiang (2001) on the *surrogate variable method* for obtaining confidence intervals for variance components, led to the realization that there was a connection between these new procedures and fiducial inference. This realization evolved through a series of works (Hannig 2009b; Hannig et al. 2006b; Iyer and Patterson 2002; Iyer et al. 2004; Patterson et al. 2004). The strengths and limitations of the fiducial approach is becoming to be better understood, see, especially, Hannig (2009b). In particular, the asymptotic exactness of fiducial confidence sets, under fairly general conditions, was established in Hannig et al. (2006b); Hannig (2009a,b).

Subsequently Hannig et al. (2003); Iyer et al. (2004); McNally et al. (2003); Wang and Iyer (2005, 2006a,b) applied this fiducial approach to derive confidence procedures in many important practical problems. Hannig (2009b) extended the initial ideas and proposed a Generalized Fiducial Inference procedure that could be applied to arbitrary classes of models, both parametric and nonparametric, both continuous and discrete. These applications include Bioequivalence Hannig et al. (2006a), Variance Components Lidong et al. (2008), Problems of Metrology Hannig et al. (2007, 2003); Wang and Iyer (2005, 2006a, b), Interlaboratory Experiments and International Key Comparison Experiments Iyer et al. (2004), Maximum Mean of a Multivariate Normal Distribution Wandler and Hannig (2009), Mixture of a Normal and Cauchy Glagovskiy (2006), Wavelet Regression Hannig and Lee (2009), ▶Logistic Regression and LD_{50} Lidong et al. (2009). Recently, other authors have also contributed to research on fiducial methods and related topics (e.g., Berger and Sun 2008; Wang 2000; Xu and Li 2006).

Generalized Fiducial Distribution

The idea underlying Generalized Fiducial Inference comes from an extended application of Fisher's fiducial argument, which is briefly described as follows. Generalized Fiducial Inference begins with expressing the relationship between the data, $\mathbf{X}$, and the parameters, $\boldsymbol{\theta}$, as

$$\mathbf{X} = G(\boldsymbol{\theta}, \mathbf{U}), \tag{2}$$

where $G(\cdot,\cdot)$ is termed structural equation, and $\mathbf{U}$ is the random component of the structural equation whose distribution is completely known. The data $\mathbf{X}$ are assumed to be created by generating a random variable $\mathbf{U}$ and plugging it into the structural equation (2).

For simplicity, this section only considers the case where the structural relation (2) can be inverted and the inverse $G^{-1}(\cdot,\cdot)$ always exists. Thus, for any observed $\mathbf{x}$ and for any arbitrary $\mathbf{u}$, $\boldsymbol{\theta}$ is obtained as $\boldsymbol{\theta} = G^{-1}(\mathbf{x}, \mathbf{u})$. Fisher's *Fiducial Argument* leads one to define the fiducial distribution for $\boldsymbol{\theta}$ as the distribution of $G^{-1}(\mathbf{x}, \mathbf{U}^\star)$ where $\mathbf{U}^\star$ is an independent copy of $\mathbf{U}$. Equivalently, a sample from the fiducial distribution of $\boldsymbol{\theta}$ can be obtained by generating $\mathbf{U}_i^\star$, $i = 1, \ldots, N$ and using $\boldsymbol{\theta}_i = G^{-1}\left(\mathbf{x}, \mathbf{U}_i^\star\right)$. Estimates and confidence intervals for $\boldsymbol{\theta}$ can be obtained based on this sample.

Hannig (2009b) has generalized this to situations where G is not invertible. The resulting fiducial distribution is called a Generalized Fiducial Distribution. To explain the idea we begin with Eq. 2 but do not assume that G is invertible with respect to $\boldsymbol{\theta}$. The inverse $G^{-1}(\cdot,\cdot)$ may not exist for one of the following two reasons: for any particular $\mathbf{u}$, either there is no $\boldsymbol{\theta}$ satisfying (2), or there is more than one $\boldsymbol{\theta}$ satisfying (2).

For the first situation, Hannig (2009b) suggests removing the offending values of $\mathbf{u}$ from the sample space and then re-normalizing the probabilities. Such an approach has also been used by Fraser (1968) in his work on structural inference. Specifically, we generate $\mathbf{u}$ conditional on the event that the inverse $G^{-1}(\cdot,\cdot)$ exists. The rationale

for this choice is that we know our data $\mathbf{x}$ were generated with some $\boldsymbol{\theta}_0$ and $\mathbf{u}_0$, which implies there is at least one solution $\boldsymbol{\theta}_0$ satisfying (2) when the "true" $\mathbf{u}_0$ is considered. Therefore, we restrict our attention to only those values of $\mathbf{u}$ for which $G^{-1}(\cdot,\cdot)$ exists. However, this set has probability zero in many practical situations leading to non-uniqueness due to the Borel paradox (Casella and Berger 2002, Section 4.9.3). The Borel paradox is the fact that when conditioning on an event of probability zero, one can obtain any answer.

The second situation can be dealt with either by selecting one of the solutions or by the use of the mechanics underlying Dempster-Shafer calculus Dempster (2008). In any case, Hannig (2009a) proved that this non-uniqueness disappears asymptotically under very general assumptions.

Hannig (2009b) proposes the following formal definition of the generalized fiducial recipe. Let $\mathbf{X} \in \mathbb{R}^n$ be a random vector with a distribution indexed by a parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Recall that the data generating mechanism for $\mathbb{X}$ is expressed by (2) where G is a jointly measurable function and $\mathbf{U}$ is a random variable or vector with a completely known distribution independent of any parameters. We define for any measurable set $A \in \mathbb{R}^n$ a set-valued function

$$Q(A,\mathbf{u}) = \{\boldsymbol{\theta} : G(\boldsymbol{\theta},\mathbf{u}) \in A\}. \tag{3}$$

The function $Q(A,\mathbf{u})$ is the generalized inverse of the function G. Assume $Q(A,\mathbf{u})$ is a measurable function of $\mathbf{u}$.

Suppose that a data set was generated using (2) and it has been observed that the sample value $\mathbf{x} \in A$. Clearly the values of $\boldsymbol{\theta}$ and $\mathbf{u}$ used to generate the observed data will satisfy $G(\boldsymbol{\theta},\mathbf{u}) \in A$. This leads to the following definition of a generalized fiducial distribution for $\boldsymbol{\theta}$:

$$Q(A,\mathbf{U}^\star) \mid \{Q(A,\mathbf{U}^\star) \neq \emptyset\}, \tag{4}$$

where $\mathbf{U}^\star$ is an independent copy of $\mathbf{U}$.

The object defined in (4) is a random set of parameters (such as an interval or a polygon) with distribution conditioned on the set being nonempty. It is well-defined provided that $P(Q(A,\mathbf{U}^\star) \neq \emptyset) > 0$. Otherwise additional care needs to be taken to interpret this distribution (c.f., Hannig 2009b). In applications, one can define a distribution on the parameter space by selecting one point out of $Q(A,\mathbf{U}^\star)$.

Examples

The following examples provide simple illustrations of the definition of a generalized fiducial distribution.

Example 1 Suppose $\mathbf{U} = (U_1,U_2)$ where U_i are i.i.d. $N(0,1)$ and $\mathbb{X} = (X_1,X_2) = G(\mu,\mathbf{U}) = (\mu+U_1,\mu+U_2)$ for some $\mu \in \mathbb{R}$. So X_i are iid $N(\mu,1)$. Given a realization $\mathbf{x} = (x_1,x_2)$ of $\mathbb{X}$, the set-valued function Q maps $\mathbf{u} = (u_1,u_2) \in R^2$ to a subset of $\mathbb{R}$ and is given by

$$Q(\mathbf{x},\mathbf{u}) = \begin{cases} \{x_1 - e_1\} & \text{if} \quad x_1 - x_2 = u_1 - u_2, \\ \emptyset & \text{if} \quad x_1 - x_2 \neq u_1 - u_2. \end{cases}$$

By definition, a generalized fiducial distribution for μ is the distribution of $x_1 - U_1^\star$ conditional on $U_1^\star - U_2^\star = x_1 - x_2$ where $\mathbf{U}^\star = \left(U_1^\star, U_2^\star\right)$ is an independent copy of $\mathbf{U}$. Hence a generalized fiducial distribution for μ is $N(\bar{x},1/2)$ where $\bar{x} = (x_1+x_2)/2$.

Example 2 Suppose $\mathbf{U} = (U_1,\ldots,U_n)$ is a vector of i.i.d. uniform $(0,1)$ random variables U_i. Let $p \in [0,1]$. Let $X = (X_1,\ldots,X_n)$ be defined by $X_i = I(U_i < p)$. So X_i are iid Bernoulli random variables with success probability p. Suppose $x = (x_1,\ldots,x_n)$ is a realization of X. Let $s = \sum_{i=1}^n x_i$ be the observed number of 1's. The mapping $Q : [0,1]^n \to [0,1]$ is given by

$$Q(x,\mathbf{u}) = \begin{cases} [0,u_{1:n}] & \text{if} \quad s = 0, \\ (u_{1:n},1] & \text{if} \quad s = n, \\ (u_{s:n},u_{s+1:n}] & \text{if} \quad s = 1,\ldots,n-1 \text{ and} \\ & \sum_{i=1}^n I(x_i = 1)I(u_i \leq u_{s:n}) = s, \\ \emptyset & \text{otherwise.} \end{cases}$$

Here $u_{r:n}$ denotes the rth order statistic among $u_1,\ldots,u_n$. So a generalized fiducial distribution for p is given by the distribution of $Q(x,\mathbf{U}^\star)$ conditional on the event $Q(x,\mathbf{U}^\star) \neq \emptyset$. By the exchangeability of $U_1^\star,\ldots,U_n^\star$ it follows that the stated conditional distribution of $Q(x,\mathbf{U}^\star)$ is the same as the distribution of $\left[0,U_{1:n}^\star\right]$ when $s = 0$, $\left(U_{s:n}^\star,U_{s+1:n}^\star\right]$ for $0 < s < n$, and $\left(U_{n:n}^\star,1\right]$ for $s = n$.

Next, we present a general recipe that is useful in many practical situations.

Example 3 Let us assume that the observations $X_1,\ldots,X_n$ are i.i.d. univariate with distribution function $F(x,\xi)$ and density $f(x,\xi)$, where ξ is a p-dimensional parameter. Denote the generalized inverse of the distribution function by $F^{-1}(\xi,u)$ and use the structural equation

$$X_i = F^{-1}(\xi,U_i) \qquad \text{for} \quad i = 1,\ldots,n. \tag{5}$$

If all the partial derivatives of $F(x,\xi)$ with respect to ξ are continuous and the Jacobian

$$\det\left(\frac{\mathbf{d}}{\mathbf{d}\xi}\left(F(x_{i_1},\xi),\ldots,F(x_{i_p},\xi)\right)\right) \neq 0$$

for al distinct $x_1, \ldots, x_p$, then Hannig (2009b) shows that the generalized fiducial distribution (4) is

$$r(\xi) = \frac{f_{\mathbf{X}}(\mathbf{x}|\xi) J(\mathbf{x}, \xi)}{\int_{\Xi} f_{\mathbf{X}}(\mathbf{x}|\xi') J(\mathbf{x}, \xi')\, d\xi'}, \tag{6}$$

where

$$J(\mathbf{x}, \xi) = \sum_{\mathbf{i}=(i_1,\ldots,i_p)} \left| \frac{\det\left(\frac{\mathbf{d}}{\mathbf{d}\xi}\left(F(x_{i_1}, \xi), \ldots, F(x_{i_p}, \xi)\right)\right)}{f(x_{i_1}, \xi) \cdots f(x_{i,p}, \xi)} \right|. \tag{7}$$

This provides a form of generalized fiducial distribution that is usable in many practical applications, see many of the papers mentioned in introduction. Moreover, if $n = p = 1$ (6) and (7) simplify to the Fisher's original definition (1).

Equation 6 is visually similar to Bayes posterior. However, the role of the prior is taken by the function $J(\mathbf{x}, \xi)$. Thus unless $J(\mathbf{x}, \xi) = k(\mathbf{x})l(\xi)$ where k and l are measurable functions, the generalized fiducial distribution is not a posterior distribution with respect to any prior. A classical example of such a situation is in Grundy (1956).

Moreover, $\binom{n}{p}^{-1} J(\mathbf{x}, \xi)$ is a ►U-statistic and therefore it often converges a.s. to

$$\pi_{\xi_0}(\xi) = E_{\xi_0} \left| \frac{\det\left(\frac{\mathbf{d}}{\mathbf{d}\xi}\left(F(X_1, \xi), \ldots, F(X_p, \xi)\right)\right)}{f(X_1, \xi) \cdots f(X_p, \xi)} \right|$$

At first glance $\pi_{\xi_0}(\xi)$ could be viewed as an interesting non-subjective prior. Unfortunately, this prior is not usable in practice, because the expectation in the definition of $\pi(\xi)$ is taken with respect to the true parameter ξ_0 which is unknown. However, since $\binom{n}{p}^{-1} J(\mathbf{x}, \xi)$ is an estimator of $\pi_{\xi_0}(\xi)$, the generalized fiducial distribution (6) could be interpreted as an empirical Bayes posterior.

Acknowledgments

The authors' research was supported in part by the National Science Foundation under Grant No. 0707037.

About the Authors

Jan Hannig is the Associate Professor of Statistics and Operations Research at The University of North Carolina at Chapel Hill. He is an Associate Editor of *Electronic Journal of Statistics* and an elected member of International Statistical Institute (ISA).

Hari Iyer is a Research Fellow at Caterpillar Inc. He is also a Professor of Statistics at Colorado State University.

Thomas C. M. Lee is a Professor of Statistics at the University of California, Davis. Before joining UC Davis, he had held regular and visiting faculty positions at University of Chicago, Chinese University of Hong Kong, Colorado State University and Harvard University. He is an elected Fellow of the American Statistical Association, and an elected Senior Member of the IEEE. He has published more than 50 papers in refereed journals and conference proceedings. Currently he is an Associate editor for *Bernoulli*, *Journal of Computational and Graphical Statistics*, and *Statistica Sinica*.

Cross References

►Behrens–Fisher Problem
►Confidence Distributions
►Statistical Inference: An Overview

References and Further Reading

Barnard GA (1995) Pivotal models and the fiducial argument. Int Stat Rev 63:309–323
Berger JO, Sun D (2008) Objective priors for the bivariate normal model. Ann Stat 36:963–982
Casella G, Berger RL (2002) Statistical inference, 2nd edn. Wadsworth and Brooks/Cole, Pacific Grove, CA
Chiang A (2001) A simple general method for constructing confidence intervals for functions of variance components. Technometrics 43:356–367
Dawid AP, Stone M (1982) The functional-model basis of fiducial inference (with discussion). Ann Stat 10:1054–1074
Dawid AP, Stone M, Zidek JV (1973) Marginalization paradoxes in Bayesian and structural inference (with discussion). J R Stat Soc Ser B 35:189–233
Dempster AP (1966) New methods for reasoning towards posterior distributions based on sample data. Ann Math Stat 37:355–374
Dempster AP (1968) A generalization of Bayesian inference (with discussion). J R Stat Soc Ser B 30:205–247
Dempster AP (2008) The Dempster-Shafer calculus for statisticians. Int J Approx Reason 48:365–377
Fisher RA (1930) Inverse probability. Proc Cambridge Philos Soc 26:528–535
Fisher RA (1933) The concepts of inverse probability and fiducial probability referring to unknown parameters. Proc R Soc Lond A 139:343–348
Fisher RA (1935) The fiducial argument in statistical inference. Ann Eugenics 6:91–98
Fraser DAS (1961a) The fiducial method and invariance. Biometrika 48:261–280
Fraser DAS (1961b) On fiducial inference. Ann Math Stat 32:661–676
Fraser DAS (1966) Structural probability and a generalization. Biometrika 53:1–9
Fraser DAS (1968) The structure of inference. Wiley, New York
Glagovskiy YS (2006) Construction of fiducial confidence intervals for the mixture of cauchy and normal distributions. Master's thesis, Department of Statistics, Colorado State University
Grundy PM (1956) Fiducial distributions and prior distributions: an example in which the former cannot be associated with the latter. J R Stat Soc Ser B 18:217–221
Hannig J (2009a) On asymptotic properties of generalized fiducial inference for discretized data. Tech. Rep. UNC/STOR/09/02, Department of Statistics and Operations Research, The University of North Carolina

Hannig J (2009b) On generalized fiducial inference. Stat Sinica 19:491–544
Hannig J, Abdel-Karim LEA, Iyer HK (2006a) Simultaneous fiducial generalized confidence intervals for ratios of means of lognormal distributions. Aust J Stat 35:261–269
Hannig J, Iyer HK, Patterson P (2006b) Fiducial generalized confidence intervals. J Am Stat Assoc 101:254–269
Hannig J, Iyer HK, Wang JC-M (2007) Fiducial approach to uncertainty assessment accounting for error due to instrument resolution. Metrologia 44:476–483
Hannig J, Lee TCM (2009) Generalized fiducial inference for wavelet regression. Biometrika 96(4):847–860
Hannig J, Wang CM, Iyer HK (2003) Uncertainty calculation for the ratio of dependent measurements. Metrologia, 4: 177–186
Iyer HK, Patterson P (2002) A recipe for constructing generalized pivotal quantities and generalized confidence intervals. Tech. Rep. 2002/10, Department of Statistics, Colorado State University
Iyer HK, Wang JC-M, Mathew T (2004) Models and confidence intervals for true values in interlaboratory trials. J Am Stat Assoc 99:1060–1071
Jeffreys H (1940) Note on the Behrens-Fisher formula. Ann Eugenics 10:48–51
Lidong E, Hannig J, Iyer HK (2008) Fiducial Intervals for variance components in an unbalanced two-component normal mixed linear model. J Am Stat Assoc 103:854–865
Lidong E, Hannig J, Iyer HK (2009) Fiducial generalized confidence interval for median lethal dose (LD50). (Preprint)
Lindley DV (1958) Fiducial distributions and Bayes' theorem. J R Stat Soc Ser B 20:102–107
McNally RJ, Iyer HK, Mathew T (2003) Tests for individual and population bioequivalence based on generalized p-values. Stat Med 22:31–53
Patterson P, Hannig J, Iyer HK (2004) Fiducial generalized confidence intervals for proportion of conformance. Tech. Rep. 2004/11, Colorado State University
Salome D (1998) Staristical inference via fiducial methods. Ph.D. thesis, University of Groningen
Stevens WL (1950) Fiducial limits of the parameter of a discontinuous distribution. Biometrika 37:117–129
Tsui K-W, Weerahandi S (1989) Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. J Am Stat Assoc 84:602–607
Tsui K-W, Weerahandi S (1991) Corrections: generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. [J Am Stat Assoc 84 (1989), no. 406, 602–607; MR1010352 (90g:62047)]. J Am Stat Assoc 86:256
Wandler DV, Hannig J (2009) Fiducial inference on the maximum mean of a multivariate normal distribution (Preprint)
Wang JC-M, Iyer HK (2005) Propagation of uncertainties in measurements using generalized inference. Metrologia 42: 145–153
Wang JC-M, Iyer HK (2006a) A generalized confidence interval for a measurand in the presence of type-A and type-B uncertainties. Measurement 39:856–863
Wang JC-M, Iyer HK (2006b) Uncertainty analysis for vector measurands using fiducial inference. Metrologia 43: 486–494
Wang YH (2000) Fiducial intervals: what are they? Am Stat 54: 105–111
Weerahandi S (1993) Generalized confidence intervals. J Am Stat Assoc 88:899–905
Weerahandi S (1994) Correction: generalized confidence intervals [J Am Stat Assoc 88 (1993), no. 423, 899–905; MR1242940 (94e:62031)]. J Am Stat Assoc 89:726
Weerahandi S (1995) Exact statistical methods for data analysis. Springer series in statistics. Springer-Verlag. New York
Wilkinson GN (1977) On resolving the controversy in statistical inference (with discussion). J R Stat Soc Ser B 39: 119–171
Xu X, Li G (2006) Fiducial inference in the pivotal family of distributions. Sci China Ser A Math 49:410–432

Financial Return Distributions

Matthias Fischer
University of Erlangen-Nürnberg, Erlangen, Germany

Describing past and forecasting future asset prices has been attracting the attention of several generations of researchers. Rather than analyzing the asset prices P_t at times $t = 1, \ldots, T$ themselves, one usually focusses on the corresponding log-returns defined by $R_t^c = \log(P_t) - \log(P_{t-1})$ for $t = 2, \ldots, T$. Considering prices (and consequently log-returns) as realizations of random variables, it seems natural to identify the underlying data-generating probability distribution. The search for an adequate model for the distribution of stock market returns dates back to the beginning of the twentieth century: Following Courtault et al. (2000), "*The date March 29, 1900, should be considered as the birthdate of mathematical finance. On that day, a French postgraduate student, Louis Bachélier, successfully defended at the Sorbonne his thesis Théorie de la Spéculation. […] This pioneering analysis of the stock and option markets contains several ideas of enormous value in both finance and probability. In particular, the theory of Brownian motion (see* ►*Brownian Motion and Diffusions*), *was initiated and used for the mathematical modelling of price movements and the evaluation of contingent claims in financial markets*".

Whereas Bachélier (1900) rests upon normally distributed return distributions, the history of heavy tails in finance began in 1963: Assuming independence of successive increments and the validity of the principle of scaling invariance, Mandelbrot (1963) advocates the (Lévy) stable distributions for price changes, supported by Fama (1965) and Fielitz (1976). In fact, tails of stable distributions are very heavy, following a power-law distribution with an exponent $\alpha < 2$. In contrast, empirical studies indicate that tails of most financial time series have to be modeled with $\alpha > 2$ (see, e.g., Lau et al. 1990 or Pagan 1996). In particular,

Akgiray et al. (1989) support this conjecture for German stock market returns. Rejecting the stable hypothesis, several proposals came up in the subsequent years: Praetz (1972), Kon (1984) or Akgiray and Booth (1987) favour finite mixtures of normal distributions, whereas, e.g., Ball and Torous (1983) propose an infinite number of normal distributions mixtures with Poisson probabilities.

Since the early seventies of the last century, the Student-*t* distribution increases in popularity (see, e.g., Blattberg and Gonedes 1974). Depending on the shape and tail parameter ν, moments of the Student-*t* distribution exist only up to a certain order depending on ν, whereas the ►moment-generating function doesn't exist. In order to increase its flexibility regarding ►skewness, peakedness and tail behavior, several generalized Student-*t* versions followed up within the past years (see, e.g., McDonald and Newey 1988; Theodossiou 1998; Hansen et al. 2003 or Adcock and Meade 2003). Finally, if both (semi-)heavy tails and existence of the corresponding moment-generating function are required, different multi-parametric distribution families with exponential tail behavior were successfully applied to financial returns: Among them, the generalized logistic distribution family (see, e.g., Bookstaber and McDonald 1987; McDonald 1991 or McDonald and Bookstaber 1991), the generalized hyperbolic secant distribution families (see, e.g., Fischer 2004, 2006) and the generalized hyperbolic distribution family (see, e.g., Eberlein and Keller 1995; Barndorff-Nielsen 1995; Küchler et al. 1999 and Prause 1999) which in turn includes a subfamily (in the limit) where one tail has polynomial and the other exponential tails, see Aas and Haff (2006).

Selecting a suitable probability distribution for a given return data sample is by far not an easy task. In general, there is no information about the tail behavior of the unknown distribution or, conversely, about the order up to which the corresponding moments exit. Discussions as to whether moments, in particular variances, do exist or not, have a long tradition in financial literature (see, for instance, Tucker 1992). In order to check whether certain moments do exist or not, Granger and Orr (1972) introduced the so-called running-variance plot. Alternatively, the test statistic of Yu (2000) – which is determined by the range of the sample interquartile and the sample standard deviation – may come to application. Recently, Pisarenko and Sornette (2006) came up with a test statistic to discriminate between exponential and polynomial tail behavior.

Cross References

►Hyperbolic Secant Distributions and Generalizations
►Statistical Modeling of Financial Markets

References and Further Reading

Aas K, Haff IH (2006) The generalized hyperbolic skew Student's t-distribution. J Financ Econom 4(2):275–309

Adcock CJ, Meade N (2003) An extension of the generalised skew Student distribution with applications to modelling returns on financial assets. Working paper, Department of Economics, University of Sheffield

Akgiray V, Booth GG (1987) Compound distribution models of stock returns: an empirical comparison. J Financ Res 10(3): 269–280

Akgiray V, Booth GG, Loistl O (1989) Statistical models of German stock returns. J Econ 50(1):17–33

Bachélier L (1900) Théorie de la spéculation. Annales Scientiéques de l'Ecole Nor-male Supérieure 17(3):21–81

Ball CA, Torous W (1983) A simplified jump process for common stock returns. J Financ Quant Anal 18:53–65

Barndorff-Nielsen OE (1995) Normal inverse Gaussian processes and the modelling of stock returns. Research Report No. 300, Department of Theoretical Statistics, University of Aarhus

Blattberg R, Gonedes N (1974) Stable and student distributions for stock prices. J Bus 47:244–280

Bookstaber RM, McDonald JB (1987) A general distribution for describing security price returns. J Bus 60(3):401–424

Courtault J-M, Kabanov J, Bru B, Crépel P (2000) Louis Bachelier – On the centenary of théorie de la spéculation. Math Financ 10(3):341–353

Eberlein E, Keller U (1995) Hyperbolic distributions in finance. Bernoulli 1(3):281–299

Fama E (1965) The behaviour of stock market prices. J Bus 38:34–105

Fielitz B (1976) Further results on asymmetric stable distributions of stock prices changes. J Financ Quant Anal 11:39–55

Fischer M (2004) Skew generalized secant hyperbolic distributions: unconditional and conditional fit to asset returns. Austrian J Stat 33(3):293–304

Fischer M (2006) A skew generalized secant hyperbolic family. Austrian J Stat 35(4):437–444

Granger CWJ, Orr R (1972) Infinite variance and research strategy in time series analysis. J Am Stat Assoc 67:275–285

Hansen CB, McDonald JB, Theodossiou P (2003) Some exible parametri models for skewed and leptokurtic data. Working paper, Department of Economics, Brigham Young University

Kon SJ (1984) Models of stock returns – A comparison. J Financ 39(1):147–165

Küchler E, Neumann K, Sørensen M, Streller A (1999) Stock returns and hyperbolic distributions. Math Comp Model 29:1–15

Lau AH, Lau KC, Wingender JR (1990) The distribution of stock returns: new evidence against the stable model. J Bus Econ Stat 8(2):217–223

Mandelbrot BB (1963) The variation of certain speculative prices. J Bus 36:394–519

McDonald JB (1991) Parametric models for partially adaptive estimation with skewed and leptokurtic residuals. Econ Lett 37: 273–278

McDonald JB, Bookstaber RM (1991) Option pricing for generalized distributions. Commun Stat Theory Meth 20(12):4053–4068

McDonald JB, Newey WK (1988) Partially adaptive estimation of regression models via the generalized-t-distribution. Econometr Theory 1(4):428–457

Pagan A (1996) The econometrics of financial markets. J Empirical Financ 3:15–102

Pisarenko V, Sornette D (2006) New statistic for financial return distributions: power-law or exponential? Physika A 366: 387–400
Praetz PD (1972) The distribution of stock price changes. J Bus 45:49–55
Prause K (1999) The generalized hyperbolic model: estimation, financial derivatives and risk measures. PhD thesis, University of Freiburg, Freiburg
Theodossiou P (1998) Financial data and the skewed generalized t distribution. Manag Sci 44:1650–1661
Tucker A (1992) A reexamination of finite- and infinite variance distributions as models of daily stock returns. J Bus Econ Stat 10:73–81
Yu J (2000) Testing for a finite variance in stock return distributions. In: Dunis CL (ed) Advances in quantitative asset management, studies in computational finance vol 1, Chap 6. Kluwer Academic, Amsterdam, pp 143–164

First Exit Time Problem

Christos H. Skiadas[1], Charilaos Skiadas[2]
[1]Professor, Director of the Data Analysis and Forecasting Laboratory
Technical University of Crete, Chania, Greece
[2]Hanover College, Hanover, IN, USA

The first exit time distribution for a stochastic process is the distribution of the times at which particles following this process cross a certain (often linear) barrier. It is often referred to also as hitting time. It is closely related to the probability density function $p(x_t, t)$ of a stochastic process x_t over time t.

For a linear horizontal barrier located at a, the first exit time density function relation is given by: $g(t) = \frac{|a|}{t} p(a, t)$.

For other types barriers (e.g., quadratic), a tangent approximation may be used to obtain a satisfactory estimate as is presented below.

The probability density function may be computed in some cases using the Fokker-Planck equation. In particular in the one-dimensional diffusion problem expressed by a stochastic differential equation of the form:

$$dx_t = \sigma dw_t,$$

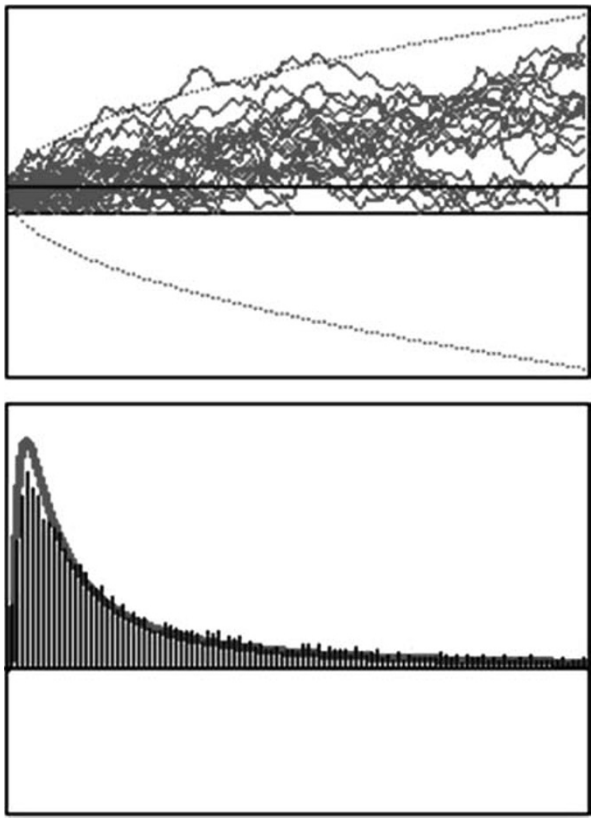
where σ is the variance and w_t is the standard Wiener process, the corresponding Fokker-Planck equation for the probability density function $p(x_t, t)$ associated to the above stochastic differential equation has the form:

$$\frac{\partial p(x_t, t)}{\partial t} = \frac{\sigma^2}{2} \frac{\partial^2 p(x_t, t)}{\partial x_t^2}$$

This partial differential equation form is also known as the one-dimensional heat equation first solved by Joseph Fourier (1822). Later on Fick (1855a; 1855b) applied this equation to express one-dimensional diffusion in solids. Albert Einstein (1905) proposed the same form for the one-dimensional diffusion for solving the Brownian motion process (see ►Brownian Motion and Diffusions). It was the first derivation and application of a probabilistic-stochastic theory to the classical Brownian motion problem that is the movement of a particle or a molecule into a liquid. He resulted in giving the development over space and time of this particle. One year later Smoluchowski (1906) proposed also a theory for solving the Brownian motion problem.

Solving this partial differential equation with the boundary conditions, $p(x_t, 0 : 0, 0) = \delta(x_t, 0)$ and $\frac{\partial p(x_t, t : 0, t)}{\partial x} = 0$ as $x_t \to \infty$ the probability density function p_t for the stochastic process results:

$$p(x_t, t) = \frac{1}{\sigma\sqrt{2\pi t}} e^{-\frac{x_t^2}{2\sigma^2 t}}.$$



First Exit Time Problem. Fig. 1 Linear barrier

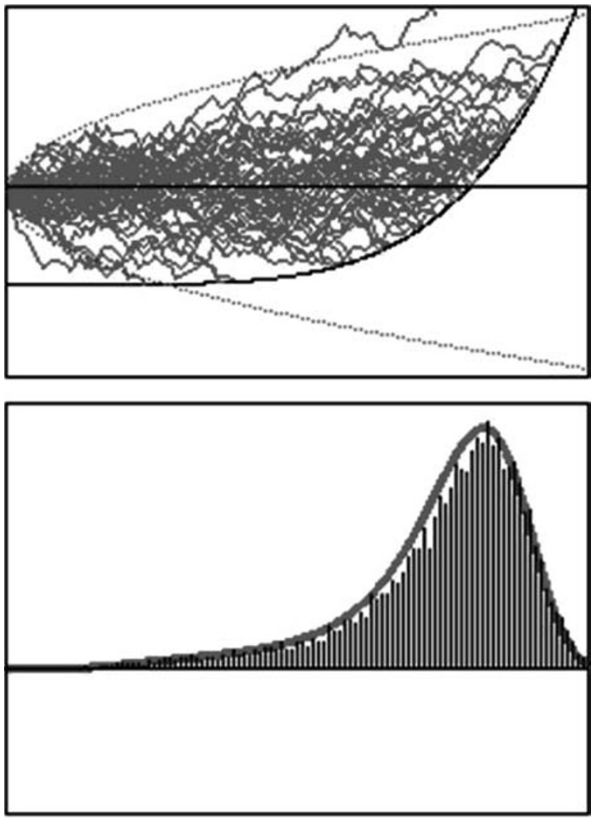
The First Exit Time Density Function

The finding of a density function expressing the distribution of the first exit time of particles escaping from a boundary is due to Schrödinger (1915) and Smoluchowski (1915) in two papers published in the same journal issue. Later on Siegert (1951) gave an interpretation closer to our modern notation whereas Jennen (1985), Lerche (1986) and Jennen and Lerche (1981) gave the most interesting first exit density function form. For the simple case presented earlier the proposed form is:

$$g(t) = \frac{|a|}{t} p(a,t) = \frac{|a|}{\sigma\sqrt{2\pi t^3}} e^{-\frac{a^2}{2\sigma^2 t}}.$$

Jennen (1985) proposed a more general form using a tangent approximation of the first exit density. Application of this theory to the mortality modeling leads to the following form (earlier work can be found in Janssen and Skiadas (1995) and Skiadas and Skiadas (2007):

$$g(t) = \frac{|H_t - tH_t'|}{t} p(t) = \frac{|H_t - tH_t'|}{\sigma\sqrt{2\pi t^3}} e^{-\frac{(H_t)^2}{2\sigma^2 t}}.$$



First Exit Time Problem. Fig. 2 Curved barrier

The last form is associated to the following stochastic process Skiadas (2010):

$$dx_t = \mu_t dt + \sigma dw_t$$

where μ_t is a function of time and there exists a function H_t related to μ_t with the differential equation: $\mu_t = dH_t/dt$.

The associated Fokker–Planck equation is:

$$\frac{\partial p(x_t,t)}{\partial t} = -\mu_t \frac{\partial p(x_t,t)}{\partial x_t} + \frac{\sigma^2}{2} \frac{\partial^2 p(x_t,t)}{\partial x_t^2}$$

and the solution is given by:

$$p(t) = \frac{1}{\sigma\sqrt{2\pi t^3}} e^{-\frac{(H_t)^2}{2\sigma^2 t}}.$$

Two realizations are provided in Figs. 1 and 2. In the first case the first exit time probability density function is provided and stochastic simulations are done for a linear barrier located at α. Figure 2 illustrates the case when a curved barrier is present.

About the Author

For biography *see* the entry ►"Chaotic modelling."

Cross References

►Brownian Motion and Diffusions
►First-Hitting-Time Based Threshold Regression
►Random Walk
►Stochastic Processes

References and Further Reading

Einstein A (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Annalen der Physik 17:549–560

Fick A (1855) Über Diffusion. Poggendorff's Annalen. 94:59–86

Fick A (1855) On liquid diffusion. Philos Mag J Sci 10:31–39

Fourier J (1822) Theorie Analytique de la Chaleur. Firmin Didot, Paris

Fourier J (1878) The analytical theory of heat. Cambridge University Press, New York

Janssen J, Skiadas CH (1995) Dynamic modelling of life-table data. Appl Stoch Model Data Anal 11(1):35–49

Jennen C (1985) Second-order approximation for Brownian first exit distributions. Ann Probab 13:126–144

Jennen C, Lerche HR (1981) First exit densities of Brownian motion through one-sided moving boundaries. Z Wahrsch uerw Gebiete 55:133–148

Lerche HR (1986) Boundary crossing of Brownian motion. Springer-Verlag, Berlin

Schrödinger E (1915) Zur theorie der fall - und steigversuche an teilchenn mit Brownsche bewegung. Phys Zeit 16:289–295

Siegert AJF (1951) On the first passage time probability problem. Phys Rev 81:617–623

Skiadas CH (2010) Exact solutions of stochastic differential equations: Gompertz, generalized logistic and revised exponential. Meth Comput Appl Probab 12(2):261–270

Skiadas CH, Skiadas C (2007) A modeling approach to life table data. In Skiadas CH (ed) Recent advances in stochastic modeling and data analysis. World Scientific, Singapore, pp 350–359
Skiadas C, Skiadas CH (2010) Development, simulation and application of first exit time densities to life table data. Comm Stat Theor Meth 39(3):444–451
Smoluchowski M (1906) Zur kinetischen theorie der Brownschen molekularbewegung und der suspensionen. Ann D Phys 21:756–780
Smoluchowski M (1915) Notizüber die berechning der Brownschen molekular-bewegung bei der ehrenhaft-millikanchen versuchsanordnung. Phys Zeit 16:318–321

First-Hitting-Time Based Threshold Regression

Xin He[1], Mei-Ling Ting Lee[2]
[1]Assistant Professor
University of Maryland, College Park, MD, USA
[2]Professor, Director, Biostatistics and Risk Assessment Center (BRAC)
University of Maryland, College Park, MD, USA

First-hitting-time (FHT) based threshold regression (TR) model is a relatively new methodology for analyzing ►survival data where the time-to-event is modeled as the first time the stochastic process of interest hits a boundary threshold. FHT models have been applied in analyzing the failure time of engineering systems, the length of hospital stay, the survival time of AIDS patients, and the duration of industrial strikes, etc.

First-Hitting-Time (FHT) Model

A first-hitting-time (FHT) model has two basic components, namely a stochastic process $\{Y(t), t \in \mathcal{T}, y \in \mathcal{Y}\}$ with initial value $Y(0) = y_0$, where $\mathcal{T}$ is the time space and $\mathcal{Y}$ is the state space of the process; and a boundary set $\mathcal{B}$, where $\mathcal{B} \subset \mathcal{Y}$. Assume that the initial value of the process y_0 lies outside the boundary set $\mathcal{B}$, then the first hitting time can be defined by the random variable

$$S = \inf\{t : Y(t) \in \mathcal{B}\},$$

where S is the first time the sample path of the stochastic process reaches the boundary set $\mathcal{B}$. In a medical context, the stochastic process $\{Y(t)\}$ may describe a subject's latent health condition or disease status over time t. The boundary set $\mathcal{B}$ represents a medical end point, such as death, or disease onset. Although the boundary set $\mathcal{B}$ is set to be fixed in time in basic FHT models, it may vary with time in some applications. The stochastic process $\{Y(t)\}$ in the FHT model may take many forms. The most commonly used process is a Wiener diffusion process with a positive initial value and a negative drift parameter. Alternative processes including the gamma process, the Ornstein-Uhlenbeck (OU) process, and the semi-Markov process have also been investigated. For a review, see Lee and Whitmore (2006) and Aalen et al. (2008).

Threshold Regression

Threshold regression (TR) is an extension of the first-hitting-time model by adding regression structures to it so as to accommodate important covariates. The threshold regression model does not required the proportional hazards assumption and hence it provides an alternative model for analyzing time-to-event data. The unknown parameters in the stochastic process $\{Y(t)\}$ and the boundary set $\mathcal{B}$ are connected to covariates using suitable regression link functions. For example, the initial state y_0 and the drift parameter μ of a Wiener diffusion process $\{Y(t)\}$ can be linked to covariates using general link functions of the form

$$y_0 = g_1(\mathbf{x})$$

and

$$\mu = g_2(\mathbf{x}),$$

where $\mathbf{x}$ is the vector of covariates (Lee et al. 2000; Lee and Whitmore 2006). Pennell et al. (2010) proposed a TR model with Bayesian random effects to account for unmeasured covariates in both the initial state and the drift. Yu et al. (2009) incorporated penalized regression and regression splines to TR models to accommodate semi-parametric nonlinear covariate effects.

Analytical Time Scale

In stead of calendar time, in many applications involving time-dependent cumulative effects, an alternative time scale can be better used in describing the stochastic process. Let $r(t|\mathbf{x})$ denote a monotonic transformation of calendar time t to analytical time r (or referred to as process time) with $r(0|\mathbf{x}) = 0$. In a medical context, the analytical time may be some time-dependent measure to describe cumulative toxic exposure or the progression of disease. The process $\{Y(r)\}$ defined in terms of analytical time r can be expressed as a subordinated process $\{Y[r(t)]\}$ in terms of calendar time t. Lee and Whitmore (1993, 2004) examined the connection between subordinated stochastic processes and analytical time.

Whitmore et al. (1998) proposed a bivariate Wiener model in which failure is governed by a latent process while auxiliary readings are available from a correlated marker process. Lee et al. (2000) extended this model to bivariate threshold regression by including CD4 cell counts as a marker process in the context of AIDS clinical trials.

Tong et al. (2008) generalized the bivariate TR model to current status data. Using Markov decomposition methods, Lee et al. (2010) generalized threshold regression to include time-dependent covariates. Lee and Whitmore (2010) discussed the connections between TR and proportional hazard regressions and demonstrated that proportional hazard functions can be generated by TR models.

About the Author

Professor Lee was named the Mosteller Statistician of the Year in 2005 by the American Statistical Association, Boston Chapter. She is Elected member of the International Statistical Institute (1995), and Elected Fellow of: Royal Statistical Society (1998), American Statistical Association (1999) and the Institute of Mathematical Statistics (2005). Professor Lee is the Founding Editor and Editor-in-Chief of the international journal *Lifetime Data Analysis*.

Cross References

►First Exit Time Problem
►Survival Data

References and Further Reading

Aalen OO, Borgan Ø, Gjessing HK (2008) Survival and event history analysis: a process point of view. Springer, New York

Lee M-LT, Whitmore GA (1993) Stochastic processes directed by randomized time. J Appl Probab 30:302–314

Lee M-LT, Whitmore GA (2004) First hitting time models for liefetime data. In: Rao CR, Balakrishnan N (eds) Advances in survival ananlysis. North Holland, Amsterdam pp 537–543

Lee M-LT, Whitmore GA (2006) Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. Stat Sci 21:501–513

Lee M-LT, Whitmore GA (2010) Proportional hazards and threshold regression: their theoretical and practical connections. Lifetime Data Anal 16:196–214

Lee M-LT, DeGruttola V, Schoenfeld D (2000) A model for markers and latent health status. J R Stat Soc B 62:747–762

Lee M-LT, Whitmore GA, Rosner B (2010) Threshold regression for survival data with time-varying covariates. Stat Med 29: 896–905

Pennell ML, Whitmore GA, Lee M-LT (2010) Bayesian random-effects threshold regression with application to survival data with nonproportional hazards. Biostat 11:111–126

Tong X, He X, Sun J, Lee M-LT (2008) Joint analysis of current status and marker data: an extension of a bivariate threshold model. Int J Biostat 4, Article 21

Whitmore GA, Crowder MJ, Lawless JF (1998) Failure inference from a marker process based on a bivariate Wiener model. Lifetime Data Anal 4:229–251

Yu Z, Tu W, Lee M-LT (2009) A semi-parametric threshold regression analysis of sexually transmitted infections in adolescent women. Stat Med 28:3029–3042

Fisher Exact Test

Peter Sprent
Emeritus Professor
University of Dundee, Dundee, UK

The Fisher Exact test was proposed by Fisher (1934) in the fifth edition of *Statistical Methods for Research Workers*. It is a test for independence as opposed to association in 2×2 contingency tables.

A typical situation where such tables arise is where we have counts of individuals categorized by each of two dichotomous attributes, e.g., one attribute may be religious affiliation dichotomized into Christian and non-Christian and the other marital status recorded as married or single.

Another example is that where one of the attributes that are dichotomized corresponds to treatments, e.g., Drug A prescribed, or Drug B prescribed, and the other attribute is the responses to those treatments, e.g., patient condition improves or patient condition does not improve.

In the latter situation if 9 patients are given Drug A and 12 patients are given drug B we might observe the following counts in cells of a 2×2 table:

	Improvement	No improvement	Row total
Drug A	8	1	9
Drug B	3	9	12
Column totals	11	10	21

Fisher pointed out that if we assume row and column totals are fixed then once we know the entry in any cell of the table (e.g., here 8 in the top left cell) then the entries in the remaining three cells are all fixed by the constraint that the marginal totals are fixed. This is usually expressed by saying the table has one degree of freedom. What Fisher noted is that under the hypothesis of independence, if we assume the marginal totals fixed then the distribution of the numbers in the first cell (or any other cell) has a hypergeometric distribution under independence for any of the common models associated with such a table as described, for example in Agresti (2002) or Sprent and Smeeton (2007). These common models are (1) that responses to each drug, for example, are binomially distributed with a common value for the binomial parameter p or (2) have a common Poisson distribution, or (3)

the four cell counts are a sample from a ►multinomial distribution.

If a general a 2×2 contingency table has cell entries n_{ij} $(i, j = 1, 2)$ and row totals n_{i+} and column totals n_{+j} and grand total of all 4 cell entries is n, then the hypergeometric distribution for the observed cell values has an associated probability

$$\frac{(n_{1+})!(n_{2+})!(n_{+1})!(n_{+2})!}{(n_{11})!(n_{12})!(n_{21})!(n_{22})!n!}$$

To perform the test one calculates these probabilities for all possible n_{11} consistent with the fixed marginal totals and computes the P-value as the sum of all such probabilities that are less than or equal to that associated with the observed configuration. For the numerical example given above n_{11} may take any integral value between 0 and 9 and the following table gives the corresponding hypergeometric probabilities:

n_{11}	Hypergeometric probability
0	0.00003
1	0.00168
2	0.02245
3	0.11788
4	0.28295
5	0.33008
6	0.18862
7	0.05052
8	0.00561
9	0.00019

From this table we see that the observed $n_{11} = 8$ has associated probability $p = 0.00561$ and the only other outcomes with this or a lower probability correspond to $n_{11} = 0, 1$ or 9. Thus the test P-value is $P = 0.00561 + 0.00168 + 0.00019 + 0.00003 = 0.00751$.

This low P-value provides very strong evidence of association, i.e., that the drugs differ in effectiveness.

In practice, except for very small samples appropriate statistical software is required to compute P. When the expected numbers in each cell assuming independence are not too small the standard chi-squared test for contingency tables gives a close approximation to the exact test P-value especially if Yates's correction (*see* Sprent and Smeeton 2007) is used.

The exact test procedure was extended by Freeman and Halton (1951) to tables with any specified numbers of rows and columns.

Some statisticians have argued that it is inappropriate to condition the test statistics on fixed marginal totals, but it is now widely accepted that in most, though not perhaps in all, situations arising in practice such conditioning is appropriate.

About the Author

For biography *see* the entry ►Sign Test.

Cross References

►Chi-Square Test: Analysis of Contingency Tables
►Exact Inference for Categorical Data
►Hypergeometric Distribution and Its Application in Statistics
►Nonparametric Statistical Inference
►Proportions, Inferences, and Comparisons

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York
Fisher RA (1934) Statistical methods for research workers, 5th edn. Oliver & Boyd, Edinburgh
Freeman GH, Halton JH (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika 38:141–149
Sprent P, Smeeton NC (2007) Applied nonparametric statistical methods, 4th edn. Chapman & Hall/CRC, Boca Raton

Fisher-Tippett Theorem

Bojan Basrak
University of Zagreb, Zagreb, Croatia

In 1928, Fisher and Tippett presented a theorem which can be considered as a founding stone of the *extreme value theory*. They identified all ►*extreme value distributions*, which means all possible nondegenerate limit laws for properly centered and scaled partial maxima $M_n = \max\{X_1, \ldots, X_n\}$, where (X_n) is a sequence of independent and identically distributed random variables. More precisely, if there exist real sequences (a_n) and (b_n) where $a_n > 0$ for all n, such that the random variables

$$\frac{M_n - b_n}{a_n} \quad \text{as } n \to \infty,$$

converge in distribution to a nondegenerate random variable with a distribution function G, then G is called an extreme value distribution. The theorem states that G (permitting centering and scaling) necessarily belongs to one of the following three classes: *Fréchet, Gumbel*, and *Weibull distributions*.

Rigorous proofs of the theorem appearing in contemporary literature are due to Gnedenko in 1943, and works of de Haan and Weissman in 1970s. The class of ►extreme value distributions coincides with the class of *max–stable distributions*, i.e. those distributions of the random variable X_1, for which there exist real constants $c_n > 0$ and d_n for each $n \geq 1$, such that $(M_n - d_n)/c_n$ has the same distribution as X_1.

To determine whether partial maxima of a given family of random variables, after scaling and centering, has asymptotically one of those distributions is one of the main tasks of extreme value analysis. Many of such questions are answered using the notion of regular variation which was introduced in mathematical analysis by Karamata, a couple of years after the publication of Fisher–Tippett theorem.

Cross References

►Extreme Value Distributions
►Statistics of Extremes

References and Further Reading

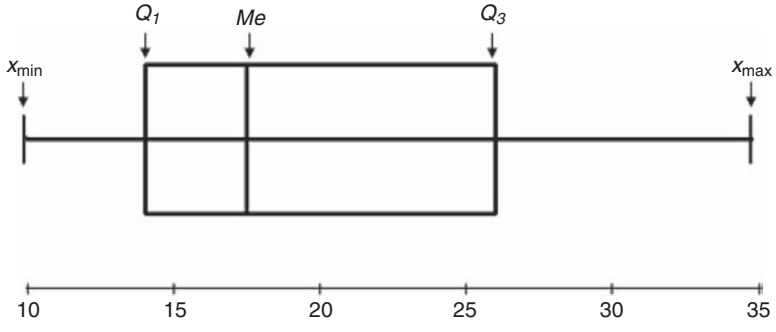
de Haan L, Ferreira A (2006) Extreme value theory: an introduction. Springer, New York
Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events for insurance and finance. Springer, Berlin
Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc Camb Philos Soc 24:180–190
Gnedenko B (1943) Sur la distribuion limite du terme maximum d'une série aléatoire. Ann Math 44:423–453
Resnick SI (1987) Extreme values, regular variation, and point processes. Springer, New York

Five-Number Summaries

Mirjana Čižmešija
Professor, Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

The five-number summary $(5'S)$ is a technique of exploratory data analyses developed with the aim of investigating one or more data sets. It consists of five descriptive measures (Anderson 2007): minimum value $(x_{\min})$, first quartile (Q_1), median (Me), third quartile (Q_3), and maximum value $(x_{\max})$. The graphical presentation of the five-number summary is the box-and-whisker plot (box-plot) developed by John Tukey (Levine 2008). In determining the $5'S$, the data set of observations on a single variable must be arranged from the smallest to the largest value, and therefore the $5'S$ are arranged as follows: $x_{\min} \leq Q_1 \leq Me \leq Q_3 \leq x_{\max}$. Each of these five parameters is important in descriptive statistics for providing information about the dispersion and skew of data sets. ►Outliers in the data set may be detected in the box-plot. In measuring dispersion, the distance between the minimum and maximum value is important (particularly in financial analyses).

The difference between the first and third quartile is the range of the middle 50% of the data in the data set (interquartile range). These differences and the differences between quartiles and the median are important in detecting the shape of the data set. In a symmetrical distribution, the difference between the first quartile and the minimum value is the same as the difference between the maximum value and the third quartile, and the difference between the median and the first quartile is the same as the difference between the third quartile and the median. In a right-skewed distribution, the difference between the first quartile and the minimum value is smaller than the



Five-Number Summaries. Fig. 1 Box-plot

difference between the maximum value and the third quartile, and the difference between the median and the first quartile is smaller than the difference between the third quartile and the median. In a left-skewed distribution, the difference between the first quartile and the minimum value is greater than the difference between the maximum value and the third quartile and the difference between the median and the first quartile is greater than the difference between the third quartile and the median. The five-number summary is a useful tool in comparing the dispersion of two or more data sets.

For example, the following data set

$$10, 11, 14, 15, 17, 18, 20, 26, 26, 35$$

can be described as $5'S$:

$$x_{\min} = 10, \quad Q_1 = 14, \quad Me = 17,5, \quad Q_3 = 26, \quad x_{\max} = 35$$

and graphically displayed by the box-plot in the Fig. 1.

Cross References

►Data Analysis
►Summarizing Data with Boxplots

References and Further Reading

Anderson DR, Sweeney DJ, Williams TA, Freeman J, Shoesmith E (2007) Statistics for business and economics. Thomson, London

Levine DM, Stephan DF, Krehbiel FC, Berenson ML (2008) Statistics for managers using Microsoft Excel, 5th edn. Personal Education International Upper Saddle River

Forecasting Principles

Kesten C. Green[1], Andreas Graefe[2], J. Scott Armstrong[3]
[1]Associate Professor
University of South Australia, Adelaide, SA, Australia
[2]Karlsruhe Institute of Technology, Karlsruhe, Germany
[3]Professor of Marketing
University of Pennsylvania, Philadelphia, PA, USA

Introduction

Forecasting is concerned with making statements about the as yet unknown. There are many ways that people go about deriving forecasts. This entry is concerned primarily with procedures that have performed well in empirical studies that contrast the accuracy of alternative methods.

Evidence about forecasting procedures has been codified as condition-action statements, rules, guidelines or, as we refer to them, *principles*. At the time of writing there are 140 principles. Think of them as being like a safety checklist for a commercial airliner – if the forecast is important, it is important to check all relevant items on the list. Most of these principles were derived as generalized findings from empirical comparisons of alternative forecasting methods. Interestingly, the empirical evidence sometimes conflicts with common beliefs about how to forecast.

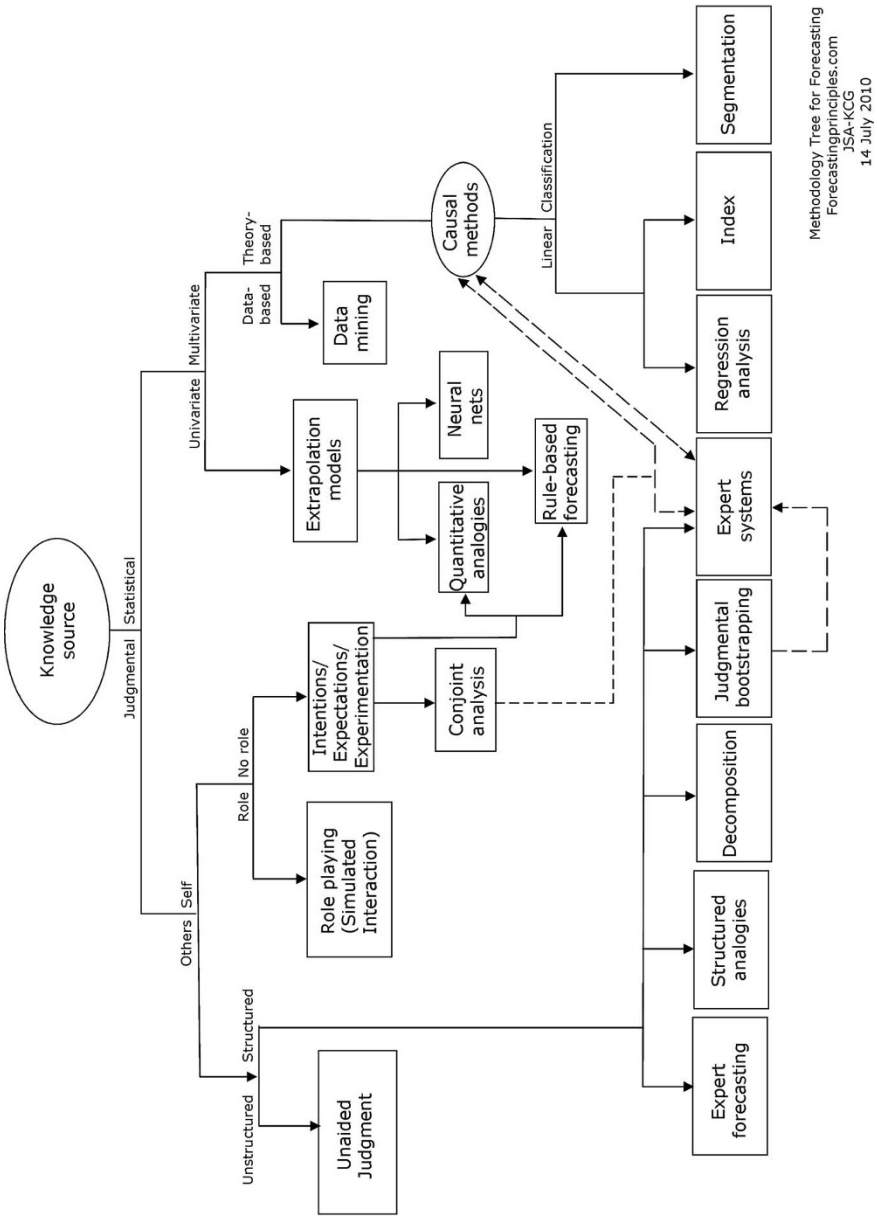
Primarily due to the strong emphasis placed on empirical comparisons of alternative methods, researchers have made many advances in forecasting since 1980. The most influential paper in this regard is the M-competition paper (Makridakis et al. 1982). This was based on a study in which different forecasters were invited to use what they thought to be the best method to forecast many time series. Entry into the competition required that methods were fully disclosed. Entrants submitted their forecasts to an umpire who calculated the errors for each method. This was only one in a series of M-competition studies, the most recent being Makridakis and Hibon (2000). For a summary of the progress that has been made in forecasting since 1980, see Armstrong (2006).

We briefly describe valid forecasting methods, provide guidelines for the selection of methods, and present the *Forecasting Canon* of nine overarching principles. The *Forecasting Canon* provides a gentle introduction for those who do not need to become forecasting experts but who nevertheless rightly believe that proper knowledge about forecasting would help them to improve their decision making. Those who wish to know more can find what they seek in *Principles of Forecasting: A Handbook for Practitioners and Researchers*, and at the Principles of Forecasting Internet site (ForPrin.com).

Forecasting Methods

As shown in Fig. 1, the *Forecasting Methodology Tree*, forecasting methods can be classified into those that are based primarily on judgmental sources of information and those that use statistical data. There is overlap between some judgmental and statistical approaches.

If available data are inadequate for quantitative analysis or qualitative information is likely to increase the accuracy, relevance, or acceptability of forecasts, one way to make forecasts is to ask experts to think about a situation and predict what will happen. If experts' forecasts are not derived using structured forecasting methods, their forecasting method is referred to as *unaided judgment*. This is the most commonly used method. It is fast, inexpensive



Forecasting Principles. Fig. 1 Methodology tree

when few forecasts are needed, and may be appropriate when small changes are expected. It is most likely to be useful when the forecaster knows the situation well and gets good feedback about the accuracy of his forecasts (e.g., weather forecasting, betting on sports, and bidding in bridge games).

Expert forecasting refers to forecasts obtained in a structured way from two or more experts. The most appropriate method depends on the conditions (e.g., time constraints, dispersal of knowledge, access to experts, expert motivation, need for confidentiality). In general, diverse experts should be recruited, questions should be chosen carefully and tested, and procedures for combining across experts (e.g., the use of medians) should be specified in advance.

The *nominal group technique* (NGT) tries to account for some of the drawbacks of traditional meetings by imposing a structure on the interactions of the experts. This process consists of three steps: First, group members work independently and generate individual forecasts. The group then conducts an unstructured discussion to deliberate on the problem. Finally, group members work independently and provide their final individual forecasts. The NGT forecast is the mean or median of the final individual estimates.

Where group pressures are a concern or physical proximity is not feasible, the *Delphi method,* which involves at least two rounds of anonymous interaction, may be useful. Instead of direct interaction, individual forecasts and arguments are summarized and reported as feedback to participants after each round. Taking into account this information, participants provide a revised forecast for the next round. The Delphi forecast is the mean or median of the individual forecasts in the final round. Rowe and Wright (2001) found that Delphi improved accuracy over unstructured groups in five studies, harmed accuracy in one, and the comparison was inconclusive in two. Delphi is most suitable if experts are expected to possess different information, but it can be conducted as a simple one-round survey for situations in which experts possess similar information. A free version of the Delphi software is available at ForPrin.com.

In situations where dispersed information frequently becomes available, *prediction markets* can be useful for providing continuously updated numerical or probability forecasts. In a prediction market, mutually anonymous participants reveal information by trading contracts whose prices reflect the aggregated group opinion. Incentives to participate in a market may be monetary or nonmonetary. Although prediction markets seem promising, to date there has been no published ▶meta-analysis of the method's accuracy. For a discussion of the relative advantages of prediction markets and Delphi, see Green et al. (2007).

With *structured analogies*, experts identify situations that are analogous to a target situation, identify similarities and differences to the target situation, and determine an overall similarity rating. The outcome or decision implied by each expert's top-rated analogy is used as the structured analogies forecast. Green and Armstrong (2007) analyzed structured analogies for the difficult problem of forecasting decisions people will make in conflict situations. When experts were able to identify two or more analogies and their closest analogy was from direct experience, 60% of structured analogies forecasts were accurate compared to 32% of experts' unaided judgment forecasts, the latter being little better than guessing.

Decomposition involves breaking down a forecasting problem into components that are easier to forecast. The components may either be multiplicative (e.g., to forecast a brand's sales, one could estimate total market sales and market share) or additive (estimates could be made for each type of product when forecasting new product sales for a division). Decomposition is most likely to be useful in situations involving high uncertainty, such as when predicting large numbers. MacGregor (2001) summarized results from three studies involving 15 tests and found that judgmental decomposition led to a 42% reduction in error under high levels of uncertainty.

Judgmental bootstrapping derives a model from knowledge of experts' forecasts and the information experts used to make their forecasts. This is typically done by regression analysis. It is useful when expert judgments have validity but data are scarce (e.g., forecasting new products) and outcomes are difficult to observe (e.g., predicting performance of executives). Once developed, judgmental bootstrapping models are a low-cost forecasting method. Armstrong (2001a) found judgmental bootstrapping to be more accurate than unaided judgment in 8 of 11 comparisons. Two tests found no difference, and one found a small loss in accuracy.

Expert systems are based on rules for forecasting that are derived from the reasoning experts use when making forecasts. They can be developed using knowledge from diverse sources such as surveys, interviews of experts, protocol analysis in which the expert explains what he is doing as he makes forecasts, and research papers. Collopy et al. (2001) summarized evidence from 15 comparisons that included expert systems. Expert systems were more accurate than unaided judgment in six comparisons, similar in one, and less accurate in another. Expert systems were less accurate than judgmental bootstrapping in two

comparisons and similar in two. Expert systems were more accurate than econometric models in one comparison and as accurate in two.

It may be possible to ask people directly to predict how they would behave in various situations. However, this requires that people have valid *intentions* or *expectations* about how they would behave. Both are most useful when (1) responses can be obtained from a representative sample, (2) responses are based on good knowledge, (3) people have no reason to lie, and (4) new information is unlikely to change behavior. Intentions are more limited than expectations in that they are most useful when (5) the event is important, (6) the behavior is planned, and (7) the respondent can fulfill the plan (e.g., their behavior is not dependent on the agreement of others). Better yet, in situations in which it is feasible to do so, conduct an experiment by changing key causal variables in a systematic way, such that the independent variables are not correlated with one another. Estimate relationships from responses to the changes and use these estimates to derive forecasts.

Role playing involves asking people to think and behave in ways that are consistent with a role and situation described to them. Role playing for the purpose of predicting the behavior of people with different roles who are interacting with each other is called *simulated interaction*. Role players are assigned roles and asked to act out prospective interactions in a realistic manner. The decisions are used as forecasts of the actual decision. Green (2005) found that 62% of simulated interaction forecasts were accurate for eight diverse conflict situations. By comparison, 31% of forecasts from the traditional approach – expert judgments unaided by structured techniques – were accurate. Game theory experts' forecasts were no better, also 31%, and both unaided judgment and game theory forecasts were little better than chance at 28% accurate.

Conjoint analysis is a method for eliciting people's preferences for different possible offerings (e.g., for alternative mobile phone designs or for different political platforms) by using combinations of features (e.g., size, camera, and screen of a mobile phone.) The possibilities can be set up as experiments where each variable is unrelated to the other variable. Regression-like analyses are then used to predict the most desirable design.

Extrapolation models use time-series data on the situation of interest (e.g., data on automobile sales from 1940–2009) or relevant cross-sectional data. For example, exponential smoothing, which relies on the principle that more recent data is weighted more heavily, can be used to extrapolate over time. Quantitative extrapolation methods do not harness people's knowledge about the data but assume that the causal forces that have shaped history will continue. If this assumption turns out to be wrong, forecast errors can be large. As a consequence, one should only extrapolate trends when they correspond to the prior expectations of domain experts. Armstrong (2001b) provides guidance on the use of extrapolation.

Quantitative analogies are similar to structured analogies. Experts identify analogous situations for which time-series or cross-sectional data are available, and rate the similarity of each analogy to the data-poor target situation. These inputs are used to derive a forecast. This method is useful in situations with little historical data. For example, one could average data from cinemas in suburbs identified by experts as similar to a new (target) suburb in order to forecast demand for cinema seats in the target suburb.

Rule-based forecasting is an expert system for combining expert domain knowledge and statistical techniques for extrapolating time series. Most series features can be identified automatically, but experts are needed to identify some features, particularly causal forces acting on trends. Collopy and Armstrong (1992) found rule-based forecasting to be more accurate than extrapolation methods.

If data are available on variables that might affect the situation of interest, causal models are possible. Theory, prior research, and expert domain knowledge provide information about relationships between the variable to be forecasted and explanatory variables. Since causal models can relate planning and decision-making to forecasts, they are useful if one wants to create forecasts that are conditional upon different states of the environment. More important, causal models can be used to forecast the effects of different policies.

Regression analysis involves estimating causal model coefficients from historical data. Models consist of one or more regression equations used to represent the relationship between a dependent variable and explanatory variables. Regression models are useful in situations with few variables and many reliable observations where the causal factors vary independently of one another. Important principles for developing regression (econometric) models are to (1) use prior knowledge and theory, not statistical fit, for selecting variables and for specifying the directions of effects (2) use simple models, and (3) discard variables if the estimated relationship conflicts with theory or prior evidence.

Real-world forecasting problems are, however, more likely to involve few observations and many relevant variables. In such situations, the *index method* can be used. Index scores are calculated by adding the values of the explanatory variables, which may be assessed subjectively,

for example as zero or one, or may be normalized quantitative data. If there is good prior domain knowledge, explanatory variables may be weighted relative to their importance. Index scores can be used as forecasts of the relative likelihood of an event. They can also be used to predict numerical outcomes, for example by regressing index scores against historical data.

Segmentation is useful when a heterogeneous whole can be divided into homogenous parts that act in different ways in response to changes, and that can be forecasted more accurately than the whole. For example, in the airline industry, price has different effects on business and personal travelers. Appropriate forecasting methods can be used to forecast individual segments. For example, separate regression models can be estimated for each segment. Armstrong (1985:287) reported on three comparative studies on segmentation. Segments were forecasted either by extrapolation or regression analysis. Segmentation improved accuracy for all three studies.

Selection of Methods

The Forecasting Method Selection Tree, shown in Fig .2, provides guidance on selecting the best forecasting method for a given problem. The Tree has been derived from evidence-based principles. Guidance is provided in response to the user's answers to questions about the availability of data and state of knowledge about the situation for which forecasts are required. The first question is whether sufficient objective data are available to perform statistical analyses. If not, the forecaster should use judgmental methods.

In deciding among judgmental procedures, one must assess whether the future is likely to be substantially different from the past, whether the situation involves decision makers who have conflicting interests, and whether policy analysis is required. Other considerations affecting the selection process are whether forecasts are made for recurrent and well-known problems, whether domain knowledge is available, and whether information about similar types of problems is available.

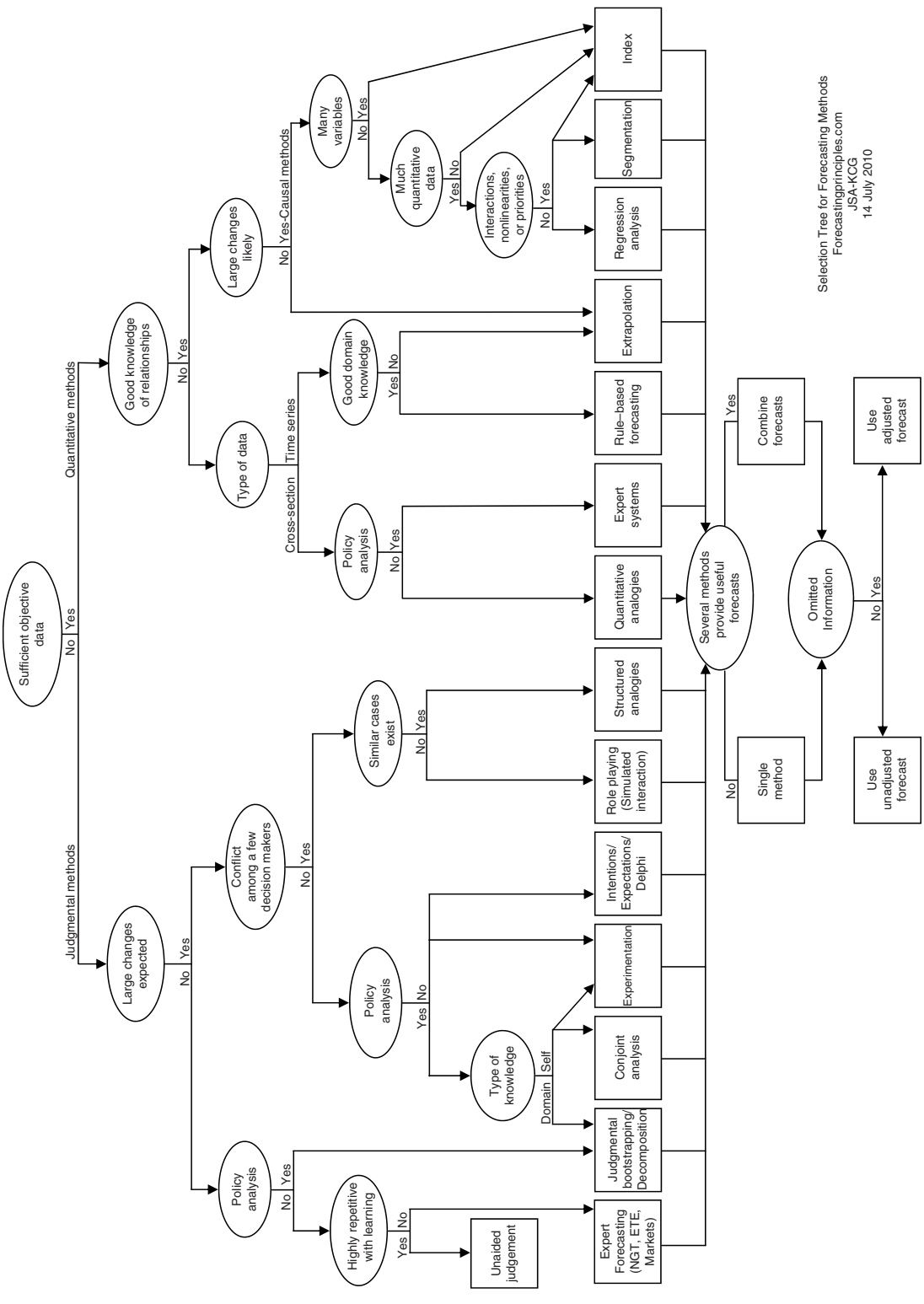
If, on the other hand, much objective data are available and it is possible to use quantitative methods, the forecaster first has to assess whether there is useful knowledge about causal relationships, whether cross-sectional or time-series data are available, and whether large changes are involved. In situations with little knowledge about empirical relationships, the next issues are to assess whether policy analysis is involved and whether there is expert domain knowledge about the situation. If there is good prior knowledge about empirical relationships and the future can be expected to substantially differ from the past, the number of variables and presence or absence of intercorrelation between them, and the number of observations determine which causal method to use. For example, regression models that rely on non-experimental data can typically use no more than three or four variables – even with massive sample sizes. For problems involving many causal variables, variable weights should not be estimated from the dataset. Instead it is useful to draw on independent sources of evidence (such as empirical studies and prior expert knowledge) for assessing the impact of each variable on the situation.

The Forecasting Method Selection Tree provides guidance but on its own, the guidance is not comprehensive. Forecasters may have difficulty identifying the conditions that apply. In such situations, one should use different methods that draw on different information and combine their forecasts according to pre-specified rules. Armstrong (2001c) conducted a meta-analysis of 30 studies and estimated that the combined forecast yielded a 12% reduction in error compared to the average error of the components; the reductions of forecast error ranged from 3% to 24%. In addition, the combined forecasts were often more accurate than the most accurate component. Studies since that meta-analysis suggest that under favorable conditions (many forecasts available for a number of different valid methods and data sources when forecasting for an uncertain situation), the error reductions from combining are much larger. Simple averages are a good starting point but differential weights may be used if there is strong evidence about the relative accuracy of the method. Combining forecasts is especially useful if the forecaster wants to avoid large errors and if there is uncertainty which method will be most accurate.

The final issue is whether there is important information that has not been incorporated in the forecasting methods. This includes situations in which recent events are not reflected in the data, experts possess good domain knowledge about future events or changes, or key variables could not be included in the model. In the absence of these conditions, one should not adjust the forecast. If important information has been omitted and adjustments are needed, one should use a structured approach. That is, provide written instructions, solicit written adjustments, request adjustments from a group of experts, ask for adjustments to be made prior to seeing the forecast, record reasons for the revisions, and examine prior forecast errors.

Forecasting Canon

The Forecasting Canon provides a summary of evidence-based forecasting knowledge, in this case in the form of



Forecasting Principles. Fig. 2 Selection tree

nine overarching principles that can help to improve forecast accuracy. The principles are often ignored by organizations, so attention to them offers substantial opportunities for gain.

Match the Forecasting Method to the Situation

Conditions for forecasting problems vary. No single best method works for all situations. The Forecasting Method Selection Tree (Fig. 2) can help identify appropriate forecasting methods for a given problem. The recommendations in the Selection Tree are based on expert judgment grounded in research studies. Interestingly, generalizations based on empirical evidence sometimes conflict with common beliefs about which forecasting method is best.

Use Domain Knowledge

Managers and analysts typically have useful knowledge about situations. While this domain knowledge can be important for forecasting, it is often ignored. Methods that are not well designed to incorporate domain knowledge include exponential smoothing, stepwise regression, ►data mining and ►neural networks.

Managers' expectations are particularly important when their knowledge about the direction of the trend in a time series conflicts with historical trends in the data (called "contrary series"). If one ignores domain knowledge about contrary series, large errors are likely.

A simple rule can be used to obtain much of the benefit of domain knowledge: when one encounters a contrary series, do not extrapolate a trend. Instead, extrapolate the latest value – this approach is known as the naive or no-change model.

Structure the Problem

One of the basic strategies in management research is to break a problem into manageable pieces, solve each piece, then put them back together. This decomposition strategy is effective for forecasting, especially when there is more knowledge about the pieces than about the whole. Decomposition is particularly useful when the forecasting task involves extreme (very large or very small) numbers.

When contrary series are involved and the components of the series can be forecasted more accurately than the global series, using causal forces to decompose the problem increases forecasting accuracy. For example, to forecast the number of people who die on the highways each year, forecast the number of passenger miles driven (a series that is expected to grow) and the death rate per million passenger miles (a series that is expected to decrease) and then multiply these forecasts.

Model the Experts' Forecasts

Expert systems represent forecasts made by experts and can reduce the costs of repetitive forecasts while improving accuracy. However, expert systems are expensive to develop.

An inexpensive alternative to expert systems is judgmental bootstrapping. The general proposition borders on the preposterous; it is that a simple model of the man will be more accurate than the man. The reasoning is that the model applies the man's rules more consistently than the man can.

Represent the Problem Realistically

Start with the situation and develop a realistic representation. This generalization conflicts with common practice, in which one starts with a model and attempt to generalize to the situation. Realistic representations are especially important when forecasts based on unaided judgment fail. Simulated interaction is especially useful for developing a realistic representation of a problem.

Use Causal Models When You Have Good Information

Good information means that the forecaster (1) understands the factors that have an influence on the variable to forecast and (2) possesses enough data to estimate a regression model. To satisfy the first condition, the analyst can obtain knowledge about the situation from domain knowledge and from prior research. Thus, for example, an analyst can draw upon quantitative summaries of research (meta-analyses) on price or advertising elasticities when developing a sales-forecasting model. An important advantage of causal models is that they reveal the effects of alternative decisions on the outcome, such as the effects of different prices on sales. Index models are a good alternative when there are many variables and insufficient data for regression analysis.

Use Simple Quantitative Methods

Complex models are often misled by noise in the data, especially in uncertain situations. Thus, using simple methods is important when there is much uncertainty about the situation. Simple models are easier to understand than complex models, and are less prone to mistakes. They are also more accurate than complex models when forecasting for complex and uncertain situations – which is the typical situation for the social sciences.

Be Conservative When Uncertain

One should make conservative forecasts for uncertain situations. For cross-sectional data, this means staying close to

the typical behavior (often called the "base rate"). In time series, one should stay close to the historical average. If the historical trend is subject to variations, discontinuities, and reversals, one should be cautious with extrapolating the historical trend. Only when a historical time series show a long steady trend with little variation should one extrapolate the trend into the future.

Combine Forecasts

Combining is especially effective when different forecasting methods are available. Ideally, one should use as many as five different methods, and combine their forecasts using a predetermined mechanical rule. Lacking strong evidence that some methods are more accurate than others, one should use a simple average of forecasts.

Conclusion

This entry gives an overview of methods and principles that are known to reduce forecast error. The Forecasting Method Selection Tree provides guidance for which method to use under given conditions. The Forecasting Canon can be used as a simple checklist to improve forecast accuracy. Further information and support for evidence-based forecasting is available from the *Principles of Forecasting* handbook and from the ForecastingPrinciples.com Internet site.

About the Authors

Dr. Green is a developer of the simulated interactions and structured analogies methods, which have been shown to provide more accurate forecasts of decisions in conflicts than does expert judgment, including the judgments of game theorists. He has been consulted by the U.S. Department of Defense and National Security Agency on forecasting matters. He is co-director of the Forecasting Principles site (ForPrin.com).

Dr. Graefe is the prediction markets editor of *Foresight - The International Journal of Applied Forecasting*. He is currently developing and testing the index method as an alternative to regression analysis, with applications to election forecasting.

Dr. Armstrong has been involved in forecasting since 1960. He has published *Long-Range Forecasting* (1978, 1985) and *Principles of Forecasting* (2001). He is a co-founder of the *Journal of Forecasting* (1982), the *International Journal of Forecasting* (1985), and the *International Symposium on Forecasting* (1981). He is a developer of new forecasting methods including: rule-based forecasting and causal forces for extrapolation. His book, *Persuasive Advertising* was published in 2010.

Cross References

►Business Forecasting Methods
►Forecasting: An Overview
►Time Series

References and Further Reading

Armstrong JS (1985) Long-range forecasting. Wiley, New York
Armstrong JS (2006) Findings from evidence-based forecasting: methods for reducing forecast error. Int J Forecasting 22: 583–598
Armstrong JS (2001a) Judgmental bootstrapping: inferring experts' rules for forecasting. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 171–192
Armstrong JS (2001b) Extrapolation for time-series and cross-sectional data. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 217–243
Armstrong JS (2001c) Combining forecasts. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 417–440
Collopy F, Armstrong JS (1992) Rule-based forecasting: development and validation of an expert systems approach to combining time-series extrapolations. Manage Sci 38:1394–1414
Collopy F, Adya M, Armstrong JS (2001) Expert systems for forecasting. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 285–300
Green KC (2005) Game theory, simulated interaction, and unaided judgement for forecasting decisions in conicts: further evidence. Int J Forecasting 21:463–472
Green KC, Armstrong JS (2007) Structured analogies for forecasting. Int J Forecasting 23:365–376
Green KC, Armstrong JS, Graefe A (2007) Methods to elicit forecasts from groups: Delphi and prediction markets compared. Foresight Int J Appl Forecasting 8:17–20
MacGregor DG (2001) Decomposition in judgmental forecasting and estimation. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 107–124
Makridakis S, Hibon M (2000) The M-3 competition: results, conclusions and implications. Int J Forecasting 16:451–476
Makridakis S, Andersen S, Carbone R, Fildes R, Hibon M, Lewandowski R, Newton J, Parzen E, Winkler R (1982) The accuracy of extrapolation (time series) methods: results of a forecasting competition. J Forecasting 1:111–153
Rowe G, Wright G (2001) Expert opinions in forecasting: the role of the Delphi technique. In: Armstrong JS (ed) Principles of forecasting, Kluwer, Boston, pp 125–144

Forecasting with ARIMA Processes

William W. S. Wei
Professor
Temple University, Philadelphia, PA, USA

One of the most important objectives in the analysis of a time series is to forecast its future values. Let us consider

the time series Z_t from the general ARIMA(p,d,q) process

$$\phi(B)(1-B)^d \dot{Z}_t = \theta(B)a_t, \tag{1}$$

where $\dot{Z}_t = (Z_t - \mu)$ if $d = 0$ and $\dot{Z}_t = Z_t$ when $d \neq 0$, $\phi(B) = (1-\phi_1 B - \cdots - \phi_p B^p)$, $\theta(B) = (1-\theta_1 B - \cdots - \theta_q B^q)$, $\phi(B) = 0$ and $\theta(B) = 0$ share no common roots that lie outside of the unit circle, and the series a_t is a Gaussian $N\left(0, \sigma_a^2\right)$ white noise process.

Minimum Mean Square Error Forecasts and Forecast Limits

Our objective is to derive a forecast with as small an error as possible. Thus, our optimum forecast will be the forecast that has the minimum mean square forecast error. Let us consider the case when $d = 0$ in Eq. (1), and express the process in the moving average representation

$$\dot{Z}_t = \psi(b)a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}, \tag{2}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = \theta(B)/\phi(B)$, and $\psi_0 = 1$. More specifically, the ψ_j can be obtained from equating the coefficients of B^j on the both sides of

$$\begin{aligned}(1-\phi_1 B - \cdots - \phi_p B^p)(1+\psi_1 B + \psi_2 B^2 + \cdots)\\ = (1-\theta_1 B - \cdots - \theta_q B^q).\end{aligned} \tag{3}$$

For $t = n + \ell$, we have $\dot{Z}_{n+\ell} = \sum_{j=0}^{\infty} \psi_j a_{n+\ell-j}$. Suppose that at time $t = n$ we have the observations $\dot{Z}_n, \dot{Z}_{n-1}, \dot{Z}_{n-2}, \ldots$ and wish to forecast ℓ-step ahead future values of $\dot{Z}_{n+\ell}$ as a linear combination of the observations $\dot{Z}_n, \dot{Z}_{n-1}, \dot{Z}_{n-2}, \ldots$. Since $\dot{Z}_t$ for $t \leq n$ can all be written in the form of (2), we can let the minimum mean square error forecast $\hat{\dot{Z}}_n(\ell)$ of $\dot{Z}_{n+\ell}$ be

$$\hat{\dot{Z}}_n(\ell) = \psi^*_\ell a_n + \psi^*_{\ell+1} a_{n-1} + \psi^*_{\ell+2} a_{n-2} + \cdots \tag{4}$$

where the ψ^*_j are to be determined. The mean square error of the forecast is

$$E\left[\dot{Z}_{n+\ell} - \hat{\dot{Z}}_n(\ell)\right]^2 = \sigma_a^2 \sum_{j=0}^{\ell-1} \psi_j^2 + \sigma_a^2 \sum_{j=0}^{\infty} \left[\psi_{\ell+j} - \psi^*_{\ell+j}\right]^2,$$

which is easily seen to be minimized when $\psi^*_{\ell+j} = \psi_{\ell+j}$. Hence,

$$\hat{\dot{Z}}_n(\ell) = \psi_\ell a_n + \psi_{\ell+1} a_{n-1} + \psi_{\ell+2} a_{n-2} + \cdots = E(\dot{Z}_{n+\ell} | \dot{Z}_t, t \leq n). \tag{5}$$

$\hat{\dot{Z}}_n(\ell)$ is usually read as the ℓ-*step* ahead forecast of $\dot{Z}_{n+\ell}$ at the forecast origin n.

The forecast error is

$$e_n(\ell) = \dot{Z}_{n+\ell} - \hat{\dot{Z}}_n(\ell) = \sum_{j=0}^{\ell-1} \psi_j a_{n+\ell-j}. \tag{6}$$

Because $E(e_n(\ell)) = 0$ the forecast is unbiased with the error variance

$$Var(e_n(\ell)) = \sigma_a^2 \sum_{j=0}^{\ell-1} \psi_j^2. \tag{7}$$

For a normal process, the $100(1-\alpha)\%$ forecast limits are

$$\hat{\dot{Z}}_n(\ell) \pm N_{\alpha/2} \left[1 + \sum_{j=0}^{\ell-1} \psi_j^2\right]^{1/2} \sigma_a, \tag{8}$$

where $N_{\alpha/2}$ is the standard normal deviate such that $P(N > N_{\alpha/2}) = \alpha/2$.

For a general ARIMA model in (1) with $d \neq 0$ the moving average representation does not exist because when we obtain the ψ_j from equating the coefficients of B^j on the both sides of

$$\begin{aligned}(1-\phi_1 B - \cdots - \phi_p B^p)(1-B)^d (1+\psi_1 B + \psi_2 B^2 + \cdots)\\ = (1-\theta_1 B - \cdots - \theta_q B^q),\end{aligned} \tag{9}$$

the resulting series of ψ_j coefficients is not convergent. However, for practical purposes, one can use Eq. (9) to find a finite number of the ψ_j coefficients. The minimum mean square error forecast is also given by $E(\dot{Z}_{n+\ell} | \dot{Z}_t, t \leq n)$ directly through the use of Eq. (1), and Eqs. (6), (7), and (8) hold also for the general ARIMA process. The main difference between the ARMA and ARIMA processes is that $\lim_{\ell \to \infty} \sum_{j=0}^{\ell-1} \psi_j^2$ exists for a stationary ARMA process but does not exist for a nonstationary ARIMA process. Hence, the eventual forecast limits for a stationary process approach two horizontal lines. For a nonstationary process since $\sum_{j=0}^{\ell-1} \psi_j^2$ increases as ℓ increases, the forecast limits become wider and wider. It implies that the forecaster becomes less certain about the result as the forecast lead time gets larger.

Computation of Forecasts

The general ARIMA process in Eq. (1) can be written as

$$(1-\Psi_1 B - \cdots - \Psi_{p+d} B^{p+d})\dot{Z}_t = (1-\theta_1 B - \cdots - \theta_q B^q)a_t, \tag{10}$$

where $(1-\Psi_1 B - \cdots - \Psi_{p+d} B^{p+d}) = \phi(B)(1-B)^d$. For $t = n+\ell$ we have

$$\begin{aligned}\dot{Z}_{n+\ell} = \Psi_1 Z_{n+\ell-1} + \cdots + \Psi_{p+d} Z_{n+\ell-p-d} + a_{n+\ell}\\ - \theta_1 a_{n+\ell-1} - \cdots - \theta_q a_{n+\ell-q}.\end{aligned}$$

Taking the conditional expectation at time origin n, we get

$$\begin{aligned}\hat{\dot{Z}}_n(\ell) = \Psi_1 \hat{\dot{Z}}_n(\ell-1) + \cdots + \Psi_{p+d} \hat{\dot{Z}}_n(\ell-p-d) + \hat{a}_n(\ell)\\ - \theta_1 \hat{a}_n(\ell-1) - \cdots - \theta_q \hat{a}_n(\ell-q),\end{aligned} \tag{11}$$

where

$$\hat{\dot{Z}}_n(j) = E(\dot{Z}_{n+j}|\dot{Z}_t, t \le n), \qquad j \ge 1,$$
$$\hat{\dot{Z}}_n(j) = \dot{Z}_{n+j}, \qquad j \le 0,$$
$$\hat{a}_n(j) = 0, \qquad j \ge 1,$$

and

$$\hat{a}_n(j) = \dot{Z}_{n+j} - \hat{\dot{Z}}_{n+j-1}(1) = a_{n+j}, \quad j \le 0.$$

Updating Forecasts

Note that from Eq. (6), we have

$$\begin{aligned} e_n(\ell+1) &= \dot{Z}_{n+\ell+1} - \hat{\dot{Z}}_n(\ell+1) \\ &= \sum_{j=0}^{\ell} \psi_j a_{n+\ell+1-j} = e_{n+1}(\ell) + \psi_\ell a_{n+1} \\ &= \dot{Z}_{n+\ell+1} - \hat{\dot{Z}}_{n+1}(\ell) + \psi_\ell a_{n+1}. \end{aligned}$$

Hence, we obtain the equation for updating forecasts,

$$\hat{\dot{Z}}_{n+1}(\ell) = \hat{\dot{Z}}_n(\ell+1) + \psi_\ell[\dot{Z}_{n+1} - \hat{\dot{Z}}_n(1)]. \tag{12}$$

Eventual Forecast Functions

When $\ell > q, \hat{\dot{Z}}_n(\ell)$ in Eq. (11) becomes

$$\Psi(B)\hat{\dot{Z}}_n(\ell) = 0, \tag{13}$$

where $\Psi(B) = \phi(B)(1-B)^d = 1 - \Psi_1 B - \cdots - \Psi_{p+d}B^{p+d}$, and $B\hat{\dot{Z}}_n(\ell) = \hat{\dot{Z}}_n(\ell-1)$. Thus, we can use the difference equation result to obtain the eventual forecast function. That is, if $\Psi(B) = \prod_{i=1}^{K}(1-R_iB)^{m_i}$ with $\sum_{i=1}^{K} m_i = (p+d)$, then

$$\hat{\dot{Z}}_n(\ell) = \sum_{i=1}^{K}\Big(\sum_{j=0}^{m_i-1} c_{ij}\ell^j\Big)R_i^\ell, \tag{14}$$

for $\ell \ge (q-p-d+1)$ where c_{ij} are constants that are functions of time origin n and known data.

An illustrative example: $(1-\phi_1B)(Z_t-\mu) = (1-\theta_1B)a_t$.

a. Computation of $\hat{Z}_n(\ell)$

For $t = n+\ell, Z_{n+\ell} = \mu + \phi_1(Z_{n+\ell-1}-\mu) + a_{n+\ell} - \theta_1 a_{n+\ell-1}$.

Hence

$$\hat{Z}_n(1) = E(Z_{n+1}|Z_t, t \le n) = \mu + \phi_1(Z_n - \mu) - \theta_1 a_t,$$

and

$$\begin{aligned} \hat{Z}_n(\ell) &= E(Z_{n+\ell}|Z_t, t \le n) \\ &= \mu + \phi_1[\hat{Z}_n(\ell-1) - \mu] \\ &= \mu + \phi_1^\ell(Z_n - \mu) - \phi_1^{\ell-1}\theta_1 a_n, \ell \ge 2. \end{aligned}$$

b. The Forecast Error Variance and Forecast Limits

From Eq. (3), $(1-\phi_1B)(1+\psi_1B+\psi_2B^2+\cdots) = (1-\theta_1B)$, and equating the coefficients of B^j on both sides, we get $\psi_j = \phi_1^{j-1}(\phi_1-\theta_1), j \ge 1$. So

$$\hat{Z}_n(\ell) \pm N_{\alpha/2}\left[1 + \sum_{j=0}^{\ell-1}\left[\phi_1^{j-1}(\phi_1-\theta_1)\right]^2\right]^{1/2}\sigma_a.$$

defines the forecast limits.

c. The Eventual Forecast Function

Since $(1-\phi_1B)(\hat{Z}_n(\ell)-\mu) = 0, \ell \ge 1$, and $|\phi_1| < 1$ we have $\hat{Z}_n(\ell) = \mu + c_1\phi_1^\ell \to \mu$ as $\ell \to \infty$. For more detailed discussions and illustrative examples on time series forecasting, we refer readers to Box, Jenkins, and Reinsel (2008), and Wei (2006).

About the Author

For biography *see* the entry ▶Time Series Regression.

Cross References

▶Box–Jenkins Time Series Models
▶Forecasting: An Overview
▶Structural Time Series Models
▶Time Series

References and Further Reading

Box GEP, Jenkins GM, Reinsel GC (2008) Time series analysis: forecasting and control, 4th edn. Wiley, New York
Wei WWS (2006) Time Series Analysis–Univariate and Multivariate Methods, 2nd edn. Pearson Addison-Wesley, Boston

Forecasting: An Overview

Rob J. Hyndman
Professor of Statistics
Monash University, Melbourne, VIC, Australia

What Can Be Forecast?

Forecasting is required in many situations: deciding whether to build another power generation plant in the next 5 years requires forecasts of future demand; scheduling staff in a call centre next week requires forecasts of call volume; stocking an inventory requires forecasts of stock requirements. Forecasts can be required several years in advance (for the case of capital investments), or only a few minutes beforehand (for telecommunication routing). Whatever the circumstances or time horizons involved, forecasting is an important aid in effective and efficient planning.

Some things are easier to forecast than others. The time of the sunrise tomorrow morning can be forecast very precisely. On the other hand, currency exchange rates are very difficult to forecast with any accuracy. The predictability of an event or a quantity depends on how well we understand the factors that contribute to it, and how much unexplained variability is involved.

Forecasting situations vary widely in their time horizons, factors determining actual outcomes, types of data patterns, and many other aspects. Forecasting methods can be very simple such as using the most recent observation as a forecast (which is called the "naïve method"), or highly complex such as ►neural networks and econometric systems of simultaneous equations. The choice of method depends on what data are available and the predictability of the quantity to be forecast.

Forecasting Methods

Forecasting methods fall into two major categories: quantitative and qualitative methods.

Quantitative forecasting can be applied when two conditions are satisfied:

1. numerical information about the past is available;
2. it is reasonable to assume that some aspects of the past patterns will continue into the future.

There is a wide range of quantitative forecasting methods, often developed within specific disciplines for specific purposes. Each method has its own properties, accuracies, and costs that must be considered when choosing a specific method.

Qualitative forecasting methods are used when one or both of the above conditions does not hold. They are also used to adjust quantitative forecasts, taking account of information that was not able to be incorporated into the formal statistical model. These are not purely guesswork – there are well-developed structured approaches to obtaining good judgmental forecasts. However, as qualitative methods are non-statistical, they will not be considered further in this article.

Explanatory Versus Time Series Forecasting

Quantitative forecasts can be largely divided into two classes: time series and explanatory models. Explanatory models assume that the variable to be forecasted exhibits an explanatory relationship with one or more other variables. For example, we may model the electricity demand (ED) of a hot region during the summer period as

$$\mathrm{ED} = f(\text{current temperature, strength of economy, population, time of day, day of week, error}). \quad (1)$$

The relationship is not exact – there will always be changes in electricity demand that can not be accounted for by the variables in the model. The "error" term on the right allows for random variation and the effects of relevant variables not included in the model. Models in this class include regression models, additive models, and some kinds of neural networks.

The purpose of the explanatory model is to describe the form of the relationship and use it to forecast future values of the forecast variable. Under this model, any change in inputs will affect the output of the system in a predictable way, assuming that the explanatory relationship does not change.

In contrast, time series forecasting uses only information on the variable to be forecast, and makes no attempt to discover the factors affecting its behavior. For example,

$$\mathrm{ED}_{t+1} = f(\mathrm{ED}_t, \mathrm{ED}_{t-1}, \mathrm{ED}_{t-2}, \mathrm{ED}_{t-3}, \ldots, \text{error}), \quad (2)$$

where t is the present hour, $t+1$ is the next hour, $t-1$ is the previous hour, $t-2$ is two hours ago, and so on. Here, prediction of the future is based on past values of a variable and/or past errors, but not on explanatory variables which may affect the system. Time series models used for forecasting include ARIMA models, exponential smoothing and structural models.

There are several reasons for using a time series forecast rather than an explanatory model for forecasting. First, the system may not be understood, and even if it was understood it may be extremely difficult to measure the relationships assumed to govern its behavior. Second, it is necessary to predict the various explanatory variables in order to be able to forecast the variable of interest, and this may be too difficult. Third, the main concern may be only to predict what will happen and not to know why it happens.

A third type of forecasting model uses both time series and explanatory variables. For example,

$$\mathrm{ED}_{t+1} = f(\mathrm{ED}_t, \text{current temperature, time of day, day of week, error}). \quad (3)$$

These types of models have been given various names in different disciplines. They are known as dynamic regression models, panel data models, longitudinal models, transfer function models, and linear system models (assuming f is linear).

The Basic Steps in a Forecasting Task

There are usually five basic steps in any forecasting task.

Step 1: Problem definition. Often this is most difficult part of forecasting. Defining the problem carefully requires

an understanding of how the forecasts will be used, who requires the forecasts, and how the forecasting function fits within the organization requiring the forecasts. A forecaster needs to spend time talking to everyone who will be involved in collecting data, maintaining databases, and using the forecasts for future planning.

Step 2: Gathering information. There are always at least two kinds of information required: (a) statistical data, and (b) the accumulated expertise of the people who collect the data and use the forecasts. Often, a difficulty will be obtaining enough historical data to be able to fit a good statistical model. However, occasionally, very old data will not be so useful due to changes in the system being forecast.

Step 3: Preliminary (exploratory) analysis. Always start by graphing the data. Are there consistent patterns? Is there a significant trend? Is seasonality important? Is there evidence of the presence of business cycles? Are there any ▶outliers in the data that need to be explained by those with expert knowledge? How strong are the relationships among the variables available for analysis?

Step 4: Choosing and fitting models. Which model to use depends on the availability of historical data, the strength of relationships between the forecast variable and any explanatory variables, and the way the forecasts are to be used. It is common to compare two or three potential models.

Step 5: Using and evaluating a forecasting model. Once a model has been selected and its parameters estimated, the model is to be used to make forecasts. The performance of the model can only be properly evaluated after the data for the forecast period have become available. A number of methods have been developed to help in assessing the accuracy of forecasts as discussed in the next section.

Forecast Distributions

All forecasting is about estimating some aspects of the conditional distribution of a random variable. For example, if we are interested in monthly sales denoted by y_t for month t, then forecasting concerns the distribution of y_{t+h} conditional on the values of $y_1, \ldots, y_t$ along with any other information available. Let $\mathcal{I}_t$ denote all other information available at time t. Then we call the distribution of $(y_{t+h} \mid y_1, \ldots, y_t, \mathcal{I}_t)$ the *forecast distribution*.

Typically, a forecast consists of a single number (known as a "point forecast"). This can be considered an estimate of the mean or median of the forecast distribution. It is often useful to provide information about forecast uncertainty as well in the form of a prediction interval. For example, if the forecast distribution is normal with mean $\hat{y}_{t+h}$ and variance σ^2_{t+h}, then a 95% prediction interval for y_{t+h} is $\hat{y}_{t+h} \pm 1.96\sigma^2_{t+h}$. Prediction intervals in forecasting are sometimes called "interval forecasts."

For some problems, it is also useful to estimate the forecast distribution rather than assume normality or some other parametric form. This is called "density forecasting."

Evaluating Forecast Accuracy

It is important to evaluate forecast accuracy using genuine forecasts. That is, it is invalid to look at how well a model fits the historical data; the accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when fitting the model. When choosing models, it is common to use a portion of the available data for testing, and use the rest of the data for fitting the model. Then the testing data can be used to measure how well the model is likely to forecast on new data.

The issue of measuring the accuracy of forecasts from different methods has been the subject of much attention. We summarize some of the approaches here. A more thorough discussion is given by Hyndman and Koehler (2006). In the following discussion, $\hat{y}_t$ denotes a forecast of y_t. We only consider the evaluation of point forecasts. There are also methods available for evaluating interval forecasts and density forecasts (Corradi and Swanson 2006).

Scale-Dependent Errors

The forecast error is simply $e_t = y_t - \hat{y}_t$ which is on the same scale as the data. Accuracy measures that are based on e_t are therefore scale-dependent and cannot be used to make comparisons between series that are on different scales.

The two most commonly used scale-dependent measures are based on the absolute error or squared errors:

$$\text{Mean absolute error (MAE)} = \text{mean}(|e_t|),$$

$$\text{Mean squared error (MSE)} = \text{mean}\left(e_t^2\right).$$

When comparing forecast methods on a single series, the MAE is popular as it is easy to understand and compute.

Percentage Errors

The percentage error is given by $p_t = 100e_t/y_t$. Percentage errors have the advantage of being scale-independent, and so are frequently used to compare forecast performance between different data sets. The most commonly used measure is:

$$\text{Mean absolute percentage error (MAPE)} = \text{mean}(|p_t|)$$

Measures based on percentage errors have the disadvantage of being infinite or undefined if $y_t = 0$ for any t in the period of interest, and having an extremely skewed distribution when any y_t is close to zero. Another problem with percentage errors that is often overlooked is that they assume a meaningful zero. For example, a percentage error makes no sense when measuring the accuracy of temperature forecasts on the Fahrenheit or Celsius scales.

They also have the disadvantage that they put a heavier penalty on positive errors than on negative errors. This observation led to the use of the so-called "symmetric" MAPE proposed by Armstrong (1985, p. 348), which was used in the M3 forecasting competition (Makridakis and Hibon 2000). It is defined by

$$\text{Symmetric mean absolute percentage error (sMAPE)} = \text{mean}\left(200\,|y_t - \hat{y}_t|/(y_t + \hat{y}_t)\right).$$

However, if y_t is zero, $\hat{y}_t$ is also likely to be close to zero. Thus, the measure still involves division by a number close to zero. Also, the value of sMAPE can be negative, so it is not really a measure of "absolute percentage errors" at all. Hyndman and Koehler (2006) recommend that the sMAPE not be used.

Scaled Errors

The MASE was proposed by Hyndman and Koehler (2006) as an alternative to the MAPE or sMAPE when comparing forecast accuracy across series on different scales. They proposed scaling the errors based on the *in-sample* MAE from the naïve forecast method. Thus, a scaled error is defined as

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|},$$

which is independent of the scale of the data. A scaled error is less than one if it arises from a better forecast than the average one-step naïve forecast computed in-sample. Conversely, it is greater than one if the forecast is worse than the average one-step naïve forecast computed in-sample. The *mean absolute scaled error* is simply

$$\text{MASE} = \text{mean}(|q_t|).$$

About the Author

Rob Hyndman is Professor of Statistics and Director of the Business and Economic Forecasting Unit at Monash University, Australia. He has published extensively in leading statistical and forecasting journals. He is co-author of the highly regarded international text on business forecasting, *Forecasting: methods and applications* (Wiley, 3rd edition 1998), and more recently *Forecasting with exponential smoothing: a state space approach* (Springer, 2008). Professor Hyndman is Editor-in-Chief of the *International Journal of Forecasting* and was previously Theory and Methods Editor of the *Australian and New Zealand Journal of Statistics* (2001–2004). He was elected to the International Statistical Institute in 2005. In 2007 he was awarded the prestigious Moran Medal from the Australian Academy of Science, for his contributions to statistical research.

Cross References

►Business Forecasting Methods
►Business Statistics
►Exponential and Holt-Winters Smoothing
►Forecasting Principles
►Forecasting with ARIMA Processes
►Fuzzy Logic in Statistical Data Analysis
►Optimality and Robustness in Statistical Forecasting
►Singular Spectrum Analysis for Time Series
►Statistical Aspects of Hurricane Modeling and Forecasting
►Statistics: An Overview
►Time Series

References and Further Reading

Armstrong JS (1985) Long-range forecasting: from crystal ball to computer. Wiley, New York

Corradi V, Swanson NR (2006) Predictive density evaluation. In: Handbook of economic forecasting, North-Holland, Amsterdam

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. *Int. J. Forecasting* 22(4):679–688

Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008) Forecasting with exponential Smoothing: the state space approach. Springer - Verlag, Berlin

Makridakis S, Hibon M (2000) The M3-competition: results, conclusions and implications. *Int. J. Forecasting* 16:451–476

Makridakis S, Wheelwright SC, Hyndman RJ (1998) Forecasting: methods and applications, 3rd edn. Wiley, New York

Forensic DNA: Statistics in

Wing Kam Fung[1], Yuk Ka Chung[2]
[1]Chair Professor
University of Hong Kong, Hong Kong, China
[2]University of Hong Kong, Hong Kong, China

Introduction

Since its introduction by Sir Alec Jeffreys in 1985, deoxyribonucleic acid (DNA) profiling, or DNA fingerprinting,

has become one of the most important tools in forensic human identification. DNA contains unique genetic information of each organism and can be found in blood, semen, hair/hair root, bone, and body fluids such as saliva and sweat. Theoretically, every individual except for identical twins can be identified by one's unique DNA sequence. However, due to technical limitations, current human identification is not based on fully sequencing the whole genome. Instead, only a number of genetic markers are used, and so the identification cannot be established without any doubt. Statistics thereby plays an important role in assessing the uncertainty in forensic identification and evaluating the weight of DNA evidence.

Random Match Probability

Suppose a crime was committed and a blood stain has been found in the crime scene and a suspect has been identified. The DNA profiles obtained from the crime stain and the blood specimen of the suspect will be compared. A DNA profile is a set of numbers representing the genetic characteristics of the forensic sample, often at nine or more DNA regions called loci. If a perfect match is found between the two DNA profiles, the suspect would not be excluded as a possible contributor to the crime stain. To evaluate the weight of the DNA evidence, the probability that another person would have the same DNA profile will be computed and reported in the courtroom. The smaller the random match probability, the stronger is the evidence to convict the suspect.

A common assumption adopted in the evaluation of the random match probability is that the population is in Hardy–Weinberg equilibrium (HWE), which means the two alleles of a genotype at a particular locus are statistically independent of each other. For example, suppose at a particular locus the alleles found in the profiles of the crime stain and the suspect are in common, say, A_iA_j. The random match probability at this particular locus can be obtained using the product rule as $2p_ip_j$ for $i \neq j$ and p_i^2 for $i = j$, where p_i and p_j are the allele frequencies of A_i and A_j, respectively. Under the assumption of linkage equilibrium, i.e., independence of alleles across all loci, multiplying the individual probabilities over all loci will give the overall random match probability, which is often found as small as one in a million or one in a billion in practice.

In some cases, the suspect is unavailable for typing and a close relative of the suspect is typed instead. In some other cases, the suspect is typed, but the prosecution about who left the crime stain involves a close relative of the suspect. Extensions of the formulas to handle these situations as well as the violation of Hardy–Weinberg and linkage equilibrium are extensively discussed in the literature.

Paternity Determination

Another application of DNA profiling is in kinship determination, which refers to the confirmation of a specific biological relationship between two individuals. In particular, a paternity test determines whether a man is the biological father of an individual. For a standard trio case in which the mother, her child, and the alleged father are typed with DNA profiles denoted by M, C, and AF respectively, the weight of evidence that the alleged father is the biological father of the child is often expressed as a likelihood ratio (LR) of the following hypotheses:

H_p : Alleged father is the biological father of the child.

H_d : The biological father is a random unrelated man.

The LR, also termed as the paternity index (PI) in the context of paternity testing, takes the form

$$LR = PI = \frac{P\left(Evidence|H_p\right)}{P\left(Evidence|H_d\right)} = \frac{P\left(M, C, AF|H_p\right)}{P\left(M, C, AF|H_d\right)}.$$

Using some results on conditional probability and the fact that the genotypes of the mother and the alleged father are not affected by the hypotheses, the index can be simplified to

$$PI = \frac{P\left(C|M, AF, H_p\right)}{P\left(C|M, H_d\right)}.$$

Suppose the genotypes at a particular locus are obtained as $C = A_1A_2$, $M = A_2A_3$, and $AF = A_1A_4$. Since the mother has half chance to pass the allele A_2 to the child and the alleged father also has half chance to pass the allele A_1 to the child under H_p, the numerator of the PI is given by $P\left(C|M, AF, H_p\right) = (1/2)\,(1/2) = 1/4$. Similarly, the denominator can be obtained as $P\left(C|M, H_d\right) = p_1\,(1/2) = p_1/2$ and as a result, $PI = 1/(2p_1)$. The overall paternity index can then be obtained by multiplying the individual PIs over all loci.

It may sometimes be argued that the alleged father is not the biological father of the child, but his relative (say, brother) is, thereby resulting in the following hypotheses:

H_p : Alleged father is the biological father of the child.

H_d : A relative (brother) of the alleged father is the biological father of the child.

The PI can still be computed by using the formula $PI = 1/[2F + 2\,(1 - 2F)\,p_1]$, where F is the kinship coefficient between the alleged father and his relative. The kinship coefficient is a measure of the relatedness between two individuals, representing the probability that two randomly sampled genes from each of them are identical. In particular, $F = 1/4$ for full siblings and, therefore, in this case, $PI = 2/(1 + 2p_1)$, which is substantially smaller

than $1/(2p_1)$, indicating that DNA profiling performs less effective in distinguishing paternity among relatives.

DNA Mixture

In practical crime cases, it is not uncommon that the biological traces collected from the crime scene are obtained as mixed stains, especially in rape cases. In general, the evaluation and interpretation of the mixed DNA sample can be very complicated due to many factors including unknown number of contributors and missing alleles. Here, we consider a simple two-person mixture problem in which the DNA mixture is assumed to be contributed by the victim and only one perpetrator. Suppose that, at a particular locus, the mixture sample contains alleles $M = \{A_1, A_2, A_3\}$, the victim has genotype $V = A_1A_2$, and the suspect has genotype $S = A_3A_3$. The following two competing hypotheses, the prosecution and defense hypotheses, about who contributes to the crime stain are considered:

H_p : The victim and the suspect are the contributors.

H_d : The victim and an unknown person are the contributors.

The weight of the evidence can be evaluated by

$$LR = \frac{P(Evidence|H_p)}{P(Evidence|H_d)} = \frac{P(M,V,S|H_p)}{P(M,V,S|H_d)} = \frac{P(M|V,S,H_p)}{P(M|V,H_d)}$$

where the last expression is obtained after some simplifications. Obviously in this case, $P(M|V,S,H_p) = 1$ as the mixture M is contributed by the victim and the suspect under H_p. Under H_d, the unknown person must have at least one A_3 allele but cannot have alleles not present in the mixture $M = \{A_1, A_2, A_3\}$. Therefore, there are only three possible genotypes for the unknown person at this locus: A_1A_3, A_2A_3, and A_3A_3. Under HWE, $P(M|V,H_d) = 2p_1p_3 + 2p_2p_3 + p_3^2$ and therefore the LR is obtained as

$$LR = \frac{1}{2p_1p_3 + 2p_2p_3 + p_3^2}$$

In the above example, the LR can be easily computed because there are only three possible genotypes for the only unknown person. In general for multiple perpetrator cases, the following general defense hypothesis may be considered:

H_d : The contributors are the victim and x unknown individuals.

For $x = 2$, there are 27 possible genotype configurations of the two unknown individuals and it is cumbersome to list them all. Over the years, general method and formulas for evaluating the LR have been developed in the literature to deal with complicated mixture problems, including situations with the presence of relatives or population substructures.

About the Author

Professor Fung is Past President, Hong Kong Statistical Society, (2003–2004), (2004–2005), (2005–2006), (2006–2007), Past Vice President, International Association for Statistical Computing, (2007–2009). He has received the Outstanding Achievement Award, Ministry of Education, China (2009) and Outstanding Researcher award, The University of Hong Kong (2001). He has been elected a Fellow of the Institute of Mathematical Statistics and a Fellow of the American Statistical Association "for significant contributions to robust statistics and forensic statistics, and for leadership in Asia for statistical research and education."

Cross References

►Bioinformatics
►Data Mining
►Medical Research, Statistics in
►Statistical Genetics
►Statistics and the Law

References and Further Reading

Balding DJ, Nichols RA (1994) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 64:125–140

Brenner C (1997) Symbolic kinship program. Genetics 145:535–542

Evett IW (1992) Evaluating DNA profiles in case where the defense "It is my brother". J Forensic Sci Soc 32:5–14

Evett IW, Weir BS (1998) Interpreting DNA evidence. Sinauer, Sunderland

Fukshansky N, Bär W (2000) Biostatistics for mixed stain: the case of tested relatives of a non-tested suspect. Int J Legal Med 114: 78–82

Fung WK (2003) User-friedly programs for easy calculations in paternity testing and kinship determinations. Forensic Sci Int 136:22–34

Fung WK, Chung YK, Wong DM (2002) Power of exclusion revisited: probability of excluding relatives of the true father from paternity. Int J Legal Med 116:64–67

Fung WK, Hu YQ (2008) Statistical DNA forensics: theory, methods and computation. Wiley, Chichester

Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific 'fingerprints' of human DNA. Nature 316:76–79

Weir BS, Triggs CM, Starling L, Stowell LI, Walsh KAJ, Buckleton J (1997) Interpreting DNA mixtures. J Forensic Sci 42:213–222

Foundations of Probability

Thomas Augustin, Marco E. G. V. Cattaneo
Ludwig Maximilian University, Munich, Germany

Introduction

Probability theory is that part of mathematics that is concerned with the description and modeling of random phenomena, or in a more general – but not unanimously accepted – sense, of any kind of uncertainty. Probability is assigned to random events, expressing their tendency to occur in a random experiment, or more generally to propositions, characterizing the degree of belief in their truth.

Probability is the fundamental concept underlying most statistical analyses that go beyond a mere description of the observed data. In statistical inference, where conclusions from a random sample have to be drawn about the properties of the underlying population, arguments based on probability allow to cope with the sampling error and therefore control the inference error, which is necessarily present in any generalization from a part to the whole. Statistical modeling aims at separating regularities (structure explainable by a model) from randomness. There, the sampling error and all the variation that is not explained by the chosen optimal model are comprised in an error probability as a residual category.

Different Interpretations and Their Consequences for Statistical Inference

The concept of probability has a very long history (see, e.g., Vallverdú 2010). Originally, the term had a more philosophical meaning, describing the degree of certainty or the persuasive power of an argument. The beginnings of a more mathematical treatment of probability are related to considerations of symmetry in games of chance (see, e.g., Hald 2003). The scope of the theory was extended by Bernoulli (1713), who applied similar symmetry considerations in the study of epistemic probability in civil, moral, and economic problems. In this connection he proved his "law of large numbers," which can be seen as the first theorem of mathematical statistics, and as a cornerstone of the *frequentist* interpretation of probability, which understands the probability of an event as the limit of its relative frequency in an infinite sequence of independent repetitions of a random experiment. Typically, the frequentist (or aleatoric) point of view is *objectivist* in the sense that it relates probability to random phenomena only and perceives probability as a property of the random experiment (e.g., rolling a dice) under consideration.

In contrast, the second of the two most common interpretations (see, e.g., Peterson (2010), for more details), the *subjective*, personalistic, or epistemic viewpoint, perceives probability as a property of the subject confronted with uncertainty. Consequently, here probability can be assigned to anything the very subject is uncertain about, and the question of whether or not there is an underlying random process vanishes. For the interpretation, in the tradition of Savage (1954) a fictive scenario is used where preferences between actions are described. In particular, the probability of an event is understood as the price at which the subject is indifferent between buying and selling a security paying 1 when the event occurs (and 0 otherwise).

The interpretation of probability predetermines to a considerable extent the choice of the statistical inference methods to learn the unknown parameters ϑ of a statistical model from the data. The frequentist perceives ϑ as an unknown but fixed quantity and seeks methods that are optimal under fictive infinite repetitions of the statistical experiment, while for the subjectivist it is straightforward to express his or her uncertainty about ϑ by a (*prior*) probability distribution, which is then, in the light of new data, updated by the so-called Bayes' rule to obtain the (*posterior*) probability distribution describing all her/his knowledge about ϑ (*Bayesian inference*).

Kolmorgorov's Axioms

While the interpretation of probability is quite important for statistical applications, the mathematical theory of probability can be developed almost independently of the interpretation of probability. The foundations of the modern theory of probability were laid by Kolmogorov (1933) in measure theory: Probability is axiomatized as a normalized measure.

More specifically (see, e.g., Merkle (2010) and Rudas (2010) for more details), let Ω be the set of elementary events under consideration (Ω is usually called *sample space*). The events of interest are described as sets of elementary events: it is assumed that they build a σ-algebra $\mathcal{A}$ of subsets of Ω (i.e., $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is nonempty and closed under complementation and countable union). A probability measure on $(\Omega, \mathcal{A})$ is a function $P : \mathcal{A} \to [0,1]$ such that $P(\Omega) = 1$ and

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n) \tag{1}$$

for all sequences of pairwise disjoint events $E_1, E_2, \ldots \in \mathcal{A}$. When Ω is uncountable, a Borel σ-algebra is usually selected as the set $\mathcal{A}$ of events of interest, because the natural choice $\mathcal{A} = \mathcal{P}(\Omega)$ would place too strong limitations

on the probability measure P, at least under the axiom of choice (see, e.g., Solovay (1970)).

Kolmogorov supplemented his axioms by two further basic definitions: the definition of *independence* of events and the definition of *conditional probability* $P(A|B)$ (i.e., the probability of event A given an event B).

From the axioms, fundamental theorems with a strong impact on statistics have been derived on the behavior of independent repetitions of a random experiment (see, e.g., Billingsley (1995) and Schervish (1995) for more details). They include different ►*laws of large numbers* (see above), the *central limit theorem* (see ►Central Limit Theorems), distinguishing the Gaussian distribution as a standard distribution for analyzing large samples, and the *Glivenko–Cantelli theorem* (see ►Glivenko-Cantelli Theorems), formulating convergence of the so-called empirical distribution function to its theoretical counterpart, which means, loosely speaking, that the true probability distribution can be rediscovered in a large sample and thus can be learned from data.

Current Discussion and Challenges

In statistical methodology, for a long time Kolmogorov's measure-theoretic axiomatization of probability theory remained almost undisputed: only countable additivity (1) was criticized by some proponents of the subjective interpretation of probability, such as De Finetti (1974–1975). If countable additivity is replaced by the weaker assumption of finite additivity (i.e., $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ for all pairs of disjoint events $E_1, E_2 \in \mathcal{A}$), then it is always possible to assign a probability to any set of elementary events (i.e., the natural choice $\mathcal{A} = \mathcal{P}(\Omega)$ does not pose problems anymore). However, without countable additivity many mathematical results of measure theory are not valid anymore.

In recent years, the traditional concept of probability has been questioned in a more fundamental way, especially from the subjectivist point of view. On the basis of severe problems encountered when trying to model uncertain expert knowledge in artificial intelligence, the role of probability as the exclusive methodology for handling uncertainty has been rejected (see, e.g., the introduction of Klir and Wierman (1999)). It is argued that traditional probability is only a one-dimensional, too reductionistic view on the multidimensional phenomenon of uncertainty. Similar conclusions (see, e.g., Hsu et al. (2005)) have been drawn in economic decision theory following Ellsberg's seminal experiments (Ellsberg 1961), where the extent of ambiguity (or non-stochastic uncertainty) has been distinguished as a constitutive component of decision making.

Such insights have been the driving force for the development of the theory of *imprecise probability* (see, e.g., Coolen et al. (2010) for a brief survey), comprising approaches that formalize the probability of an event A as an interval $[\underline{P}(A), \overline{P}(A)]$, with the difference between $\overline{P}(A)$ and $\underline{P}(A)$ expressing the extent of ambiguity. Here $\underline{P}$ and $\overline{P}$ are non-additive set-functions, often called *lower* and *upper probabilities*. In particular, Walley (1991) has partially extended De Finetti's framework (De Finetti 1974–1975) to a behavioral theory of imprecise probability, based on an interpretation of probability as possibly differing buying and selling prices, while Weichselberger (2001) has developed a theory of *interval-probability* by generalizing Kolmogorov's axioms.

About the Authors

Dr Thomas Augustin is Professor of Statistics at the Ludwig Maximilian University (LMU), Munich. He is Head of the group "Methodological Foundations of Statistics and their Applications." Dr Marco Cattaneo is Assistant Professor at the Ludwig Maximilian University (LMU), Munich.

Cross References

►Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Measure Theory in Probability
►Philosophical Foundations of Statistics
►Philosophy of Probability
►Probability Theory: An Outline
►Probability, History of
►Statistics, History of

References and Further Reading

Bernoulli J (1713) Ars conjectandi. Thurneysen Brothers, Basel
Billingsley P (1995) Probability and measure, 3rd edn. Wiley, New York
Coolen FPA, Troffaes M, Augustin T (2010) Imprecise probability. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin
De Finetti B (1974–1975) Theory of probability. Wiley, New York
Ellsberg D (1961) Risk, ambiguity, and the Savage axioms. Quart J Econ 75:643–669
Hald A (2003) A history of probability and statistics and their applications before 1750. Wiley, New York
Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF (2005) Neural systems responding to degrees of uncertainty in human decision-making. Science 310:1680–1683
Klir GJ, Wierman MJ (1999) Uncertainty-based information. Physica, Heidelberg
Kolmogorov A (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin
Merkle M (2010) Measure theory in probability. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin

Peterson M (2010) Philosophy of probability. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin
Rudas T (2010) Probability theory: an outline. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin
Savage LJ (1954) The foundations of statistics. Wiley, New York
Schervish MJ (1995) Theory of statistics. Springer, Berlin
Solovay RM (1970) A model of set-theory in which every set of reals is Lebesgue measurable. Ann Math (2nd series) 92:1–56
Vallverdú J (2010) History of probability. In: Lovric M (ed) International encyclopedia of statistical sciences. Springer, Berlin
Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman & Hall, London
Weichselberger K (2001) Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Physica, Heidelberg

Frailty Model

Paul Janssen[1], Luc Duchateau[2]
[1]Professor, President of the Belgian Statistical Society (2008–2010), Vice-rector of research at UHasselt (2008–2012)
Hasselt University, Diepenbeek, Belgium
[2]Professor and Head, President of the Quetelet Society (Belgian branch of IBS) (2010–2012)
Ghent University, Ghent, Belgium

►Survival data are often clustered; it follows that the independence assumption between event times does not hold. Such survival data occur, for instance, in cancer clinical trials, where patients share the same hospital environment. The shared frailty model can take such clustering in the data into account and provides information on the within cluster dependence. In such a model, the frailty is a measure for the relative risk shared by all observations in the same cluster. The model, a conditional hazard model, is given by

$$h_{ij}(t) = h_0(t)u_i \exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right) = h_0(t)\exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta} + w_i\right)$$

where $h_{ij}(t)$ is the conditional (on u_i or w_i) hazard function for the jth observation ($j=1,\ldots,n_i$) in the ith cluster ($i=1,\ldots,s$): $h_0(t)$ is the baseline hazard, $\boldsymbol{\beta}$ is the fixed effects vector of dimension p, $\boldsymbol{x}_{ij}$ is the vector of covariates and w_i (u_i) is the random effect (frailty) for the ith cluster. The w_i's (u_i's) are the actual values of a sample from a density $f_W(.)$ ($f_U(.)$). Clustered survival data will be denoted by the observed (event or censoring) times $\boldsymbol{y} = (y_{11},\ldots,y_{sn_s})^t$ and the censoring indicators $(\delta_{11},\ldots,\delta_{sn_s})^t$. Textbooks references dealing with shared frailty models include Hougaard (2000) and Duchateau and Janssen (2008).

The one-parameter gamma density function $f_U(u) = \frac{u^{1/\theta-1}\exp(-u/\theta)}{\theta^{1/\theta}\Gamma(1/\theta)}$ (with mean one and variance θ) is often used as frailty density as it simplifies model fitting, especially if a parametric baseline hazard (parameterized by $\boldsymbol{\xi}$) is assumed. The marginal likelihood for the ith cluster of the gamma frailty model can easily be obtained by first writing the conditional likelihood for the ith cluster and by then integrating out the gamma distributed frailty. With $\boldsymbol{\zeta} = (\boldsymbol{\xi}, \theta, \boldsymbol{\beta})$, we have

$$L_{marg,i}(\boldsymbol{\zeta}) = \int_0^\infty \prod_{j=1}^{n_i} \left(h_0(y_{ij})u_i \exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)\right)^{\delta_{ij}} \exp\left(-H_0(y_{ij})u_i \exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)\right) \times \frac{u_i^{1/\theta-1}}{\theta^{1/\theta}\Gamma(1/\theta)} \exp\left(-u_i/\theta\right) du_i$$

There exists a closed form for this expression. Taking the logarithm and summing over the s clusters we obtain (Klein 1992; Duchateau and Janssen 2008, Chap. 2)

$$l_{marg}(\boldsymbol{\zeta}) = \sum_{i=1}^{s}\left[d_i\log\theta - \log\Gamma(1/\theta) + \log\Gamma(1/\theta + d_i) - (1/\theta + d_i)\log\left(1 + \theta\sum_{j=1}^{n_i} H_{ij,c}(y_{ij})\right) + \sum_{j=1}^{n_i}\delta_{ij}\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta} + \log h_0(y_{ij})\right)\right] \quad (1)$$

where $H_{ij,c}(y_{ij}) = H_0(y_{ij})\exp\left(\boldsymbol{x}_{ij}^t\boldsymbol{\beta}\right)$ and $d_i = \sum_{j=1}^{n_i}\delta_{ij}$, the number of events in the ith cluster. The marginal loglikelihood does no longer contain the frailties and can therefore be maximized to obtain parameters estimates $\hat{\boldsymbol{\zeta}}$. The asymptotic variance-covariance matrix can also be obtained using the marginal loglikelihood expression. The preferred model in survival analysis is a (conditional) hazards model with unspecified baseline hazard (a semiparametric model, a Cox model). Leaving $h_0(.)$ and $H_0(.)$ in (1) unspecified we obtain a semiparametric gamma frailty model. For such model direct maximization of the marginal likelihood is not possible. Both the EM-algorithm (Klein 1992) and penalized likelihood maximization (Therneau et al. 2003) have been proposed to fit such models; both approaches use the fact that closed form expressions can be obtained for the expected values of the frailties.

An alternative representation of the marginal likelihood (1) for the parametric gamma frailty model is based on the Laplace transform of the gamma frailty density

$\mathcal{L}(s) = \mathrm{E}(\exp(-Us)) = (1+\theta s)^{-1/\theta}$. With $\boldsymbol{t}_{n_i} = (t_1, \ldots, t_{n_i})$ and $H_{i,c}(\boldsymbol{t}_{n_i}) = \sum_{j=1}^{n_i} H_{ij,c}(t_i)$, the joint survival function for the ith cluster is given by

$$S_{i,f}(\boldsymbol{t}_{n_i}) = \int_0^\infty \exp(-u_i H_{i,c}(\boldsymbol{t}_{n_i})) f_{U_i}(u_i) du_i$$
$$= \mathcal{L}(H_{i,c}(\boldsymbol{t}_{n_i})) = (1+\theta H_{i,c}(\boldsymbol{t}_{n_i}))^{-1/\theta}$$

The likelihood contribution of the ith cluster, with $\boldsymbol{y}_{n_i} = (y_{i1}, \ldots, y_{in_i})$ and the first l observations uncensored, is then

$$(-1)^l \frac{\partial^l}{\partial t_1 \ldots \partial t_l} S_{i,f}(\boldsymbol{y}_{n_i}) = (-1)^l \mathcal{L}^{(l)}(H_{i,c}(\boldsymbol{y}_{n_i}))$$
$$\prod_{j=1}^{l} h_0(y_{ij}) \exp(\boldsymbol{x}_{ij}^t \boldsymbol{\beta}) \qquad (2)$$

For the gamma frailty model the explicit form of (2) is

$$\prod_{j=1}^{n_i} h_{\boldsymbol{x}_{ij},c}^{\delta_{ij}}(y_{ij}) (1+\theta H_{i,c}(\boldsymbol{y}_{n_i}))^{-1/\theta - d_i} \prod_{l=0}^{d_i-1} (1+l\theta)$$

with $\prod_{l=0}^{d_i-1}(1+l\theta) = 1$ for $d_i = 0$.

For frailty densities different from the gamma frailty density, for which the Laplace transform exists, expression (2) is the key to obtain the appropriate marginal loglikelihood expression. Frequently used frailty densities, such as the inverse Gaussian and the positive stable densities (Hougaard 1986a), have indeed simple Laplace transforms. More complex two-parameter frailty densities are the power variance function densities (Hougaard 1986b) and the compound Poisson densities (Aalen 1992). Although the lognormal density is also used as frailty density, it does not have a simple Laplace transform; its use mainly stems from mixed models ideas (McGilchrist and Aisbett 1991), and different techniques, such as ▶numerical integration, have to be used to fit this model (Bellamy et al. 2004).

The choice of the frailty density determines the type of dependence between the observations within a cluster. A global dependence measure is Kendall's τ (Kendall 1938). For two randomly chosen clusters i and k of size two with event times (T_{i1}, T_{i2}) and (T_{k1}, T_{k2}) and no covariates, τ is defined as $\mathrm{E}[\mathrm{sign}((T_{i1}-T_{k1})(T_{i2}-T_{k2}))]$ where $\mathrm{sign}(x) = -1, 0, 1$ for $x < 0, x = 0, x > 0$. Kendall's τ can be expressed as a function of the Laplace transform. Global dependence measures do not allow us to investigate how dependence changes over time. An important local dependence measure is the cross ratio function (Clayton 1978). An interesting feature of this function is its relation with a local version of Kendall's τ (see Duchateau and Janssen 2008, Chap. 4). The positive stable distribution and the ▶gamma distribution characterize early and late dependence respectively, with the ▶inverse Gaussian distribution taking a position in between the two.

So far we discussed the shared frailty model, which is the most simple model to handle within cluster dependence. The shared frailty model can be extended in different ways. First, a frailty term can be assigned to each subject, resulting in a univariate frailty model which can be used to model overdispersion (Aalen 1994). Another extension is the correlated frailty model in which the subjects in a cluster do not share the same frailty term although their respective frailties are correlated (Yashin and Iachine 1995). Finally the model can be extended to multifrailty and multilevel frailty models. In a multifrailty model, two different frailties occur in one and the same cluster. A good example is the study of the heterogeneity of a prognostic index over hospitals in cancer clinical trials, with each hospital (cluster) containing a frailty term for the hospital effect and a frailty term for the prognostic index effect (Legrand et al. 2007). Multilevel frailty models have two or more nesting levels, with a set of smaller clusters contained in a large cluster. Fitting such models is discussed in Ripatti and Palmgren (2000) and Rondeau et al. (2006).

Cross References

▶Demographic Analysis: A Stochastic Approach
▶Hazard Ratio Estimator
▶Hazard Regression Models
▶Modeling Survival Data
▶Survival Data

References and Further Reading

Aalen OO (1992) Modelling heterogeneity in survival analysis by the compound Poisson distribution. Ann Appl Probab 2:951–972

Aalen OO (1994) Effects of frailty in survival analysis. Stat Methods Med Res 3:227–243

Bellamy SL, Li Y, Ryan LM, Lipsitz S, Canner MJ, Wright R (2004) Analysis of clustered and interval-censored data from a community-based study in asthma. Stat Med 23:3607–3621

Clayton DG (1978) A model for association in bivariate life tables and its application in epidemiological studies of family tendency in chronic disease incidence. Biometrika 65:141–151

Duchateau L, Janssen P (2008) The frailty model. Springer, New York

Hougaard P (1986a) A class of multivariate failure time distributions. Biometrika 73:671–678

Hougaard P (1986b) Survival models for heterogeneous populations derived from stable distributions. Biometrika 73:387–396

Hougaard P (2000) Analysis of multivariate survival data. Springer, New York

Kendall MG (1938) A new measure of rank correlation. Biometrika 30:81–93

Klein JP (1992) Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics 48: 795–806

Legrand C, Duchateau L, Janssen P, Ducrocq V, Sylvester R (2007) Validation of prognostic indices using the frailty model. Lifetime Data Anal 15:59–78
McGilchrist CA, Aisbett CW (1991) Regression with frailty in survival analysis. Biometrics 47:461–466
Ripatti S, Palmgren J (2000) Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 56:1016–1022
Rondeau V, Filleul L, Joly P (2006) Nested frailty models using maximum penalized likelihood estimation. Stat Med 25:4036–4052
Therneau TM, Grambsch PM, Pankratz VS (2003) Penalized survival models and frailty. J Comput Graph Stat 12:156–175
Yashin AI, Iachine IA (1995) Genetic analysis of durations: correlated frailty model applied to survival of Danish twins. Genet Epidemiol 12:529–538

Fraud in Statistics

Vassiliy Simchera
Director of Rosstat's Statistical Research Institute
Moscow, Russia

Fraud is an intentional distortion of the truth, whether it is a deliberate omission or false elimination, or an exaggeration or fabrication.

The aim of one who commits fraud is always self-benefit and self-interest. The main reasons for fraud are immorality, impunity, and anarchy, and the methods are deception and betrayal. From these it gives rise to the gravest of crimes – violation of law, murder, mutinies and wars. The tools to overcome fraud are law and order, auditing and control, morals, science, prosecution, and adequate punishment.

Fraud is a man-made phenomenon. The substance of fraud is unknown in the natural world. A lack of knowledge or limited knowledge, unpremeditated actions as well as unobserved phenomena (including deliberate, but legal and justified by jury's verdict, but still obviously criminal actions and perjuries) on the modern level of social mentality do not belong to the substance of fraud, and they are definitely not a subject for a scientific research.

Science, as opposed to jurisprudence, is much more liberal (despite some exceptions for genius of Galileo Galilei, Giordano Bruno, Jan Hus and Nicolaus Copernicus).

The most efficient tool for not only revealing but also overcoming fraud in economy (and further in socio-economic activity) is statistics. Its accurate methods of observation and auditing, powerful databases and knowledge bases, advanced software, and technical provision, as well as the intellectual culture verified by hundreds of years of qualitative data collection and data processing, allow the guarantee of controlled completeness, credibility, and accessibility for a wide range of people.

Being the world's most powerful information system with regulated branches in center and local areas controlled by hundred of quality criteria including those provided by IMF, the modern statistics is by its nature, as any other meter device is free from necessity to lie but at the same time it is surrounded by various kinds of lies and in turn reflects them, and as any other domain of empirical knowledge cannot be free of it.

Fraud in statistics is distortion of data, resulting from two different types of causes: 1) distortion of random errors, caused by poor observation and calculation, the characteristics of which are analyzed in another chapter in this text 2) deliberate (premeditated) distortion of data, resulting from different kinds of systematic causes and giving rise to effects beyond the statistics domain; these cannot be eliminated by methods or techniques.

The main sources of the data distortions (or in simple words – improper data reflections) are unknown, unobservable, and immeasurable phenomena. Such phenomena are not and cannot be discussed due to the objective reasons by observed phenomena; they are actually published and reflected in an incomplete and distorted form and thus they rather characterize themselves but do not reflect the real situation.

The most widespread sources of distortion in modern statistics are:

- evasion from participation in the preparation and submission of the obligatory current statistical reports;
- failure of respondents to answer ▶questionnaire for periodical and random statistical samplings;
- use of obsolete registers of individuals as well as legal entities, omissions, including deliberate omissions of the observed units and the reports units;
- underestimation (or overestimation) of the statistical data registration and reporting.

The particular type of data distortion in modern statistics, statistical estimates, connected with substitution of concepts or estimates obtained with use of inadequate techniques and algorithms that cannot be verified by the existing criteria of their credibility or with the application of other control methods, which are suitable for solving similar class of problems.

The biggest domains of fraud in statistics today are activities that cannot be prohibited and there is little means to prevent these activities. They are as follows:

- all types of illegal activities, including terrorism, counterfeiting, money laundering, corruption, smuggling, drug dealing, arms trafficking, illegal mining of rare metals, trafficking of toxic agents illegal organ transplants, and child abduction;
- illegal activity of individuals and legal entities;
- illegal business of unregistered organizations, institutions, and individuals;
- production and rendering services for self-consumption by households and individuals;
- unrecorded and omitted by statistical observations types of activities, statistical errors, rounding, errors and discrepancies, underestimated or overestimated estimates;
- second-hand activity, tolling, venture enterprises, intellectual activity; intermediate consumption, subsidies for production; offshore activity;
- transactions charges, fees, tips, VAT return, barter, payment in goods for labour, single payments, taxes, losses, including anthropogenic impacts;
- doctoring, imputed value, royalty, goodwill, repeat count of raw materials, commodities, services and capitals;
- other reasons, their identification is imposed and acceptable within the limits of standards and methods of effective statistical reporting and accounts.

The demonstrative example of fraud is fictitious estimates of capitalization of the world markets which, against the background of their real assets estimates (2008) would not exceed $70 trillion USD, and today account for over $700 trillion of USD.

The phantom of fraud in the modern world is also represented by estimates of banks assets. According to these estimates by US Statistics on Banking, 2007 (table 1136) which is considered as the most reliable one, the aggregate assets of all 730,000 of American banks in 2007 were estimated for $14.0 trillion USD (the precise sum is $13,792.5 billion of USD), whilst according to public information the allied assets of JP Morgan Bank solely at the same year were estimated for $97.5 trillion USD, Goldman and Sachs – $50 trillion USD, and HSBC – $108 trillion USD, which exceeded their accounted real equity capital by 30–40 times or more.

Another example of fraud is tax evasion, in particular VAT, the size of which reaches one-third of its total volume in the world, including over $20 billion USD per year in England or $50 billion USD in the United States.

However, there are no ideal measurements or absolutely precise estimations in science and life. Even those which seem to be absolutely accurate values obtained from the variables such as lengths, speed, weight, and temperature (degrees), are just conventional but not the absolute truth itself.

On the other hand, not all inaccurate values (estimates) are distorted ones and hence not all distorted values are false. In accordance with existing criteria in statistics, the inaccurate estimates are such and only distorted estimates that deteriorate the true core of measured phenomena and turn it into its opposite, that is to say, a lie. Inaccurate and reasonably distorted estimates, which by the way prevail in modern statistics (actually they prevailed in the past, too), are called approximated and they are widely used with reserve of some errors as acceptable asymptotic or approximations estimates.

About the Author

For biography *see* the entry ►Actuarial Methods.

Cross References

►Banking, Statistics in
►Misuse of Statistics
►Pyramid Schemes
►Statistical Fallacies
►Statistical Fallacies: Misconceptions, and Myths

References and Further Reading

Keen M, Smith S (2007) VAT fraud and evasion: what do we know, and what can be done. IMF Working Paper. 07/31

Mauro P (2002) The persistence of corruption and slow economic growth. IMF Working Paper /02/213

OECD (2002) Measuring the non-observed economy: a handbook. OECD Publications, Paris

Simchera VM (2003) Statistical information and economic disinformation. Federalism Magazine, Russia, 3:91–116

Simchera VM (2006) Moral economy. TENS Publishing House, Russia

Yehoue EB, Ruhashyankiko JF (2006) Corruption and technology-induced private sector development. IMFWorking Paper /06/198

Frequentist Hypothesis Testing: A Defense

Shlomo Sawilowsky
Professor and Assistant Dean
Wayne State University, Detroit, MI, USA

John Graunt, William Petty, René Déscartes, Blaise Pascal, Pierre Fermat, James Gregory, Christiaan Huygens,

Isaac Newton, Gottfried Leibniz, Jakob Bernoulli, Johann Bernoulli, Abraham DeMoivre, Daniel Bernoulli, Leonhard Euler, Joseph Lagrange, Pierre Laplace, Siméon Poisson, Jean Fourier, Friedrich Bessel, Carl Jacobi, Carl Gauss, Augustin Cauchy, Gustav Dirichlet, Georg Riemann, Michel Chasles, Augustus DeMorgan, Lambert Quetelet, Joseph Liouville, Pafnuty Chebyshev, Charles Hermite, and Andrei Markov. It was the work of these men, among others, that led to the development of the grand theorems, the mathematical inerrancies.

True, they were initially prompted by real world problems, such as winning games of chance or determining actuarial odds; and esoteric problems, such as proving the existence of a Divine plan by confirming a slightly greater proportion of male births to ensure the survival of the species. However, the grand theorems ascended to their elevated place in history because they are elegant, not because they were particularly useful in solving the problems for which they were created.

Moreover, they dictated the types of problems that are considered worthy, relegating those not subsumed under the cleverness of what mankind can solve as being intractable and designated as an eternal mystery of the universe. Their importance was buttressed by their utility for the few problems that they could solve, not for the problems that needed to be solved. Nunnally (1978) wrote mathematics "is purely an abstract enterprise that need have nothing to do with the real world… Thus the statement *iggle wug drang flous* could be a legitimate mathematical statement in a set of rules stating that when any *iggle* is *wugged* it *drang* a *flous*…Of course … [this] might not be of any practical use" (p. 9–10).

Woodward (1906) observed "since the beginning of the eighteenth century almost every mathematician of note has been a contributor to or an expositor of the theory of probability" (p. 8). But the focus on probability eventually moved away from populations and the grand theorems, and settled on just very large samples, such as, e.g., the work of Charles Darwin and Francis Galton.

Darwin collected his data between December 26, 1831 and February 27, 1832 while on the Cherokee class ten gun brig-sloop H. M. S. Beagle, sailing under the command of Captain Robert Fitzroy. Most of Darwin's data were obtained in St. Jago (Santiago) in the Cape Verde Islands from January 16 – February 7. Galton (1885) collected 17 discreet data points on 9,337 people. They were measured in a cubicle 6 feet wide and 36 feet long with the assistance of Serjeant Williams, Mr. Gammage the optician, and a doorkeeper who made himself useful. The data were obtained in the anthropometric laboratory at the International Health Exhibition and subsequently deposited at the South Kensington Museum.

Darwin's and Galton's lack of mathematical training limited their ability to quantify and analyze their trophies, but that limitation was resolved with the brilliance of Karl (née Carl) Pearson. With their data in hand, and the more immediate problem of huge data sets from the biologist/zoologist Walter Weldon, Pearson set to work. By 1900, he provided the rigor that had eluded his colleagues with the discovery of both r and χ^2, and the world was at peace. Well, at least scholars, the intelligencia, and their paparazzi were comforted.

K. Pearson (1978) assuredly knew the limitations of the grand theorems. After all, he quipped

> "As I understand De Moivre the 'Original Design' is the mean occurrence on an indefinite number of trials…The Deity fixed the 'means' and 'chance' provided the fluctuations…There is much value in the idea of the ultimate laws of the universe being statistical laws… [but] it is not an exactly dignified conception of the Deity to suppose him occupied solely with first moments and neglecting second and higher moments!" (p. 160)

But, alas and alack, as the first champion of statistics, K. Pearson was the inheritor of the grand theorems. As a co-founding editor of *Biometrika* he strove to stay above controversy by minimizing, if not ignoring, ordinary problems. And indeed there are those who still pine for the days of yore with its grand theorems, as Tukey (1954) nostalgically noted "Once upon a time the calculation of the first four moments was an honorable art in statistics" (p. 717).

But the ordinary person readily intuited that real world problems are not asymptotic in nature. William Gosset's Monte Carlo study published in 1908 with numbers written on pieces of poster board was conducted because he wasn't sure mathematicians could help him with real, small samples problems.

How could the recipe of his employer, Arthur Guinness, be improved? How many barrels of malt or hops are needed to approximate a population, or at least a large number? 2? 3? Are 4 barrels close to infinity? This chemist (whose sole credential was his undergraduate dual major in chemistry and mathematics from New College, Oxford) sounded all the great mathematical minds of his day, who assured him that he could rely on the grand theorems, and that he need not trouble himself with matters above his pay grade. Daydreams, it seems he was told, were more profitable than the time spent fretting on how large is large.

Gosset ("Student") neither wrote the t formula in the form that it appears in undergraduate textbooks today, nor

did he create the experimental design in which it would be applied. Ronald Fisher provided both the design and the null hypothesis that brought Gosset's Monte Carlo experiments and intuitive mathematics to fruition. Then, Pearson, Gosset, and Fisher became a quintet with the addition of Jerzy Neyman and Egon Pearson, who cemented the frequentist approach to hypothesis testing with the creation of an alternative hypothesis. Not satisfied, Neyman later re-expressed frequentist hypothesis testing into confidence intervals based on the same theory of probability.

Doubts were immediately raised, such as Bowley (1934), who asked and answered, "I am not at all sure that the 'confidence' is not a 'confidence trick'… Does it really take us any further?... I think it does not" (p. 609). Many scholars have adopted the shortcut to notoriety by rushing to follow in Bowley's footsteps, proclaiming the sky is falling on hypothesis testing.

But K. Pearson's development of the χ^2 test is surely listed among the greatest practical achievements of three millennia of mathematical inquiry. He captured the ability, regardless of the direct object, to quantify the difference between human observation and expectation. Remarkable! Was it wrong? Of course. Fisher had to modify the degrees of freedom. Again, remarkable! Was it still wrong? Of course. Frank Yates had to modify the method for small values of expectation. Once again, remarkable! Was it nevertheless wrong? Of course. The correction was found to sap statistical power. Where does the χ^2 test stand today? Statistical software can produce exact p values regardless of how small the expectation per cell. Remarkable!

Has society, therefore, improved with advent of the evolution of the χ^2 test? Most assuredly not:

- We live in a χ^2 society due to political correctness that dictates equality of outcome instead of equality of opportunity. The test of independence version of this statistic is accepted *sans voire dire* by many legal systems as the single most important arbiter of truth, justice, and salvation. It has been asserted that any statistical difference between (often even nonrandomly selected) samples of ethnicity, gender, or other demographic as compared with (often even inaccurate, incomplete, and outdated) census data is *primae faciea* evidence of institutional racism, sexism, or other ism. A plaintiff allegation that is supportable by a significant χ^2 is often accepted by the court (judges and juries) *praesumptio iuris et de iure*. (Sawilowsky 2010).

But is this really the fault of the χ^2 test? Any device can be lethal in the hands of a lunatic, as Mosteller (1968) warned, "I fear that the first act of most social scientists upon seeing a contingency table is to compute a chi-square for it" (p. 1).

What discipline has not followed an evolutionary path? Has agriculture, archeology, architecture, anthropology, biology, botany, chemistry, computer science, education, engineering, genetics, medicine, nursing, pharmacology, physics, psychology, sociology, and zoology always been as they exist today? Do we blame statistics for its ignoble development more so because the content disciplines were dependent on it?

There have been antagonists of Fisher–Neyman/Pearson hypothesis testing for three quarters of a century since Bowley. And it is understandable with the following analogy:

I had a summer job in 1972, working in Florida for a major manufacturer of fiberglass yachts. The hulls of the larger boats were made by laminating two 1/2 hull pieces together. The exterior paint was sprayed inside the two molds and set to dry. Next, fiberglass chop mixed with resin and methyl ethyl ketone peroxide was sprayed into the mold, laminators worked out the air bubbles, and it was set to harden.

The two 1/2 hull shells were then aligned, and held in place with many C clamps with the aid of a powerful air compressor. This was necessary because over time the molds changed in shape and they no longer matched. A small, temporary seam was laminated inside the hull to keep the two parts together. When the clamps were released, one could almost see the two 1/2 hulls trembling, working against the thin fiberglass seam to separate and go their separate ways.

My job was to be lowered inside the hull, and lay down a successively wider series of fiberglass mats and resign/catalyst, to strengthen the seam. On one boat I had laminated perhaps five or six of the required ten mats when it was quitting time. The crew chief told me I could continue the next day where I had left off.

To my chagrin, when I arrive early the next day I discovered that the night shift personnel had taken my hull down the production line, and the boat by now had floors, carpet, sink, and other amenities already installed, obviating the ability to bond the final fiberglass mats to strengthen the hull's seam. I protested to my crew chief, who nonchalantly replied not to worry myself about such things. "After all," he said, "the naval architects who designed the boat allowed considerable tolerance that should handle situations such as this." I made that my last day on the job at that company, and since then I've often wondered how that yacht fared in the middle of the Gulf of Mexico.

So too, the juxtaposition of Fisher's null with Neyman and E. Pearson's alternative leads to trembling, each part of the statistical hypothesis seemingly working against each other to go their separate ways. But this was not the end of the development of the frequentist theory. It surpassed E. Pearson (1962), who admitted "through the lack of close contact with my partner during the last 20 years, it would be a little difficult to say where precisely the Neyman and Pearson theory stands today" (p. 53). The same sentiment was also expressed by those who followed the age of the pioneers, such as Savage (1962) who echoed, "What I, and many other statisticians, call the Neyman-Pearson view may, for all I know, never have been held by Professor Neyman or by Professor Pearson" (p. 62). Wilks (1948) concluded that by now the "modern statistical method is a science in itself" (p. 1).

In truth, many of the foibles in hypothesis testing, since being admitted to the country club of mature disciplines, are traceable back to the statistician, not to the statistics. Fallacies, misconceptions, and myths abound. Which of the disciplines listed above are immune to this, and why is there an expectation that statistics should fare any better?

Yes, even under the best of circumstances there are those who have no use for hypothesis testing. Ernst Rutherford (cited in N T J Bailey 1967) said, "If your experiment needs statistics, you ought to have done a better experiment" (p. 23). But, Albert Einstein (cited in Shankland 1973) countered, "I thank you very much for sending me your careful study about the [Dayton] Miller experiments. Those experiments, conducted with so much care, merit, of course, *a very careful statistical investigation*," (p. 2283, italics added for emphasis).

Much of the criticism against hypothesis testing would presumably vanish if workers heeded the advice of Finney (1953), who advised "when you are experienced enough to make your own statistical analyses, be sure you choose the right technique and not merely any one that you can remember!" (p. 174). The sciences, physical and social, should be placated with McNemar's (1949) advice that "the student should be warned that he cannot expect miracles to be wrought by the use of statistical tools" (p. 3).

Proper selection of statistical tests based on their small samples properties, along with an understanding of their respective null and alternative hypotheses, research design, random sampling, nominal α, Type I and II errors, statistical power, and effect size would eliminate attacks against hypothesis testing from all save perhaps those who, as Bross (1969) characterized it, base their science on "a Bayesian *t*-test using an informationless prior" (p. 52). Has the world benefitted from frequentist hypothesis testing?

> The question is silly. No reputable quantitative physical, behavioral, or social scientist would overlook the breadth and depth of scholarly knowledge and its impact on society that has accrued from over a century of hypothesis testing. The definitive evidence: William Sealy Gosset created the *t* test to make better beer. (Sawilowsky 2003, p. 469)

About the Author

Shlomo S. Sawilowsky is Professor and Assistant Dean in the College of Education, and Wayne State University Distinguished Faculty Fellow. He is the author of *Statistics Through Monte Carlo Simulation With Fortran* (2003) and *Real Data Analysis* (2008), and over 100 peer-reviewed articles on applied data analysis. He is the founding editor of the *Journal of Modern Applied Statistical Methods* (http://tbf.coe.wayne.edu/jmasm). He has served as major professor on 52 doctoral dissertations, Co-advisor on 18 dissertations, 2*nd* advisor on 37 doctoral dissertations, Cognate advisor on 2 doctoral dissertations, and advisor on 23 Master's theses in applied data analysis. Approximately 1/2 of his graduates are female and 1/4 are African American. Professor Sawilowsky has won many teaching and research awards. He was the recipient of the 1998 Wayne State University Outstanding Graduate Mentor Award, and the College of Education's Excellence in Teaching Award. "Professor Sawilowsky's exceptional record as an academician is reflected in the excellence with which he mentors graduate students" (AMSTAT News, October 1998). Professor Sawilowsky was the 2008 President of the American Educational Research Association/SIG Educational Statisticians.

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning
►Confidence Interval
►Effect Size
►Full Bayesian Significant Test (FBST)
►Null-Hypothesis Significance Testing: Misconceptions
►Presentation of Statistical Testimony
►Psychology, Statistics in
►P-Values
►Role of Statistics
►Significance Testing: An Overview
►Significance Tests, History and Logic of
►Significance Tests: A Critique
►Statistical Evidence
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Inference: An Overview

►Statistical Significance
►Statistics: Controversies in Practice

References and Further Reading

Bailey NTJ (1967) The mathematical approach to biology and medicine. Wiley, New York
Bowley A (1934) Discussion on Dr. Neyman's paper. J Roy Stat Soc 97:607–610
Bross I (1969) Applications of probability: science versus pseudoscience. J Am Stat Assoc 64:51–57
Finney D (1953) An introduction to statistical science in agriculture. Ejnar Munksgaard, Copenhagen
Galton F (1885) On the anthropometric laboratory at the late international health exhibition. Journal of the Anthropological Institute of Grand Britain and Ireland, 14:205–221
McNemar Q (1949) Psychological statistics. Wiley, New York
Mosteller F (1968) Association and estimation in contingency tables. J Am Stat Assoc 63:1–28
Nunnally JC (1978) Psychometric theory, 2nd edn. McGraw-Hill, New York
Pearson ES (1962) The foundations of statistical inference. Methuen, London
Pearson K (ed. Pearson ES) (1978) The history of statistics in the 17th and 18th centuries against the changing background of intellectual, scientific and religious thought: Lectures given at University College London during the academic sessions 1921–1923. Macmillan, New York
Savage LJ (1962) The foundations of statistical inference. Methuen, London
Sawilowsky S (2010) Statistical fallacies, misconceptions, and myths, this encyclopedia
Sawilowsky S (2003) Deconstructing arguments from the case against hypothesis testing. J Mod Appl Stat Meth 2(2):467–474
Shankland R (1973) Michelson's role in the development of relativity. Appl Optics 12(10):2280–2287
Tukey JW (1954) Unsolved problems of experimental statistics. J Am Stat Assoc 49:706–731
Wilks SS (1948) Elementary statistical analysis. Princeton University Press, Princeton
Woodward R (1906) Probability and theory of errors. Wiley, New York

Full Bayesian Significant Test (FBST)

Carlos Alberto de Bragança Pereira
Professor, Head, Instituto de Matemática e Estatística
Universidade de São Paulo, São Paulo, Brazil

Introduction

Significance testing of precise (or sharp) hypotheses is an old and controversial problem: it has been central in statistical inference. Both frequentist and Bayesian schools of inference have presented solutions to this problem, not always prioritizing the consideration of fundamental issues such as the meaning of precise hypotheses or the inferential rationale for testing them. The Full Bayesian Significance Test, FBST, is an alternative solution to the problem, which attempts to ease some of the questions met by frequentist and standard Bayes tests based on Bayes factors. FBST was introduced by Pereira and Stern (1999) and reviewed by Pereira et al. (2008).

The discussion here is restricted to univariate parameter and (sufficient statistic) sample spaces;

$$\Theta \subset \mathcal{R} \text{ and } X \subset \mathcal{R}$$

A sharp hypothesis H is then a statement of the form $H : \theta = \theta_0$ where $\theta_0 \in \Theta$. The posterior probability (density) for θ is obtained after the observation of $x \in X$. While a frequentist looks for the set, C, of sample points at least as inconsistent with θ_0 as x is, a Bayesian could look for the tangential set T of parameter points that are more consistent with x than θ_0 is. This understanding can be interpreted as a partial duality between sampling and Bayesian theories. The evidence in favor of H is for frequentists the usual p-value, while for Bayesian it should be $ev = 1 - \underline{ev}$:

$$pv = Pr\{x \in C|\theta_0\} \text{ and } ev = 1 - \underline{ev} = 1 - Pr\{\theta \in T|x\}.$$

The larger pv and ev, the stronger the evidence favoring H.

In the general case, the posterior distribution is sufficient for ev to be calculated, without any complication due to dimensionality of neither the parameter nor of the sample space. This feature ceases the need for nuisance parameters elimination, a problem that disturbs some statisticians (Basu 1977). If one feels that the goal of measuring consistency between data and a null hypothesis should not involve prior opinion about the parameter, the normalized likelihood, if available, may replace the posterior distribution. The computation of ev needs no asymptotic methods, although numerical optimization and integration may be needed.

The fact that the frequentist and Bayesian measures of evidence, pv and ev, are probability values – therefore defined in a zero to one scale – does not easily help to answer the question "How small is *significant*?". For ►p-values, the NP lemma settles the question by means of subjective arbitration of critical values. For Bayesian assessment of significance through evaluation of ev, decision theory again clears the picture. Madruga et al. (2001) show that there exist loss functions the minimization of which render a test of significance based on ev into a formal Bayes test.

The FBST has successfully solved several relevant problems of statistical inference: see Pereira et al. (2008) for a list of publications.

FBST Definition

Significance FBST was created under the assumption that a significance test of a sharp hypothesis had to be performed. At this point, a formal definition of a sharp hypothesis is presented.

Consider general statistical spaces, where $\Theta \subset \mathcal{R}^m$ is the parameter space and $X \subset \mathcal{R}^k$ is the sample space.

Definition 1 A *sharp* hypothesis H states that θ belongs to a sub-manifold Θ_H of smaller dimension than Θ.

The subset Θ_H has null Lebesgue measure whenever H is sharp. A probability density on the parameter space is an ordering system, notwithstanding having every point probability zero. In the FBST construction, all sets of same nature are treated accordingly in the same way. As a consequence, the sets that define sharp hypotheses keep having nil probabilities. As opposed to changing the nature of H by assigning positive probability to it, the tangential set T of points, having posterior density values higher than any θ in Θ_H, is considered. H is rejected if the posterior probability of T is *large*. The formalization of these ideas is presented below.

Let us consider a standard parametric statistical model; i.e., for an integer m, the parameter is $\theta \in \Theta \subset \mathcal{R}^m$, $g(\bullet)$ a probability prior density over Θ, x is the observation (a scalar or a vector), and $L_x(\bullet)$ is the likelihood generated by data x. Posterior to the observation of x, the sole relevant entity for the evaluation of the Bayesian evidence ev is the posterior probability density for θ given x, denoted by

$$g_x(\theta) = g(\theta|x) \propto g(\theta)L_x(\theta).$$

Of course, one is restricted to the case where the posterior probability distribution over Θ is absolutely continuous; i.e., $g_x(\theta)$ is a density over Θ. For simplicity, H is used for Θ_H in the sequel.

Definition 2 (evidence) Consider a sharp hypothesis $H : \theta \in \Theta_H$ and

$$g^* = \sup_H g_x(\theta) \text{ and } T = \{\theta \in \Theta : g_x(\theta) > g^*\}.$$

The *Bayesian evidence value against* H is defined as the posterior probability of the tangential set, i.e.,

$$\underline{ev} = Pr\{\theta \in T|x\} = \int_T g_x(\theta)d\theta.$$

One must note that the evidence value supporting H, $ev = 1 - \underline{ev}$, is not an evidence against A, the alternative hypothesis (which is not sharp anyway). Equivalently, $\underline{ev}$ is not evidence in favor of A, although it is against H.

Definition 3 (test) The *FBST* (Full Bayesian Significance Test) is the procedure that rejects H whenever $ev = 1 - \underline{ev}$ is small.

The following example illustrates the use of the FBST and two standard tests, McNemar and Jeffreys' Bayes Factor. Irony et al. (2000) discuss this inference problem introduced by McNemar (1955).

Example 1 McNemar vs. FBST Two professors, Ed and Joe, from the Department of Dentistry evaluated the skills of 224 students in dental fillings preparation. Each student was evaluated by both professors. The evaluation result could be approval (A) or disapproval (F). The Department wants to check whether the professors are equally exigent. Table 1 presents the data.

This is a four-fold classification with probabilities p_{11}, p_{12}, p_{21}, and p_{22}. Using standard notation, the hypothesis to be tested is $H : p_{1\cdot} = p_{\cdot 1}$ which is equivalent to $H : p_{12} = p_{21}$ (against $A : p_{12} \neq p_{21}$). In order to have the likelihood function readily available, we will consider a uniform prior, i.e., a *Dirichlet* density with parameter $(1, 1, 1, 1)$.

The McNemar exact significance for this data set is $pv = .064$. Recall that this test is based in a partial likelihood function, a binomial with $p = p_{12}(p_{12} + p_{21})^{-1}$ and $n = 66$. With the normal approximation, the pv become .049 with the partial likelihood used by McNemar, the FBST evidence is $ev = .045$. The value of the Bayes Factor under the same uniform prior is $BF = .953$. If one assigns probability 1/2 to the sharp hypothesis H, its posterior probability attains $\pi = .488$. Hence, the posterior probability π barely differs from 1/2, the probability previously assigned to H, while pv and ev seem to be more conclusive against H. While, in the three dimension full model, $ev = 0.265$ may seem to be a not low value and the test cannot be performed without a criterion. In other

Full Bayesian Significant Test (FBST). Table 1 Results of the evaluation of 224 students

	Joe		
Ed	A	F	Total
A	62	41	103
F	25	96	121
Total	87	137	224

words, a decision is not made until *ev* is compared to a "critical value." The derivation of such a criterion – resulting from the identification of the FBST as a genuine Bayes procedure – is the subject of Madruga et al. (2001).

The strong disagreement among the values of *ev*, *pv*, and *BF* seldom occurs in situations where Θ is a subset of the real line. The speculation is that this is related to the elimination of nuisance parameters: By conditioning in McNemar case and by marginalization in the Bayes Factor case. In higher dimension, elimination of nuisance parameters seems to be problematic, as pointed by Basu (1977).

FBST Theory

From a theoretical perspective, on the other hand, it may be propounded that if the computation of *ev* is to have any inferential meaning, then it ought to proceed to a declaration of significance (or not). To this – in a sense – simultaneously NPW and Fisherian viewpoint can be opposed the identification of ev as an estimator of the indicator function $\phi = I(\theta \in \Theta_H)$. In fact, Madruga et al. (2001) show that there are loss functions the minimization of which makes *ev* a Bayes estimator of ϕ (see Hwang et al. 1992).

Madruga et al. (2001) prove that the FBST procedure is the posterior minimization of an expected loss λ defined as follows:

$$\lambda(\text{Rejection of } H, \theta) = a\{1 - I[\theta \in T]\} \text{ and}$$

$$\lambda(\text{Acceptance of } H, \theta) = b + dI[\theta \in T].$$

Here, a, b and d are positive real numbers. The operational FBST procedure is given by the criterion according to which H is to be rejected if, and only if, the evidence *ev* is smaller than $c = (b+d)/(a+d)$. One should notice that the evidence *ev* is the Bayesian formal test statistic and that positive probability for H is never required. A complete discussion of the above approach can be found in Pereira et al. (2008).

Final Remarks

The following list states several desirable properties attended by *ev*:

1. *ev* is a probability value derived from the posterior distribution on the full parameter space.
2. Both *ev* and FBST possesses versions which are invariant for alternative parameterizations.
3. The need of approximations in the computation of *ev* is restricted to numerical maximization and integration.
4. FBST does not violate the Likelihood Principle.
5. FBST neither requires nuisance parameters elimination nor the assignment of positive prior probabilities to sets of zero Lebesgue measure.
6. FBST is a formal Bayes test and therefore has critical values obtained from considered loss functions.
7. *ev* is a possibilistic support for sharp hypotheses, complying with the Onus Probandi juridical principle (In Dubio Pro Reo rule), Stern (2003).
8. Derived from the full posterior distribution, *ev* is a homogeneous computation calculus with the same two steps: constrained optimization and integration with the posterior density.
9. Computing time was not a great burden whenever FBST was used. The sophisticated numerical algorithms used could be considered a more serious obstacle to the popularization of the FBST.

ev was developed to be the Bayesian *pv* alternative, while maintaining the most desirable (known or perceived) properties in practical use. The list presented above seems to respond successfully to the challenge: the FBST is conceptually simple and elegant, theoretically coherent, and easily implemented for any statistical model, as long as the necessary computational procedures for numerical optimization and integration are available.

About the Author

Dr Carlos Pereira is a Professor and Head, Department of Statistics, University of São Paulo, Brazil. He is Past President of the Brazilian Statistical Society (1998–1990). He was the Director of the Institute of Mathematic and Statistics, São Paulo, Brazil (1994–1998). He was also Director of the Bioinformatic Scientific Center, University of São Paulo (2006–2009). He is an Elected member of the International Statistical Institute. He has authored and co-authored more than 150 papers and 4 books, including *Bayesian Analysis* (in Portuguese) in 1982 – the first Bayesian book published in Latin America. Professor Pereira has received the Ralph Bradley award from Florida State University in 1980. He was a research engineer at IEOR in Berkeley at the University of California (1986–1988). He was Associate editor of *Entropy*, *Environmetrics*, and *Brazilian J of Probability and Statistics*. Currently, he is the Statistical editor of the *Brazilian J of Psychiatry*. He was a member of both the Environmetrics Society and Board of Directors of *Entropy*.

Cross References

►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Significance Testing: An Overview

References and Further Reading

Basu D (1977) On the elimination of nuisance parameters. JASA 72:355–366

Hwang JT, Casella G, Robert C, Wells MT, Farrel RG (1992) Estimation of accuracy in testing. Ann Stat 20:490–509

Irony TZ, Pereira CA de B, Tiwari RC (2000) Analysis of opinion swing: comparison of two correlated proportions. Am Stat 54(1):57–62

Madruga MR, Esteves LG, Wechsler S (2001) On the Bayesianity of Pereira-Stern tests. Test 10:291–299

McNemar Q (1955) Psychological statistics. Wiley, New York

Pereira CA de B, Stern JM (1999) Evidence and credibility: full Bayesian significance test for precise hypotheses. Entropy 1: 69–80

Pereira CA de B, Stern JM, Wechsler S (2008) Can a significance test be genuinely Bayesian? Bayesian Anal 3(1):79–100

Stern JM (2007) Cognitive constructivism, eigen-solutions, and sharp statistical hypotheses. Cybernetics Human Knowing 14(1):9–36

Functional Data Analysis

Hans-Georg Müller
Professor of Statistics
University of California-Davis, Davis, CA, USA

Functional data analysis (FDA) refers to the statistical analysis of data samples consisting of random functions or surfaces, where each function is viewed as one sample element. Typically, the random functions contained in the sample are considered to be independent and smooth. FDA methodology is essentially nonparametric, utilizes smoothing methods, and allows for flexible modeling. The underlying random processes generating the data are sometimes assumed to be (non-stationary) ▶Gaussian processes.

Functional data are ubiquitous and may involve samples of density functions (Kneip and Utikal 2001) or hazard functions (Chiou and Müller 2009). Application areas include growth curves, econometrics, evolutionary biology, genetics and general kinds of longitudinal data. FDA methodology features functional principal component analysis (Rice and Silverman 1991), warping and curve registration (Gervini and Gasser 2004) and functional regression (Ramsay and Dalzell 1991). Theoretical foundations and asymptotic analysis of FDA are closely tied to perturbation theory of linear operators in Hilbert space (Bosq 2000). Finite sample implementations often require to address ill-posed problems with suitable regularization. A broad overview of applied aspects of FDA can be found in the textbook Ramsay and Silverman (2005).

The basic statistical methodologies of ANOVA, regression, correlation, classification and clustering that are available for scalar and vector data have spurred analogous developments for functional data. An additional aspect is that the time axis itself may be subject to random distortions and adequate functional models sometimes need to reflect such time-warping. Another issue is that often the random trajectories are not directly observed. Instead, for each sample function one has available measurements on a time grid that may range from very dense to extremely sparse. Sparse and randomly distributed measurement times are frequently encountered in longitudinal studies. Additional contamination of the measurements of the trajectory levels by errors is also common. These situations require careful modeling of the relationship between the recorded observations and the assumed underlying functional trajectories (Rice and Wu 2001; James and Sugar 2003; Yao et al. 2005). Initial analysis of functional data includes exploratory plotting of the observed functions in a "spaghetti plot" to obtain an initial idea of functional shapes, check for ▶outliers and identify "landmarks." Preprocessing may include outlier removal and curve alignment (registration) to adjust for time-warping.

Basic objects in FDA are the mean function μ and the covariance function G. For square integrable random functions $X(t)$,

$$\mu(t) = E(Y(t)),\ G(s,t) = \mathrm{cov}\{X(s), X(t)\},\quad s,t \in \mathcal{T}, \tag{1}$$

with auto-covariance operator $(Af)(t) = \int_{\mathcal{T}} f(s)G(s,t)\,ds$. This linear operator of Hilbert-Schmidt type has orthonormal eigenfunctions ϕ_k, $k = 1, 2, \ldots$, with associated ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$, such that $A\,\phi_k = \lambda_k\,\phi_k$. The foundation for functional principal component analysis is the Karhunen-Loève representation of random functions $X(t) = \mu(t) + \sum_{k=1}^{\infty} A_k\,\phi_k(t)$, where $A_k = \int_{\mathcal{T}}(Y(t) - \mu(t))\phi_k(t)\,dt$ are uncorrelated centered random variables with $var(A_k) = \lambda_k$.

Estimators employing smoothing methods (local least squares or splines) have been developed for various sampling schemes (sparse, dense, with errors) to obtain a data-based version of this representation, where one regularizes by truncating at a finite number K of included components. The idea is to borrow strength from the entire sample of functions rather than estimating each function separately. The functional data are then represented by the subject-specific vectors of score estimates $\hat{A}_k$, $k = 1, \ldots, K$, which can be used to represent individual trajectories and

for subsequent statistical analysis. Useful representations are alternatively obtained with pre-specified fixed basis functions, notably B-splines and wavelets.

Functional regression models may include one or several functions among the predictors, responses, or both. For pairs (X, Y) with centered random predictor functions X and scalar responses Y, the linear model is

$$E(Y|X) = \int_{\mathcal{T}} X(s)\beta(s)\, ds.$$

The regression parameter function β is usually represented in a suitable basis, for example the eigenbasis, with coefficient estimates determined by ►least squares or similar criteria. A variant, which is also applicable for classification purposes, is the generalized functional linear model $E(Y|X) = g\{\mu + \int_{\mathcal{T}} X(s)\beta(s)\, ds\}$ with link function g. The link function (and an additional variance function if applicable) is adapted to the (often discrete) distribution of Y; the components of the model can be estimated by quasi-likelihood.

The class of useful functional regression models is large. A flexible extension of the functional linear model is the functional additive model. Writing centered predictors as $X = \sum_{k=1}^{\infty} A_k \phi_k$, it is given by

$$E(Y|X) = \sum_{k=1}^{\infty} f_k(A_k)\phi_k$$

for smooth functions f_k with $E(f_k(A_k)) = 0$. Of practical relevance are models with varying domains, with more than one predictor function, and functional (autoregressive) time series models. In addition to the functional trajectories themselves, their derivatives are of interest to study the dynamics of the underlying processes.

Acknowledgments

Research partially supported by NSF Grant DMS-0806199.

About the Author

Hans-Georg Müller is Professor at the Department of Statistics, University of California, Davis, USA. For additional information, papers, and software go to http://www.stat.ucdavis.edu/mueller/.

Cross References

►Components of Statistics

References and Further Reading

Bosq D (2000) Linear processes in function spaces: theory and applications. Springer, New York

Chiou J-M, Müller H-G (2009) Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. J Am Stat Assoc 104:572–585

Gervini D, Gasser T (2004) Self-modeling warping functions. J Roy Stat Soc B Met 66:959–971

Hall P, Hosseini-Nasab M (2006) On properties of functional principal components analysis. J Roy Stat Soc B Met 68:109–126

James GM, Sugar CA (2003) Clustering for sparsely sampled functional data. J Am Stat Assoc 98:397–408

Kneip A, Utikal KJ (2001) Inference for density families using functional principal component analysis. J Am Stat Assoc 96: 519–542

Ramsay JO, Dalzell CJ (1991) Some tools for functional data analysis. J Roy Stat Soc B Met 53:539–572

Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer series in statistics. Springer, New York

Rice JA, Silverman BW (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. J Roy Stat Soc B Met 53:233–243

Rice JA, Wu CO (2001) Nonparametric mixed effects models for unequally sampled noisy curves. Biometrics 57:253–259

Yao F, Müller H-G, Wang J-L (2005) Functional data analysis for sparse longitudinal data. J Am Stat Assoc 100:577–590

Functional Derivatives in Statistics: Asymptotics and Robustness

Luisa Turrin Fernholz
Professor Emerita of Statistics
Temple University, Philadelphia, PA, USA

Introduction

Given a sample $X_1, \ldots, X_n$ of i.i.d. random variables with common distribution function (df) F and empirical df F_n, a statistic $S(X_1, \ldots, X_n)$ is called a *statistical functional* if it can be written in terms of a functional T, independent of n, such that $S(X_1, \ldots, X_n) = T(F_n)$ for all $n \geq 1$. The domain of T contains at least the population df F and the empirical df F_n for all $n \geq 1$. In this setting the statistic $T(F_n)$ estimates the parameter $T(F)$.

The sample mean is a statistical functional since $\overline{X} = 1/n \sum_1^n X_i = \int x dF_n(x) = T_1(F_n)$ which estimates the parameter $T_1(F) = \int x dF(x)$. The statistical functional corresponding to the sample median is $T_2(F_n) = F_n^{-1}(1/2) = \text{med}\{X_1, \ldots, X_n\}$ estimating the population median $T_2(F) = F^{-1}(1/2)$. Most statistics of interest are statistical functionals. They can be defined explicitly, such as T_1 and T_2, or implicitly, such as maximum likelihood type estimators or M-estimators which are solutions of equations in θ of the form $\int \psi(x, \theta)\, dF_n(x) = 0$.

Statistical functionals were introduced by von Mises (1947), who proposed the use of a functional derivative

called the Volterra derivative along with the corresponding Taylor expansion to obtain the asymptotic distribution of a statistic. However, the technical details were obscure with intractable notation and complicated regularity conditions. Consequently, the results appeared difficult to implement, and the von Mises theory was neglected until the late 1960s and 1970s with the surge of ▶robust statistics associated mainly with the work of Huber (1964, 1977) and Hampel (1968, 1974). For these new statistics, the statistical functional setting was found to be optimal for the study of robustness properties and the von Mises approach seemed to provide a natural environment for deriving the asymptotic distribution of the proposed robust estimates. During these robustness years the functional analysis concepts of differentiability and continuity were used to investigate the robustness aspects of the new statistics in addition to the asymptotics. In particular, the introduction of the influence function made a connection between robustness and classical asymptotics.

The Influence Function

Given a statistical functional T and a df F, the *influence function* of T at F is the real valued function $IF_{T,F}$ defined by

$$IF_{T,F}(x) = \lim_{t\to 0^+} \frac{T((1-t)F + t\Delta_x) - T(F)}{t},$$

where Δ_x is the d.f. of the point mass one at x. This function is normalized by setting $IF(x) = IF_{T,F}(x) - E_F(IF_{T,F}(X))$ so that $E_F(IF(X)) = 0$.

The influence function has played an important role in robust statistics. It was introduced by Hampel (1974), who observed that for large n, $IF_{T,F}(x)$ measures the effect on $T(F_n)$ of a single additional observation with value x. A bounded influence function indicates robustness of the statistic. For example, for the sample mean $T_1(F_n)$ as defined above, the influence function is $IF(x) = x - T_1(F)$. For the sample median $T_2(F_n)$, if $f = F'$, we have

$$IF(x) = \left[-1/(2f(T_2(F))\right] I_{\{x<(T_2(F)\}}(x) + \left[1/(2f(T_2(F))\right] I_{\{x\geq T_2(F)\}}(x).$$

Hence, the sample median with bounded influence function is more robust than the sample mean whose influence function is not bounded. A complete treatment of the robustness measures derived from the influence function can be found in Hampel et al. (1986).

In the framework of statistical functionals, the influence function can be viewed as a weak form of a functional derivative. Stronger derivatives were defined to analyze the asymptotic behavior of a statistic, but in all these derivatives the influence function is the crucial ingredient. It also provides a link between robustness and asymptotics as will be shown below.

Functional Derivatives

Consider a statistical functional T with domain an open set which lies in a normed vector space and contains a df F. A continuous linear functional T'_F is the *derivative* of T at F when

$$\lim_{t\to 0} \frac{T(F+tH) - T(F) - T'_F(tH)}{t} = 0, \qquad (1)$$

for H in subsets of the domain of T.

If (1) holds pointwise for each H, then T'_F is the *Gâteaux derivative*.

If (1) holds uniformly for all H in compact subsets of the domain of T, then T'_F is the *Hadamard derivative*.

If (1) holds uniformly for all H in bounded subsets of the domain of T, then T'_F is the *Fréchet derivative*.

Clearly Fréchet differentiability implies Hadamard differentiability, which implies Gâteaux differentiability. In all cases, the influence function is the central ingredient for any derivative since $T'_F(H) = \int IF_{T,F}(x)\, dH(x)$. Now, consider the Taylor expansion of T at F:

$$T(F+tF) - T(F) = T'_F(tH) + Rem$$

with

$$Rem = Rem(T, H, t) = o(t).$$

This remainder tends to zero either pointwise or uniformly according to whether F is Gâteaux, Hadamard, or Fréchet differentiable. For Hadamard derivatives see Reeds (1976) or Fernholz (1983) and for Fréchet derivatives see Huber (1981) or Serfling (1981).

When $t = 1/\sqrt{n}$ and $H = \sqrt{n}\,(F_n - F)$, the linear term of the Taylor expansion of T is

$$\int IF_{T,F}(x)\, d(F_n - F)(x) = \frac{1}{n}\sum_1^n IF(X_i),$$

where IF has been normalized, and the von Mises expansion of T at F is

$$T(F_n) = T(F) + \frac{1}{n}\sum_1^n IF(X_i) + Rem$$

or

$$\sqrt{n}(T(F_n) - T(F)) = \frac{1}{\sqrt{n}}\sum_1^n IF(X_i) + \sqrt{n}\, Rem.$$

When T is Hadamard or Fréchet differentiable, $\sqrt{n}\, Rem \to 0$ in probability, so that under certain regularity conditions for F we have the ▶asymptotic normality,

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \sigma^2).$$

In this case the influence function gives the asymptotic variance $\sigma^2 = E_F[IF(X)]^2$.

Remarks

The derivative used by von Mises for these calculations was similar to the Gâteaux derivative, so several strong regularity conditions had to be imposed on T to obtain its asymptotic normality. With the Hadamard or Fréchet derivatives these extra conditions are not needed.

It is important to note that the influence function plays a key role in these von Mises calculations. Note also that the influence function provides a link between robustness and asymptotics, and for this reason the von Mises approach via the influence function has become a useful method for obtaining asymptotic normality results.

The use of the Hadamard and Fréchet derivatives translates the problem of asymptotics into a problem of functional differentiability. Since Hadamard and Fréchet derivatives enjoy the chain rule property, we can show that a statistic $T(F_n)$ is asymptotically normal if the functional T is a composition of Hadamard or Fréchet differentiable functional components, where each component has a simple form. For references see Reeds (1976) or Fernholz (1983).

Higher Order Derivatives

The influence function is also called the *first kernel* since higher order derivatives can be defined for a real valued function T. If we set $\varphi_1(x) = IF(x)$ for the first kernel, the *second kernel* is

$$\varphi_2(x, y) = \frac{\partial^2}{\partial s\, \partial t} T\big(F(1 - s - t) + t\Delta_x + s\Delta_y\big)|_{t=0,s=0},$$

and in general, the *kernel of order* k is

$$\varphi_k(x_1, x_2, \ldots, x_k) = \frac{\partial^k}{\partial t_1\, \partial t_2 \ldots \partial t_k} T\left(F\left(1 - \sum_1^k t_i\right) + t_1\Delta_{x_1} + \ldots + t_k\Delta_{x_k}\right)\Big|_{(0,\ldots,0)}.$$

These kernels constitute the main ingredients for general Fréchet, Hadamard or Gâteaux higher order derivatives of T at F and for the corresponding higher order Taylor expansions. Hence, for $k \geq 2$ the k-th order von Mises expansion of $T(F_n)$ at F is:

$$T(F_n) - T(F) = \frac{1}{n}\sum_i \varphi_1(X_i) + \frac{1}{2n^2}\sum_{i,j}\varphi_2(X_i, X_j) + \cdots + \frac{1}{k!n^k}\sum_{i_1,\ldots,i_k}\varphi_k(X_{i_1}, \ldots, X_{i_k}) + Rem_k,$$

where, under certain differentiability conditions for T, the remainder of order k satisfies $Rem_k = o_P(n^{-k/2})$.

Higher order von Mises expansions were used to study the asymptotic distribution of a statistic when it is not normal (see von Mises 1947; Filippova 1972; Reeds 1976). These expansions are also useful to study the bias of a statistic since $E_F(T(F_n)) = T(F) + E_F(Rem_1)$, where

$$Rem_1 = \frac{1}{2n^2}\sum_{i,j}\varphi_2(X_i, X_j) + \cdots + \frac{1}{k!n^k}\sum_{i_1,\ldots,i_k}\varphi_k(X_{i_1}, \ldots, X_{i_k}) + Rem_k.$$

Results in this direction can be found in Sen (1988) and Fernholz (2001).

Multivariate Functionals

The formal von Mises calculations outlined above can be carried out for functionals of p variables after generalizing some basic rules of elementary calculus for the case of functional derivatives. Thus, if $T : \mathbb{R}^p \to \mathbb{R}$ and we have p samples of sizes $n_1, n_2, \ldots, n_p$ from the populations $F_1, \ldots, F_p$ respectively, we can consider the corresponding empirical df's $F_{n_1}, \ldots, F_{n_p}$. Then, the multivariate statistical functional $T(F_{n_1}, \ldots, F_{n_p})$ has p first order partial derivatives given by the corresponding multivariate influence function $\varphi_1 = (\varphi_{11}, \varphi_{12}, \ldots, \varphi_{1p})$, where for $1 \leq i \leq p$ the components are

$$\varphi_{1i}(x) = \frac{\partial T(F_1, \ldots, F_{i-1}, (1-\varepsilon)F_i + \varepsilon\Delta_x, F_{i+1}, \ldots, F_p)}{\partial \varepsilon}\Big|_{\varepsilon=0}.$$

Higher order partial derivatives can be found with the corresponding higher order von Mises expansions. For details, examples, and applications see Filippova (1972), Reeds (1976), and Fernholz (2001).

Statistical Functionals and the Bootstrap

Statistical functionals played a key role in the development of the bootstrap (see ▶Bootstrap Methods) introduced by Efron (1979). The "plug in" principle of Efron is essentially the study of a statistic in the setting of statistical functionals. After the bootstrap was introduced, the functional derivatives provided the answer for one of the basic asymptotic questions regarding the consistency of the bootstrap estimators $T(F_n^*)$. Does the bootstrap work when the von Mises method works? That is, does

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow{\mathcal{D}} N(0, \sigma^2) \text{ imply}$$
$$\sqrt{n}\left(T\left(F_n^*\right) - T(F_n)\right) \xrightarrow{\mathcal{D}} \mathrm{N}(0, \sigma^2) \ ?$$

The affirmative answer was given by R. Gill (1989) where he used von Mises expansions with Hadamard derivatives to show the asymptotic consistency of the bootstrap.

Smoothed Versions of Statistical Functionals

Using the convolution of F_n with a smooth df kernel sequence K_n we can obtain the smoothed version $\widetilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_n(x - X_i)$ of F_n. For a given statistical functional $T(F_n)$ estimating $T(F)$, we can consider the corresponding smoothed functional $T(\widetilde{F}_n)$ which, for continuous populations, may give a better estimate for $T(F)$. Some robustness aspects of $T(\widetilde{F}_n)$ can be analyzed through the influence function of T, and under reasonable regularity conditions for K_n, the ▶asymptotic normality of the smoothed version $T(\widetilde{F}_n)$ can be obtained when T is Hadamard differentiable. See Fernholz (1991, 1993).

About the Author

Luisa Turrin Fernholz is Professor Emerita of Statistics at Temple University, Philadelphia, PA (USA). She is currently the director of the Minerva Research Foundation and a member of the Advisory Council Committee for the Department of Mathematics at Princeton University as well as a member of the School of Mathematics Council for the Institute for Advanced Study, at Princeton, NJ (USA). She has previously held faculty positions at Princeton University, University of Pennsylvania, and the University of Buenos Aires. She is an elected member of the International Statistical Institute and a member of the ASA, the IMS, and the Bernoulli Society. She authored the research monograph *Von Mises Calculus for Statistical Functionals* (Springer Verlag, Lecture Notes in Statistics, Vol. 19, 1983). She co-edited several statistics volumes on Data Analysis and Robustness, and has published articles on probability theory as well as asymptotic expansions, robustness, functional derivatives, bias and variance reduction, and bootstrap applications, among other topics.

Cross References

▶Bootstrap Methods
▶Multivariate Technique: Robustness
▶Target Estimation: A New Approach to Parametric Estimation

References and Further Reading

Cabrera J, Fernholz LT (1999) Target estimation for bias and mean square error reduction. Ann Statist 27(3):1080–1104
Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Statist 7:1–26
Fernholz LT (1983) Von Mises calculus for statistical functionals. Lecture notes in statistics, vol 19. Springer, New York
Fernholz LT(1991) Almost sure convergence of smoothed empirical distribution functions. Scand J Statist 18:255–262
Fernholz LT (1993) Smoothed versions of statistical functionals. In: Morgenthaler S, Ronchetti E, Stahel W (eds) New directions in statistical data analysis and robustness, Birkhauser, London, pp 61–72
Fernholz LT (2001) On multivariate higher order von Mises expansions. Metrika 53(2):123–140
Filippova AA (1962) Mises theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. Theory Prob Appl 7:24–57
Gill RD (1989) Non- and semi-parametric maximum likelihood estimators and the von Mises Method, (part I). Scand J Statist 16:97–128
Hampel F (1974) The influence curve and its role in robust estimation. J Am Statist Assoc 69:383–393
Hampel F, Ronchetti E, Rousseeuw P, Stahel W (1986) Robust statistics. The approach based on the influence function. Wiley, New York
Huber P (1964) Robust estimation of a location parameter. Ann Math Statist 35:73–101
Huber P (1977) Robust Statistical Procedures, vol 27, Regional conference series in applied mathematics. SIAM, Philadelphia
Huber P (1981) Robust statistics. Wiley, New York
Reeds JA (1976) On the definition of von Mises functionals. PhD dissertation, Harvard University, Cambridge
Sen PK (1988) Functional jackknifing: rationality and general asymptotics. Ann Statist 16:450–469
von Mises R (1947) On the asymptotic distribution of differentiable statistical functions. Ann Math Statist 18:309–348

Fuzzy Logic in Statistical Data Analysis

Efendi N. Nasibov
Professor, Head of Department of Computer Science,
Faculty of Science and Arts
Dokuz Eylul University, Imzir, Turkey

Probability and Statistics with Fuzziness

Fuzzy logic and fuzzy sets theory first discussed in 1965 by Zadeh (Zadeh 1965). In classical sets theory, classifications are precise and the subject either belongs to a set or not. On the contrary, in fuzzy sets theory, the subject located in the border both belongs and does not belong to a set simultaneously. As a mathematical representation, in classical

set theory, if the object is the member of a set, it takes the membership value of 1; otherwise it takes the membership value of 0. However, in fuzzy logic, objects could have membership degrees between 0 and 1. In fuzzy logic, for example, a 30-year-old person could be the member of both the "young people" set with a membership degree of 0.6 and the "not young people" set with a membership degree of 0.4.

The relation between "fuzzy sets theory" and "statistics and probability theory" is and important research areas. In probability theory, realization of events is based on the classical 0–1 logic, i.e., an event occurs or does not occur. When the boundaries of classes that reflect the events are precise, such logic is valid. For example, when a dice has been rolled, the event of "coming up 1 or 2" is a precise event and it has a precise probability. But the event of "coming up a little number" is an imprecise event since its boundaries can not be stated; consequently, its probability can not be designated. In such situations, using probability theory together with fuzzy logic and fuzzy sets theory provides more admissible results.

Another important utilization of fuzzy logic and fuzzy sets theory is in statistical data analysis. With improvement of fuzzy sets theory, many studies have been made to combine statistical analysis methods and fuzzy sets theory. An analysis in which fuzzy logic is used is more robust than the classical logic. Furthermore, more reliable results can be obtained by a fuzzy approach (Rubin 1998).

There are many instances where fuzzy logic is used in statistical data analysis, including clustering, classification, regression, ►principal component analysis (PCA), independent component analysis (ICA), ►multidimensional scaling, ►time series, hypothesis tests, and confidence intervals (Coppi et al. 2006; Pop 2004; Taheri 2003; Mares 2007).

Fuzzy Clustering and Classification

Bellman et al. (1966) and Ruspini (1969) are the pioneers who used fuzzy sets theory in cluster analysis. Afterwards, many approaches were proposed on the use of fuzzy logic in cluster analysis. The most widely used approaches are based on fuzzy partitioning.

Fuzzy partitioning: The fuzzy partitioning of the data set $X = x_1, x_2, \ldots, x_n$ into fuzzy clusters $C_1, C_2, \ldots, C_c (1 < c < n)$ is denoted by the matrix $U_f = (u_{ij}) = (\mu_{C_i}(x_j))$, which satisfies the conditions given below:

$$0 \le u_{ij} \le 1, \quad \forall i \in \{1, 2, \ldots, c\}, \quad \forall j \in \{1, 2, \ldots, n\}, \tag{1}$$

$$0 < \sum_{j=1}^{n} u_{ij} < n, \quad \forall i \in \{1, 2, \ldots, c\} \tag{2}$$

where u_{ij} is the membership degree of the element x_j to the cluster C_i. In most cases, the following normalization condition as well as (1) and (2) is required for fuzzy partitioning:

$$\sum_{i=1}^{c} u_{ij} = 1, \quad \forall j \in \{1, 2, \ldots, n\}. \tag{3}$$

The first solution algorithms for the clustering approach based on fuzzy partitioning were proposed by Dunn (1973) and improved by Bezdek (1973).

In the most widely used fuzzy c-means (FCM) algorithm, the optimal fuzzy partitioning is obtained by minimizing the following function:

$$J_f(U_f, v_1, \ldots, v_c) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d(v_i, x_j)^2 \tag{4}$$

where c is the predetermined number of clusters, $d(v_i, x_j)$ is the distance between the cluster center v_i and the object x_j, and $m (m > 1)$ is the fuzziness index. The solution of (1)–(4) is found through the iterative computation of membership degrees and cluster centers:

$$u_{ij} = \frac{d(v_i, x_j)^{-2/(m-1)}}{\sum_{t=1}^{c} d(v_t, x_j)^{-2/(m-1)}}, \; i = 1, \ldots, c; \; j = 1, \ldots, n \tag{5}$$

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, \; i = 1, \ldots, c \tag{6}$$

The FCM algorithm is successful in finding spherical-shaped cluster structures. The Gustafson-Kessel algorithm based on FCM can find ellipsoidal cluster structures by using a covariance matrix. Fuzzy maximum likelihood estimation (FMLE) and the expectation maximization (EM) algorithms are also widely used fuzzy clustering algorithms (Doring et al. 2006).

Another approach for using fuzzy logic in cluster analysis is based on fuzzy neighborhood relations (FDBSCAN, FJP, FN-DBSCAN). In such an approach, the data are handled as fuzzy points and the classes are formed as crisp level sets based on the fuzziness level. Different clustering structures are obtained in different fuzziness levels. This approach could also be conceived as hierarchical clustering. The main point is to find the optimal hierarchy level and the optimal cluster structure convenient to this hierarchy level. In the fuzzy joint points based algorithms such as FJP, NRFJP, MFJP, such a problem has been solved by using an integrated cluster validity mechanism (Nasibov and Ulutagay 2007). The superiority of such algorithms over the FCM-based algorithms is not only the possibility

for finding arbitrarily shaped rather than only spherical-shaped clusters, but also not needing to determine the number of clusters in advance. On the other hand, using a fuzzy neighborhood relation among data can increase the robustness of clustering algorithm (Nasibov and Ulutagay 2009).

Classification is referred to as a supervised classification while clustering is referred to as an unsupervised classification. In a fuzzy supervised classification, the fuzzy partition X of elements is given in advance. One must specify the class of a datum x^* which is handled afterward. To do this, many approaches are used, including fuzzy inference system (FIS), fuzzy linear discriminant analysis, fuzzy k-nearest neighbors, and fuzzy-Bayesian classifier.

Fuzzy Regression

Fuzzy regression analysis is done by applying fuzzy logic techniques to classical regression models. There is no need for the assumptions of classical regression to hold for fuzzy regression. Moreover, fuzzy regression does not require normally distributed data, stability tests, or large samples. Classical regression analysis is able to respond to the needs of numerical science working with precise information. But in the social sciences, in which the personal perceptions are important, it is not easy to estimate the assumed appropriate and consistent estimators, because the concerned data are composed of imprecise, i.e., fuzzy data. In such situations, fuzzy logic can provide approximate ideas to reach adequate conclusions. There are various fuzzy regression models based on either fuzziness of the values of independent/dependent variables or fuzziness of the regression function (Näther 2006).

Usually, the fuzzy regression equation is as follows:

$$\tilde{y}_i = \tilde{b}_0 \oplus \tilde{b}_1 \odot \tilde{x}_{i1} \oplus \ldots \oplus \tilde{b}_p \odot \tilde{x}_{ip},\ i = 1, \ldots, n \qquad (7)$$

where $\oplus$ and $\odot$ are addition and multiplication processes on fuzzy numbers, $(\tilde{b}_0, \tilde{b}_1, \ldots, \tilde{b}_p)$ are fuzzy regression coefficients, $\tilde{y}_i$ is fuzzy response, and $(\tilde{x}_{i1}, \tilde{x}_{i2}, \ldots, \tilde{x}_{ip})$ are fuzzy explanatory variables.

The first study of fuzzy logic in regression analysis was made by Tanaka as fuzzy linear regression model (Tanaka et al. 1979, 1980, 1982). In Tanaka's approach, regression line is formed as a fuzzy linear function of data. Linear programming has been used to determine the parameters of this fuzzy function.

Another approach to fuzzy logic in regression analysis minimizes the sum of squares error between the observed and the predicted data which take fuzzy values. In determining the distance between fuzzy data, various fuzzy distances can be used and various models can be constructed (Diamond 1988; D'Urso 2003; Kim and Bishu 1998; Nasibov 2007).

As a third approach, Fuzzy c-Regression Models (FcRM), which arose from the technique of fuzzy clustering application on regression, can be specified. This approach, also called the switching regression, was proposed by Hathaway and Bezdek (Hathaway and Bezdek 1993). In this approach, all data are partitioned into distinct clusters since it is easier to express the structure with partial lines instead of a single regression function. The process works as in the FCM algorithm. The only difference is that, not only the membership degrees, but also the parameters of regression lines of the clusters are updated instead of cluster centers. The optimal value of the parameters has been found using the weighted least squares approach. For synthesis of the results, Fuzzy Inference Systems (FIS) such as Mamdani, Takagi-Sugeno-Kang (TSK), etc., can be used (Jang et al. 1997).

Fuzzy Principal Component Analysis

►Principal component analysis (PCA) is a preferred analysis method to reduce the dimension of the feature space and to extract information. PCA determines the linear combinations that describe maximum variability among the original data measurements. However, it is influenced by ►outliers, loss of data, and poor linear combinations. To resolve this problem, PCA models are created using fuzzy logic and the results are handled more efficiently than classical principal component analysis (Sarbu and Pop 2005). As with fuzzy regression, whole data sets are divided into fuzzy subsets in order to create better PCA models. Thus, the influence of outliers, which have minimum membership degree to clusters, is reduced.

The fuzzy covariance matrix for cluster A_i is constructed as follows:

$$C_{kl}^{(i)} = \frac{\sum_{j=1}^{n} [A_i(x^j)]^2 (x_{jk} - \bar{x}_k)(x_{jl} - \bar{x}_l)}{\sum_{j=1}^{n} [A_i(x^j)]^2}, \qquad (8)$$

where $A_i(x^j)$ indicates the membership degree of an object x^j to the cluster A_i and is inversely proportional with the distance between the object and the independent component.

One of the first studies about the fuzzy PCA was performed by Yabuuch and Watada in the construction of the principal component model using fuzzy logic for the elements in the fuzzy groups (Yabuuch and Watada 1997). The fuzzy PCA allows us to analyze the features of vague data samples. Hence, the fuzzy PCA gives more reliable results. Afterwards, the local fuzzy PCA method is used to reduce the dimension of feature vectors effectively. In this method, data space is partitioned into clusters using fuzzy clustering and then PCA is applied by constructing a fuzzy covariance

matrix (Lee 2004). In the study performed by Hsieh and Yang, fuzzy clustering is applied to find the hidden information in a DNA sequence by combining PCA and fuzzy adaptive resonance theory (fuzzy-ART) (Hsieh et al. 2008).

Fuzzy Independent Component Analysis

The recently developed and widely used independent component analysis (ICA) method is used to find linear form of non-Gaussian and statistically independent variables, and to extract information from databases (Hyvarinen et al. 2001). ICA is closely related to the blind source separation (BSS) method. The measurements $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of m unknown source signals ($\mathbf{s} = (s_1, s_2, \ldots, s_m)$) composed by unknown linear mixtures ($\mathbf{A}$) are performed:

$$\mathbf{x} = \mathbf{A}\mathbf{s}. \tag{9}$$

For the computational ease, $m = n$ is assumed. Hence, the matrix $\mathbf{A}$ is estimated by using the advantage of being an independent and non-Gaussian of source data and, using the matrix $\mathbf{W}$ (inverse of $\mathbf{A}$), source signals can be calculated from the equation below:

$$\mathbf{s} = \mathbf{W}\mathbf{x}. \tag{10}$$

The most widely used ICA algorithm is the Fast-ICA algorithm in terms of ease of use and speed. Honda et al. (2000) improved the Fast-ICA algorithm as the fuzzy Fast-ICA algorithm. In the fuzzy Fast-ICA algorithm, the Fuzzy c-Varieties (FCV) clustering method, which separates data into linear clusters using linear variability, is applied and then the local independent components in the fuzzy clusters are estimated by Fast-ICA algorithm.

Honda and Ichihashi (2008) have also proposed the fuzzy local ICA model as the improved version of the local ICA model. In the fuzzy local ICA model, fuzzy clustering, PCA, and multiple regression analysis are used simultaneously.

Gait biometrics have great advantages in comparison with the widely used biometrics such as face, fingerprint, and iris. In order to recognize gait, Lu et al. have developed a simple method based on human silhouette using genetic fuzzy vector machine (GFVM) and independent component analysis (Lu and Zhang 2007).

Fuzzy Time Series

The term "fuzzy time series" was first coined by Song and Chissom (Song and Chissom 1993ab, 1994).

Let $Y(t) \in R^1 (t = \ldots, 0, 1, 2, \ldots)$ be the universe of discourse on which fuzzy sets $f_i(t)(i = 1, 2, \ldots)$ are defined. Let $F(t)$ be a collection on $f_i(t)(i = 1, 2, \ldots)$. Then, $F(t)$ is called a fuzzy time series on $Y(t)(t = \ldots, 0, 1, 2, \ldots)$. In other words, fuzzy time series $F(t)$ is a chronological sequence of imprecise or fuzzy data ordered by time. Fuzzy time series are regarded as realizations of fuzzy random processes. In the fuzzy time series, fuzzy data as well as time-dependent dynamic relation can be considered as fuzzy:

$$F(t) = F(t-1) \circ \tilde{R}(t-1, t) \tag{11}$$

n^{th}-order fuzzy time series forecasting model, can be represented as follows:

$$F(t-1), F(t-2), \ldots, F(t-n) \rightarrow F(t) \tag{12}$$

For modeling of fuzzy time series, fuzzy ARMA, ARIMA processes, or fuzzy artificial neural networks are applied (Tseng et al. 2001; Zhang et al. 1998). Fuzzy time series can be analyzed and forecast by specifying an underlying fuzzy random process with the aid of generally applicable numerical methods.

Statistical Hypothesis Tests and Confidence Intervals

The main purpose of the traditional hypothesis test is to separate $\theta \in \Theta$ parameter space into two regions such as ω and $\Theta \backslash \omega$. The null and alternative hypotheses are as follows:

$$\begin{cases} H_0 : \theta \in \omega & (\textit{null hypothesis}) \\ H_1 : \theta \in \Theta \backslash \omega, & (\textit{alternative hypothesis}) \end{cases} \tag{13}$$

If the boundaries of ω and $\Theta \backslash \omega$ regions are assumed to be fuzzy, the fuzzy hypothesis test can be constructed as follows (Coppi et al. 2006):

$$\begin{cases} H_0 : \mu_\omega(\theta), & (\textit{null hypothesis}) \\ H_1 : \mu_{\Theta \backslash \omega(\theta)}, & (\textit{alternative hypothesis}) \end{cases} \tag{14}$$

Data handled in daily life are usually imprecise, i.e., fuzzy. For instance, water level of the river may not be fully measured due to fluctuations. In such a case, the well-known crisp hypothesis tests will not give reliable results.

Different approaches related to statistical hypothesis tests have been developed using fuzzy sets theory. First, Casals et al. (1986ab) and Casals and Gil (1989) have developed the ►Neyman-Pearson Lemma and Bayes method for statistical hypothesis tests with fuzzy data. There are two approaches to analyze statistical hypothesis tests: (1) observations are ordinary (crisp) and hypotheses are fuzzy (Arnold 1998), (2) both observations and hypotheses are fuzzy (Wu 2005). There may be some problems in applying classical statistical hypothesis to fuzzy observations. For instance, θ might be "approximately one" or θ might be "very large" and so on, where θ is any tested parameter. Bayes method might be useful for such types of hypothesis tests (Taheri and Behboodian 2001). However, if the fuzzy data are observed, the most appropriate method will be to apply fuzzy set theory to establish the statistical model.

In some approaches to using fuzzy logic in hypothesis tests, the estimators as fuzzy numbers are obtained using confidence intervals. If the estimator is a fuzzy number, the test statistic in hypothesis testing will also be a fuzzy number. Thus, the critical value at the hypothesis test is a fuzzy number. The result of this approach might be more realistic than a crisp hypothesis test (Buckley 2005). These results may be evaluated with probability theory (Hryniewicz 2006).

Fuzzy sets theory through the fuzzy random variables is applied to statistical confidence intervals for unknown fuzzy parameters. When the sample size is sufficiently large, an approximate fuzzy confidence interval could be constructed through a central limit theorem (Wu 2009). In case of fuzzy data, an interval estimation problem is formulated and the relation between fuzzy numbers and random intervals is found in (Coral et al. 1988; Gil 1992).

About the Author

Dr Efendi Nasibov is a Professor and Head, Department of Computer Science, Dokuz Eylul University, Turkey. He was a Professor and Head, Theoretical Statistics Division of the Department of Statistics, Dokuz Eylul University, Turkey (2006–2009). He was also the Head of the Department of Decision Making Models and Methods (2003–2009) of the Institute of Cybernetics, National Academy of Sciences of Azerbaijan. He is an elected member of the Academy of Modern Sciences named after Lotfi Zadeh, Baku, Azerbaijan. He has authored and co-authored more than 130 papers and 5 books.

Cross References

►Cluster Analysis: An Introduction
►Confidence Interval
►Data Analysis
►Expert Systems
►Forecasting: An Overview
►Fuzzy Set Theory and Probability Theory: What is the Relationship?; Fuzzy Sets: An Introduction
►Fuzzy Sets: An Introduction
►Hierarchical Clustering
►Multicriteria Clustering
►Neyman-Pearson Lemma
►Principal Component Analysis
►Statistical Methods for Non-Precise Data

References and Further Reading

Arnold BF (1998) Testing fuzzy hypothesis with crisp data. Fuzzy Set Syst 94:323–333
Bellman RE, Kalaba RE, Zadeh LA (1966) Abstraction and pattern classification. J Math Anal Appl 2:581–586
Bezdek JC (1973) Fuzzy mathematics in pattern classification. PhD Thesis, Cornell University, Ithaca, New York
Bezdek JC (1974) Cluster validity with fuzzy sets. J Cybernetics 3:58–73
Buckley JJ (2005) Fuzzy statistics: hypothesis testing. Soft Comput 9:512–518
Casals MR, Gil MA (1989) A note on the operativeness of Neyman–Pearson tests with fuzzy information. Fuzzy Set Syst 30:215–220
Casals MR, Gil MA, Gil P (1986a) On the use of Zadeh's probabilistic definition for testing statistical hypotheses from fuzzy information. Fuzzy Set Syst 20:175–190
Casals MR, Gil MA, Gil P (1986b) The fuzzy decision problem: An approach to the problem of testing statistical hypotheses with fuzzy information. Euro J Oper Res 27:371–382
Coppi R, Gil MA, Kiers HAL (2006) The fuzzy approach to statistical analysis. Comput Stat Data Anal 51:1–14
Corral N, Gil MA (1988) A note on interval estimation with fuzzy data. Fuzzy Set Syst 28:209–215
D'Urso P (2003) Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. Comput Stat Data Anal 42:47–72
Diamond P (1988) Fuzzy least squares. Inform Sci 46:141–157
Doring C, Lesot MJ, Kruse R (2006) Data analysis with fuzzy clustering methods. Comput Stat Data Anal 51:192–214
Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybernetics 3:32–57
Gil MA (1992) A note on the connection between fuzzy numbers and random intervals. Stat Prob Lett 13:311–319
Hathaway RJ, Bezdek JC (1993) Switching regression models and fuzzy clustering. IEEE Trans Fuzzy Syst 3:195–204
Honda K, Ichihashi H (2008) Fuzzy local ICA for extracting independent components related to external criteria. Appl Math Sci 2(6):275–291
Honda K, Ichihashi H, Ohue M, Kitaguchi K (2000) Extraction of local independent components using fuzzy clustering. In Proceedings of 6th International Conference on Soft Computing, pp 837–842
Hryniewicz O (2006) Possibilistic decisions and fuzzy statistical tests. Fuzzy Set Syst 157:2665–2673
Hsieh KL, Yang IC (2008) Incorporating PCA and fuzzy-ART techniques into achieve organism classification based on codon usage consideration. Comput Biol Med 38:886–893
Hyvarinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York
Jang JSR, Sun CT, Mizutani E (1997) Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice-Hall, Englewood Cliffs
Kim B, Bishu RR (1998) Evaluation of fuzzy linear regression models by comparing membership functions. Fuzzy Set Syst 100:343–352
Lee KY (2004) Local fuzzy PCA based GMM with dimension reduction on speaker identification. Pattern Recogn Lett 25: 1811–1817
Lu J, Zhang E (2007) Gait recognition for human identification based on ICA and fuzzy SVM through multiple views fusion. Pattern Recogn Lett 28:2401–2411
Mares M (2007) Fuzzy data in statistics. Kybernetica 43(4): 491–502
Nasibov EN (2007) Fuzzy least squares regression model based on weighted distance between fuzzy numbers. Automat Contr Comput Sci 41(1):10–17

Nasibov EN, Ulutagay G (2007) A new unsupervised approach for fuzzy clustering. Fuzzy Set Syst 158:2118–2133
Nasibov EN, Ulutagay G (2009) Robustness of density-based clustering methods with various neighborhood relations. Fuzzy Set Syst 160:3601–3615
Näther W (2006) Regression with fuzzy random data. Comput Stat Data Anal 51:235–252
Pop HF (2004) Data analysis with fuzzy sets: a short survey. Studia University of Babes-Bolyai, Informatica XLIX(2):111–122
Rubin SH (1998) A fuzzy approach towards inferential data mining. Comput Ind Eng 35(1–2):267–270
Ruspini EH (1969) A new approach to clustering. Inform Contr 15:22–32
Sarbu C, Pop HF (2005) Principal component analysis versus fuzzy principal component analysis: a case study: the quality of Danube water (1985–1996). Talanta 65:1215–1220
Song Q, Chissom BS (1993a) Forecasting enrollments with fuzzy time series-part I. Fuzzy Set Syst 54:1–9
Song Q, Chissom BS (1993b) Fuzzy time series and its models. Fuzzy Set Syst 54:269–277
Song Q, Chissom BS (1994) Forecasting enrollments with fuzzy time series-part II. Fuzzy Set Syst 62:1–8
Taheri SM (2003) Trends in fuzzy statistics. Austr J Stat 32(3):239–257
Taheri SM, Behboodian J (2001) A Bayesian approach to fuzzy hypothesis testing. Fuzzy Set Syst 123:39–48
Tanaka H, Okuda T, Asai K (1979) Fuzzy information and decision in statistical model. In Gupta MM et al. (eds) Advances in fuzzy set theory and applications. North-Holland, Amsterdam, pp 303–320
Tanaka H, Uejima S, Asai K (1980) Fuzzy linear regression model. IEEE Trans Syst Man Cybernet 10:2933–2938
Tanaka H, Uejima S, Asai K (1982) Linear regression analysis with fuzzy model. IEEE Trans Syst Man Cybernet 12:903–907
Tseng FM, Tzeng GH, Yu HC, Yuan BJC (2001) Fuzzy ARIMA model for forecasting the foreign exchange market. Fuzzy Set Syst 118:9–19
Wu HC (2005) Statistical hypotheses testing for fuzzy data. Inform Sci 175:30–56
Wu HC (2009) Statistical confidence intervals for fuzzy data. Expert Syst Appl 36:2670–2676
Yabuuch Y, Watada J (1997) Fuzzy principal component analysis and its application. Biomedical Fuzzy Hum Sci 3(1):83–92
Zadeh LA (1965) Fuzzy sets. Inform Contr 8(3):338–353
Zhang GP, Eddy PB, Hu YM (1998) Forecasting with artificial neural networks: the state of the art. Int J Forecast 14:35–62

Fuzzy Set Theory and Probability Theory: What is the Relationship?

Lotfi A. Zadeh
Professor Emeritus
University of California-Berkeley, Berkeley, CA, USA

Relationship between probability theory and fuzzy set theory is associated with a long history of discussion and debate. My first paper on fuzzy sets was published in 1965 (Zadeh 1965). In a paper published in 1966, Loginov suggested that the membership function of a fuzzy set may be interpreted as a conditional probability (Loginov 1966). Subsequently, related links to probability theory were suggested and analyzed by many others (Coletti and Scozzafava 2004; Freeling 1981; Hisdal 1986a, b; Nurmi 1977; Ross et al. 2002; Singpurwalla and Booker 2004; Stallings 1977; Thomas 1995; Viertl 1987; Yager 1984). Among such links are links to set-valued random variables (Goodmanm and Nguyen 1985; Orlov 1980; Wang and Sanchez 1982) and to the Dempster–Shafer theory (Dempster 1967; Shafer 1976). A more detailed discussion of these links may be found in my l995 paper "Probability theory and fuzzy logic are complementary rather than competitive," (Zadeh 1995).

In reality, probability theory and fuzzy set theory are distinct theories with different agendas. Scientific theories originate in perceptions. Primitive perceptions such as perceptions of distance, direction, weight, loudness, color, etc. crystallize in early childhood. Among basic perceptions which crystallize at later stages of development are those of likelihood, count, class, similarity and possibility. Fundamentally, probability theory may be viewed as a formalization of perceptions of likelihood and count; fuzzy set theory may be viewed as a formalization of perceptions of class and similarity; and possibility theory may be viewed as a formalization of perception of possibility. It should be noted that perceptions of likelihood and possibility are distinct. Fuzzy set theory and possibility theory are closely related (Zadeh 1978). A key to a better understanding of the nature of the relationship between probability theory and fuzzy set theory is the observation that probability theory is rooted in perceptions of likelihood and count while fuzzy set theory is rooted in perceptions of class and similarity.

In debates over the nature of the relationship between probability theory and fuzzy set theory, there are four schools of thought. The prevailing view within the Bayesian community is that probability theory is sufficient for dealing with uncertainty and imprecision of any kind, implying that there is no need for fuzzy set theory. An eloquent spokesman for this school of thought is an eminent Bayesian, Professor Dennis Lindley. Here is an excerpt of what he had to say on this subject.

> *The only satisfactory description of uncertainty is probability. By this I mean that every uncertainty statement must be in the form of a probability; that several uncertainties must be combined using the rules of probability; and that the calculus of probabilities is adequate to handle all situations involving*

uncertainty… probability is the only sensible description of uncertainty and is adequate for all problems involving uncertainty. All other methods are inadequate… anything that can be done with fuzzy logic, belief functions, upper and lower probabilities, or any other alternative to probability can better be done with probability (Lindley 1987).

The second school of thought is that probability theory and fuzzy set theory are distinct theories which are complementary rather than competitive. This is the view that is articulated in my l995 Technometrics paper (Zadeh 1995). The third school of thought is that standard probability theory, call it PT, is in need of generalization through addition to PT of concepts and techniques drawn from fuzzy set theory and, more generally, from fuzzy logic, with the understanding that fuzzy set theory is a branch of fuzzy logic. Basically, fuzzy logic, FL, is a precise system of reasoning, computation and deduction in which the objects of discourse are fuzzy sets, that is, classes in which membership is a matter of degree. Thus, in fuzzy logic everything is, or is allowed to be, a matter of degree.

It is important to observe that any bivalent-logic-based theory, T, may be generalized through addition of concepts and techniques drawn from fuzzy logic. Such generalization is referred to as FL-generalization. The view that standard probability theory, PT, can be enriched through FL-generalization is articulated in my 2002 paper "Toward a perception-based theory of probabilistic reasoning" (Zadeh 2002), 2005 paper "Toward a generalized theory of uncertainty (GTU) – an outline" (Zadeh 2005) and 2006 paper "Generalized theory of uncertainty (GTU) – principal concepts and ideas" (Zadeh 2006). The result of FL-generalization, call it PTp, is a generalized theory of probability which has a key capability – the capability to deal with information which is described in a natural language and, more particularly, with perception-based probabilities and relations which are described in a natural language. What is not widely recognized is that many, perhaps most, real-world probabilities are perception-based. Examples: What is the probability that Obama will succeed in solving the financial crisis? What is the probability that there will be a significant increase in the price of oil in the near future? Such probabilities are perception-based and non-numerical. Standard probability theory provides no facilities for computation and reasoning with non-numerical, perception-based probabilities.

The fourth school of thought is that FL-generalization of probability theory should be accompanied by a shift in the foundations of probability theory from bivalent logic to fuzzy logic. This is a radical view which is put forth in my 2004 paper "Probability theory and fuzzy logic – a radical view" (Zadeh 2004).

Is probability theory sufficient for dealing with any kind of uncertainty and imprecision? Professor Lindley's answer is: Yes. In a paper published in 1986 entitled "Is probability theory sufficient for dealing with uncertainty in AI: A negative view," (Zadeh 1986) I argued that the answer is: No. In contradiction to Professor Lindley's assertion, here are some simple examples of problems which do not lend themselves to solution through the use of standard probability theory.

In these examples X is a real-valued variable.
X is larger than approximately a
X is smaller than approximately b

What is the probability that X is approximately c?

Usually X is larger than approximately a
Usually X is smaller than approximately b

What is the probability that X is approximately c?

Usually X is much larger than approximately a
Usually X is much smaller than approximately b

What is the probability that X is approximately c?
What is the expected value of X?

Usually it takes Robert about an hour to get home from work
Robert left work at about 5 pm

What is the probability that Robert is home at 6:l5 pm?

In these examples, question-relevant information is described in natural language. What these examples underscore is that, as was alluded to earlier, standard probability theory does not provide methods of deduction and computation with information described in natural language. Lack of this capability is a serious limitation of standard probability theory, PT. To add this capability to PT it is necessary to FL-generalize PT through addition to PT of concepts and techniques drawn from fuzzy logic.

What would be gained by going beyond FL-generalizat ion of PT, and shifting the foundations of PT from bivalent logic to fuzzy logic? There is a compelling reason for such a shift. At this juncture, most scientific theories, including probability theory, are based on bivalent logic. In bivalent-logic-based theories, the basic concepts are defined as bivalent concepts, with no shades of truth allowed. In reality, most basic concepts are fuzzy, that is, are a matter of degree. For example, in probability theory the concept of independence is defined as a bivalent concept, meaning that two events A and B are either independent or not independent,

with no degrees of independence allowed. But what is quite obvious is that the concept of independence is fuzzy rather than bivalent. The same applies to the concepts of event, stationarity and more generally, to most other basic concepts within probability theory. A shift in the foundations of probability theory would involve a redefinition of bivalent concepts as fuzzy concepts. Such redefinition would enhance the ability of probability theory to serve as a model of reality.

What is widely unrecognized at this juncture is that (a) the capability of probability theory to deal with real-world problems can be enhanced through FL-generalization. Even more widely unrecognized is that (b) the ability of probability theory to serve as a model of reality can be further enhanced through a shift in the foundations of probability theory from bivalent logic to fuzzy logic. But as we move further into the age of machine intelligence and automated decision-making the need for (a) and (b) will become increasingly apparent. I believe that eventually (a) and (b) will gain acceptance.

Acknowledgments

Research supported in part by ONR N00014-02-1-0294, BT Grant CT1080028046, Omron Grant, Tekes Grant, Chevron Texaco Grant, The Ministry of Communications and Information Technology of Azerbaijan and the BISC Program of UC Berkeley.

About the Author

Lotfi Zadeh was born in 1921 in Baku, Azerbaijan. After his PhD from Columbia University in 1949 in Electrical Engineering, he taught at Columbia for ten years till 1959 where he was a full professor. He joined the Electrical Engineering Department at the University of California, Berkeley in 1959 and served as its chairman from 1963 to 1968. Presently, he is a professor in the Graduate school serving as the Director of BISC (Berkeley Initiative in Soft Computing). He authored a seminal paper in fuzzy sets in 1965. This landmark paper initiated a new direction, which over the past three decades led to a vast literature, and a rapidly growing number of applications ranging from consumer products to subway trains and decision support systems. For this seminal contribution he received the Oldenburger medal from the American Society of Mechanical Engineers in 1993, the IEEE Medal of Honor in 1995, the Okawa prize in 1996, the B. Bolzano Medal from the Academy of Sciences of the Czech Republic, and the Benjamin Franklin Medal in Electrical Engineering. He is a member of the National Academy of Engineering and a Foreign Member of the Polish, Finnish, Korean, Bulgarian, Russian and Azerbaijan Academy of Sciences. He has single-authored over two hundred papers and serves on the editorial boards of over fifty journals. Dr. Zadeh is a recipient of twenty-six honorary doctorates.

Cross References

►Fuzzy Logic in Statistical Data Analysis; Fuzzy Sets: An Introduction
►Fuzzy Sets: An Introduction
►Philosophy of Probability
►Probability Theory: An Outline

References and Further Reading

Coletti G, Scozzafava R (2004) Conditional probability, fuzzy sets, and possibility: a unifying view. Fuzzy Set Syst 144(1):227–249
Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38:325–329
Dubois D, Nguyen HT, Prade H (2000) Possibility theory, probability and fuzzy sets: misunderstandings, bridges and gaps. In: Dubois D, Prade H (eds) Fundamentals of fuzzy sets. The handbooks of fuzzy sets series. Kluwer, Boston, MA, pp 343–438
Freeling ANS (1981) Possibilities versus fuzzy probabilities – Two alternative decision aids. Tech. Rep. 81–6, Decision Science Consortium Inc., Washington, DC
Goodman IR, Nguyen HT (1985) Uncertainty models for knowledge-based systems. North Holland, Amsterdam
Hisdal E (1986) Infinite-valued logic based on two-valued logic and probability. Part 1.1: Difficulties with present-day fuzzy-set theory and their resolution in the TEE model. Int J Man-Mach Stud 25(1):89–111
Hisdal E (1986) Infinite-valued logic based on two-valued logic and probability. Part 1.2: Different sources of fuzziness. Int J Man-Mach Stud 25(2):113–138
Lindley DV (1987) The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. Statistical Science 2:17–24
Loginov VJ (1966) Probability treatment of Zadeh membership functions and their use in pattern recognition, Eng Cybern 68–69
Nurmi H (1977) Probability and fuzziness: some methodological considerations. Unpublished paper presented at the sixth research conference on subjective probability, utility, and decision making, Warszawa
Orlov AI (1980) Problems of optimization and fuzzy variables. Znaniye, Moscow
Ross TJ, Booker JM, Parkinson WJ (eds) (2002) Fuzzy logic and probability applications: bridging the gap. Society for Industrial and Applied Mathematics, Philadelphia, PA
Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton, NJ
Singpurwalla ND, Booker JM (2004) Membership functions and probability measures of fuzzy sets. J Am Stat Assoc 99:467
Stallings W (1977) Fuzzy set theory versus Bayesian statistics. IEEE Trans Syst Man Cybern, SMC-7:216–219
Thomas SF (1995) Fuzziness and probability, ACG Press, Wichita KS
Viertl R (1987) Is it necessary to develop a fuzzy Bayesian inference? In: Viertl R (ed) Probability and Bayesian statistics. Plenum, New York, pp 471–475

Wang PZ, Sanchez E (1982) Treating a fuzzy subset as a projectable random set. In: Gupta MM, Sanchez E (eds) Fuzzy information and decision processes. North Holland, Amsterdam, pp 213–220
Yager RR (1984) Probabilities from fuzzy observations. Inf Sci 32:1–31
Zadeh LA (1965) Fuzzy sets. Inform Contr 8:338–353
Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Set Syst 1:3–28
Zadeh LA (1986) Is probability theory sufficient for dealing with uncertainty in AI: a negative view. In: Kanal LN, Lemmer JF (eds) Uncertainty in artificial intelligence. North Holland, Amsterdam
Zadeh LA (1995) Probability theory and fuzzy logic are complementary rather than competitive. Technometrics 37:271–276
Zadeh LA (2002) Toward a perception-based theory of probabilistic reasoning with imprecise probabilities. J Stat Plan Inference 105:233–264
Zadeh LA (2004) Probability theory and fuzzy logic – a radical view. J Am Stat Assoc 99(467):880–881
Zadeh LA (2005) Toward a generalized theory of uncertainty (GTU) – an outline. Inf Sci 172:1–40
Zadeh LA (2006) Generalized theory of uncertainty (GTU) – principal concepts and ideas. Comput Stat Data Anal 51:15–46

Fuzzy Sets: An Introduction

Madan Lal Puri
Professor Emeritus
King Fahd University of Petroleum and Minerals,
Dhahran, Saudi Arabia
Indiana University, Bloomington, IN, USA

Some of the basic properties and implications of the concepts of fuzzy set theory are presented. The notion of a fuzzy set is seen to provide a convenient point of departure for the construction of a conceptual framework which parallels in many respects the framework used in the case of ordinary sets but is more general than the latter. The material presented is from the basic paper of Zadeh (1965) who introduced the notion of fuzzy sets. The reader is also referred to Rosenfeld (1982) for a brief survey of some of the concepts of fuzzy set theory and its application to pattern recognition (see ►Pattern Recognition, Aspects of and ►Statistical Pattern Recognition Principles).

Introduction

In everyday life we often deal with imprecisely defined properties or quantities–e.g., "a few books," "a long story," "a popular teacher," "a tall man," etc. More often than not, the classes of objects which we encounter in the real physical world do not have precisely defined criteria of membership. For example, consider the class of animals. This class clearly includes dogs, horses, birds, etc. as its members, and clearly excludes rocks, fluids, plants, etc. However, such objects as starfish, bacteria, etc. have an ambiguous status with respect to the class of animals. The same kind of ambiguity arises in the case of a number such as 10 in relation to the "class" of all numbers which are much greater than 1.

Clearly, the class of all real numbers which are much greater than 1, or "the class of tall men" do not constitute classes in the usual mathematical sense of these terms. Yet, the fact remains that such imprecisely defined "classes" play an important role in human thinking, particularly, in the domain of pattern recognition, communication of information, decision theory, control theory and medical diagnosis, among others.

The purpose of this note is to provide in a preliminary way some of the basic properties and implications of a concept which is being used more and more in dealing with the type of "classes"mentioned above. The concept in question is that of a "fuzzy set" with a continuum of grades of membership, the concept introduced by Zadeh (1965) in order to allow imprecisely defined notions to be properly formulated and manipulated.

Over the past 20–25 years there has been a tremendous growth of literature on fuzzy sets amounting by now to over 2,000 papers and several textbooks; there is even a journal devoted to this subject.

This note is intended to provide a brief survey of some of the basic concepts of fuzzy sets and related topics.

We begin with some basic definitions.

Definitions

Let $\mathcal{X}$ be a space of points (objects), with a generic element of $\mathcal{X}$ denoted by x. Thus, $\mathcal{X} = \{x\}$.

A fuzzy set (class) A in $\mathcal{X}$ is characterized by a *membership (characteristic) function* $f_A(x)$ which associates with each point x in $\mathcal{X}$ a real number in the interval $[0,1]$, with the value of $f_A(x)$ at x representing the "grade of membership" or "the degree of membership" of x in A. The key idea in fuzzy set theory is that an element has a "degree of membership" in a fuzzy set, and we usually assume that this degree is a real number between 0 and 1. The nearer the value of $f_A(x)$ to unity, the higher the degree of membership of x in A. In the case of the "fuzzy set" of tall men, the elements are men, and their degrees of membership depend on their heights; e.g., a man who is 5 ft tall might have degree 0, a man who is $6\frac{1}{2}$ ft tall might have degree 1, and men of intermediate heights might have intermediate degrees. Analogous remarks apply to such fuzzy sets as the set of young women, the set of rich people, the set of first rate mathematicians, and so on. When A is a set in the ordinary sense of the term, its membership function can take

on only two values 0 and 1, with $f_A(x) = 1$ if $x \in A$ or 0 if $x \notin A$. Thus, in this case, $f_A(x)$ reduces to the familiar characteristic function or indicator function of a so-called crisp set A.

It may be noted that the notion of a fuzzy set is completely nonstatistical in nature, and the rules which are commonly used for manipulating fuzzy set memberships are not the same as the rules for manipulating probabilities.

The Algebra of Fuzzy Subsets

The rules for combining and manipulating fuzzy subsets of $\mathcal{X}$ (Blurrian algebra) should reduce to the rules of ordinary subset algebra when subsets are crisp. This motivates the fuzzy subset algebra introduced by Zadeh (1965), where $\leq$, sup ($\vee$) and inf ($\wedge$) play the roles of $\subseteq$, $\cup$, and $\cap$, respectively. We say that

1. A fuzzy set is empty iff (if and only if) its membership function is $\equiv 0$ on $\mathcal{X}$.
2. Two fuzzy sets A and B are *equal* (and we write $A = B$) iff $f_A(x) = f_B(x)\ \forall\ x \in \mathcal{X}$. (Instead of writing $f_A(x) = f_B(x)\ \forall\ x \in \mathcal{X}$, we shall write $f_A = f_B$).
3. The *complement* of a fuzzy set A is denoted by A' and is defined as $f'_A = 1 - f_A$.
4. $A \subset B$ iff $f_A \leq f_B$.
5. The *union* of two fuzzy sets A and B with respect to respective membership functions $f_A(x)$ and $f_B(x)$ is a fuzzy set C, (and we write $C = A \cup B$) whose membership function is related to those of A and B by

$$f_C(x) = \max[f_A(x), f_B(x)] \quad \forall\ x \in \mathcal{X} \text{ i.e., } f_C = f_A \vee f_B. \tag{I}$$

(Note that the union has the associative property, i.e., $A \cup (B \cup C) = (A \cup B) \cup C$. Also note that a more intuitive way of defining the union is the following: The union of A and B is the smallest fuzzy set containing both A and B. More precisely, if D is any fuzzy set which contains both A and B, then it also contains the union of A and B.)

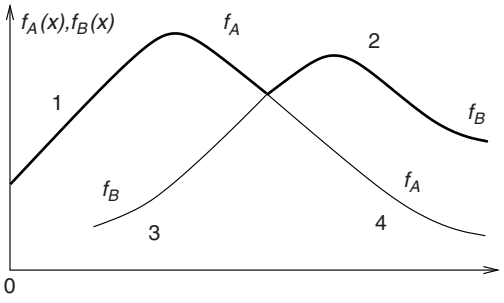
6. The *intersection* of two fuzzy sets A and B with respect to their respective membership functions $f_A(x)$ and $f_B(x)$ is a fuzzy set C (written as $C = A \cap B$) whose membership function f_C is related to those of A and B by $f_C(x) = \min[f_A(x), f_B(x)]\ \forall\ x \in \mathcal{X}$ i.e.,

$$f_C = f_A \wedge f_B.$$

As in the case of union, it is easy to show that the intersection of A and B is the *largest* fuzzy set which is contained in both A and B.

7. A and B are *disjoint* if $A \cap B = C$ is empty, i.e., $f_C(x) \equiv 0\ \forall\ x \in \mathcal{X}$.

Note that $\cap$, like union, has the associative property. Also note that the notion of "belonging" which plays a fundamental role in the case of ordinary sets, does not have the same role in the case of fuzzy sets. Thus, it is not meaningful to speak of a point x "belonging" to a fuzzy set A except in the trivial sense of $f_A(x)$ being positive. Less trivially, one can introduce two levels α and β ($0 < \alpha < 1$, $\beta < \alpha$, $0 < \beta < 1$) and agree to say that (1) $x \in A$ if $f_A(x) \geq \alpha$, (2) $x \notin A$ if $f_A(x) \leq \beta$; and (3) x has an intermediate status relative to A, if $\beta < f_A(x) < \alpha$. This leads to a three valued logic with three truth values: T ($f_A(x) \geq \alpha$), F ($f_A(x) \leq \beta$), and U ($\beta < f_A(x) < \alpha$). Note that the empty and universal (fuzzy) subsets are just the constant functions 0 and 1; they are in fact non-fuzzy!



(Curve segments 1 and 2 comprise the membership function of the union (heavy lines). Curve segments 3 and 4 comprise the membership function of the intersection).

It is clear that these definitions are the extensions of the definitions of $\subseteq$, $\cup$ and $\cap$ for ordinary sets. It is also trivial to verify that they have properties analogous to those of $\subseteq$, $\cup$ and $\cap$, e.g., $A \cup B$ is the $\cap$ of all fuzzy sets $C \supset A \subset C$ and $B \subset C$; and $A \cap B$ is the union of all fuzzy sets $C \supset C \subset A$ and $C \subset B$. Evidently $\subseteq$ is a partial order relation, and $\cup$ and $\cap$ are commutative, associative and distributive over each other. It is also easy to extend many of the basic identities which hold for ordinary sets to fuzzy sets. As examples, we have the De Morgan's Laws:

$$(A \cup B)' = A' \cap B' \tag{1}$$

$$(A \cap B)' = A' \cup B'. \tag{2}$$

To prove (1), for example, note that the left hand side

$$= 1 - \max[f_A, f_B]$$

$$= \min[1 - f_A, 1 - f_B] = \min[f_{A'}, f_{B'}]$$

$$= \text{Right hand side.}$$

This can easily be verified by testing it for two possible cases: $f_A(x) > f_B(x)$ and $f_A(x) < f_B(x)$. Essentially fuzzy sets in $\mathcal{X}$ constitute a distributive lattice with a 0 and 1 (Birkoff, 1948).

8. The *algebraic product* of A and B is denoted by AB, and is defined in terms of membership functions of A and B by the relation $f_{AB} = f_A \cdot f_B$. Clearly $AB \subset A \cap B$.
9. The *algebraic sum* of A and B is denoted by $A + B$ and is defined as

$$f_{A+B} = f_A + f_B$$

provided $f_A + f_B \leq 1$.

Thus, unlike the algebraic product, the algebraic sum is meaningful only if $f_A(x) + f_B(x) \leq 1 \; \forall \; x \in X$.

10. The *absolute difference* of A and B is denoted by $|A - B|$ and is defined as $f_{|A-B|} = |f_A - f_B|$. Note that in the case of ordinary sets, $|A - B|$ reduces to the relative complement of $A \cap B$ in $A \cup B$. ($|A - B|$ is the symmetric difference $A \triangle B = (A - B) \cup (B - A)$).
11. The dual of algebraic product is the sum $A \oplus B = (A'B')' = A + B - AB$. (Note that for ordinary sets $\cap$ and the algebraic product are equivalent operations, as are $\cup$ and $\oplus$.)

Convex Combination

By a convex combination of two vectors f and g is usually meant a linear combination of f and g of the form $\lambda f + (1 - \lambda)g$ where $0 \leq \lambda \leq 1$. The mode of combining f and g can be generalized to fuzzy sets in the following manner:

Let A, B and C be arbitrary fuzzy sets. *The convex combination* A, B and C is denoted by $(A, B; C)$ and is defined by the relation

$$(A, B; C) = CA + C'B$$

where C' is the complement of C. In terms of membership functions, this means

$$f_{(A,B;C)}(x) = f_C(x)f_A(x) + [1 - f_C(x)]f_B(x), \quad x \in \mathcal{X}.$$

A basic property of the convex combination of A, B and C is expressed as

$$A \cap B \subset (A, B; C) \subset A \cup B \quad \forall \; C.$$

This is an immediate consequence of

$$\min[f_A(x), f_B(x)] \leq \lambda f_A(x) + (1 - \lambda)f_B(x) \leq \max[f_A(x), f_B(x)], \quad x \in \mathcal{X}$$

which holds for all λ in $[0, 1]$. It is interesting to observe that given any fuzzy set C satisfying $A \cap B \subset C \subset A \cup B$, one can always find a fuzzy set $D \ni C = (A, B; D)$. The membership function of this set D is given by

$$f_D(x) = \frac{f_C(x) - f_B(x)}{f_A(x) - f_B(x)}, \quad x \in \mathcal{X}.$$

Functions

What about functions? Let f be a function from $\mathcal{X}$ into T, and let A be a fuzzy subset of $\mathcal{X}$ with membership function μ_A. Then, the image of A under f is defined in terms of its membership function μ_A by

$$[f(\mu_A)](y) \equiv \begin{cases} \sup\limits_{f(x)=y} \mu_A(x) \; \forall \; y \in T & \text{i.e., } \sup\limits_{x \in \inf f^{-1}(y)} \mu_A(x) \; \forall \; y \in T \\ 0 & \text{if } f^{-1}(y) = \emptyset. \end{cases}$$

Similarly if B is a fuzzy subset of T, then the preimage or inverse image of B under f is defined in terms of its membership function μ_B by

$$[f^{-1}(\mu_B)](x) \equiv \mu_B(f(x)) \quad \forall \; x \in \mathcal{X} \text{ i.e., } f^{-1}(\mu_B) \equiv \mu_B \circ f.$$

Explanation

Let $f : X \to T$. Let B be a fuzzy set in T with membership $\mu_B(y)$. The inverse mapping f^{-1} induces a fuzzy set A in $\mathcal{X}$ whose membership function is defined by

$$\mu_A(x) = \mu_B(y), \quad y \in T$$

for all x in $\mathcal{X}$ which are mapped by f into y.

Consider now the converse problem. Let A be a fuzzy set in $\mathcal{X}$, and as before, let $f : \mathcal{X} \to T$. Question: What is the membership function for the fuzzy set B in T which is induced by this mapping? If f is not $1 : 1$, then an ambiguity arises when two or more distinct points in $\mathcal{X}$, say x_1 and x_2, with different grades of membership in A, are mapped into the same point y in T. In this case, the question is: What grade of membership in B should be assigned to y? To resolve this ambiguity, we agree to assign the larger of the grades of membership to y. More generally, the membership function for B will be defined by

$$\mu_B(y) = \max_{x \in f^{-1}(y)} \mu_A(x), \quad y \in T$$

where $f^{-1}(y) = \{x; x \in X; f(X) = y\}$. Evidently these definitions generalize the standard definitions of the image and the preimage of a subset, and one can verify that these definitions are compatible with fuzzy subset algebra in the usual ways, e.g., one can show that f and f^{-1} have the following properties:

(a) $f^{-1}\left(\bigvee\limits_{i \in I} \mu_{A_i}\right) = \bigvee\limits_{i \in I} f^{-1}(\mu_{A_i})$

(b) $f^{-1}\left(\bigwedge\limits_{i \in I} \mu_{A_i}\right) = \bigwedge\limits_{i \in I} f^{-1}(\mu_{A_i})$

also

(c) $f\left(\underset{i\in I}{\vee}\mu_{A_i}\right) = \underset{i\in I}{\vee} f(\mu_{A_i})$

(d) $f\left(\underset{i\in I}{\wedge}\mu_{A_i}\right) \le \underset{i\in I}{\wedge} f(\mu_{A_i})$

(e) $\overline{f(\mu_A)} \le f(\overline{\mu}_A); \overline{f^{-1}(\mu_B)} = f^{-1}(\overline{\mu}_B).$

$\overline{\quad}$ means complement.

(f) $f(f^{-1}(\mu_B)) \le \mu_B; f^{-1}(f(\mu_A)) \ge \mu_A.$

Proof (a): $f^{-1}\left(\underset{i\in I}{\vee}\mu_A\right) = \left(\underset{i\in I}{\vee}\mu_{A_i}\right)\circ f = \underset{i\in I}{\vee}(\mu_{A_i}\circ f) = \underset{i\in I}{\vee} f^{-1}(\mu_{A_i})$ and so on.

(c)
$$f\left[\underset{i}{\vee}\mu_{A_i}\right](y) = \begin{cases} \underset{f(x)=y}{\sup}\left[\underset{i}{\vee}\mu_{A_i}\right](x) \\ 0 \qquad \text{if } x = f^{-1}(y) = \emptyset \end{cases}$$
$$= \begin{cases} \underset{i}{\vee}\ \underset{x\in f^{-1}(y)}{\sup}\ [\mu_{A_i}(x)] = \underset{i}{\vee} f(\mu_{A_i}) \\ 0 \end{cases}$$

Fuzzy Relations

The concept of a *relation* has a natural extension to fuzzy sets and plays an important role in the theory of such sets and their applications, just as it does in the case of ordinary sets. Ordinarily, a relation is defined as a set of ordered pairs, e.g., the set of all ordered pairs of real numbers x and y such that $x \ge y$. In the context of fuzzy sets, a *fuzzy relation* in $\mathcal{X}$ is a fuzzy set in the product space $\mathcal{X}\times\mathcal{X}$, e.g., the relation denoted by $x \gg y$; $x, y \in \mathbb{R}$ may be regarded as a fuzzy set A in $\mathbb{R}^2$, with the membership function of A, $f_A(x,y)$ having the following (subjective) representative values: $f_A(10,5) = 0; f_A(100,10) = 0.7, f_A(100,1) = 1$, etc.

More generally, one can define an *n–ary fuzzy relation* in $\mathcal{X}$ as a fuzzy set A in the product space $\mathcal{X}\times\mathcal{X}\times\ldots\times\mathcal{X}$. For such relations, the membership function is of the form $f_A(x_1,\ldots,x_n)$ where $x_i \in \mathcal{X}$, $i = 1,\ldots,n$.

In the case of binary fuzzy relations, the composition of two fuzzy relations A and B is denoted by $B\circ A$, and is defined as a fuzzy relation in $\mathcal{X}$ whose membership function is related to those of A and B by

$$f_{B\circ A}(x,y) = \sup_v \min[f_A(x,v), f_B(v,y)]$$
$$= \underset{v}{\vee}[f_A(x,v)\wedge f_B(v,y)]$$

$\forall\ x, y$ and v in $\mathcal{X}$. (Note also that this generalizes the usual definition.) Note that the operation of composition has the associative property: $A\circ(B\circ C) = (A\circ B)\circ C$.

Convexity

A fuzzy set A is *convex* iff the sets Γ_α defined by

$$\Gamma_\alpha = \{x; f_A(x) \ge \alpha\} \tag{3}$$

are convex for all α in the interval $(0,1]$.

An alternative and more direct definition of convexity is the following: A fuzzy set A is *convex* iff

$$f_A[\lambda x_1 + (1-\lambda)x_2] \ge \min[f_A(x_1), f_A(x_2)] \tag{4}$$

for all x_1 and x_2 in $\mathcal{X}$ and all λ in $[0,1]$.

Note that this definition does not imply that the function $f_A(x)$ must be a convex function of x.

It can be seen that the two definitions are equivalent (see Zadeh 1965).

A basic property of convex fuzzy sets is:

Theorem *If A and B are fuzzy convex, then $A\wedge B$ is also fuzzy convex.*

Boundedness

A fuzzy set A is bounded iff the sets $\Gamma_\alpha = \{x; f_A(x) \ge \alpha\}$ are bounded for all $\alpha > 0$; i.e., for all $\alpha > 0$, $\exists$ a finite $R(\alpha) \ni \|x\| \le R(\alpha)$ for all x in Γ_α.

If A is a bounded set, then for all $\varepsilon > 0$, $\exists$ a hyperplane $H \ni f_A(x) \le \varepsilon\ \forall\ x$ on the side of H which does not contain the origin. For example consider the set $\Gamma_\varepsilon = \{x; f_A(x) \ge \varepsilon\}$. By hypothesis this set is contained in a sphere S of radius $R(\varepsilon)$. Let H be any hyperplane supporting S. Then, all points on the side of H which does not contain the origin lie outside or on S, and hence for all such points $f_A(x) \le \varepsilon$.

Preliminary

As a preliminary, let A and B be two bounded fuzzy sets and let H be a hypersurface in $\mathbb{R}^{(n)}$ defined by the equation $h(x) = 0$ with all points for which $h(x) \ge 0$ being on one side of H and all points for which $h(x) \le 0$ being on the other side. Let K_H be a number dependent on $H \ni f_A(x) \le K_H$ on one side of H and $f_B(x) \le K_B$ on the other side. Let $M_H = \inf K_H$. The number $D_H = 1 - M$ is called the *degree of separation* of A and B by H.

In general one is concerned not with a given hypersurface H, but with a family of hypersurfaces $\{H_\lambda\}$, with λ

ranging over $\mathbb{R}^{(m)}$. The problem then is to find a member of this family which realizes the highest degree of separation.

A special case of this problem is one where the H_λ are hyperplanes in $\mathbb{R}^{(n)}$, with λ ranging over $\mathbb{R}^{(n)}$. In this case we define the *degree of separation* of A and B by

$$D = 1 - M \quad \text{where} \quad M = \inf_H M_H.$$

Separation of Convex Fuzzy Sets

The classical separation theorem for ordinary convex sets states, in essence, that if A and B are disjoint convex sets, then there exists a separating hyperplane H such that A is on one side of H and B is on the other side. This theorem can be extended to convex fuzzy sets, without requiring that A and B be disjoint, since the condition of disjointedness is much too restrictive in the case of fuzzy sets. It turns out that the answer is in the affirmative.

Theorem *Let A and B be bounded convex fuzzy sets in $\mathbb{R}^{(n)}$, with maximal grades M_A and M_B respectively i.e., $M_A = \sup_x f_A(x)$, $M_B = \sup_x M_B(x)$. Let M be the maximal grade for the intersection $A \cap B$ (i.e., $M = \sup_x \min[f_A(x), f_B(x)]$). Then $D = 1 - M$. (D is called the degree of separation of A and B by the hyperplane H).*

In other words, the theorem states that the highest degree of separation of two convex fuzzy sets that can be achieved with a hyperplane in $R^{(n)}$ is one minus the maximal grade in the intersection $A \cap B$. Zadeh has applied these types of results in the problems of optimization, pattern discrimination, etc.

Concluding Remarks

The concepts of fuzzy sets and fuzzy functions have been found useful in many applications, notably in pattern recognition, clustering, information retrieval and systems analysis, among other areas (cf. Negoita and Ralescu 1975). Motivated by some of these applications and related problems, Puri and Ralescu (1982, 1983) introduced the integration on fuzzy sets and differentials of fuzzy functions. This led to the study of fuzzy random variables, their expectations, concept of normality for fuzzy random variables and different limit theorems for fuzzy random variables (cf. Puri and Ralescu (1985, 1986, 1991), Klement, Puri and Ralescu (1984, 1986), and Proske and Puri (2002a, b and the references cited in these papers).

About the Author

Professor Puri was ranked the fourth most prolific statistician in the world for his writings in the top statistical journals in a 1997 report by the Natural Sciences and Engineering Research Council of Canada. Among statisticians in universities which do not have separate departments of statistics, Puri was ranked number one in the world by the same report. Puri has received a great many honors for his outstanding contributions to statistics and we mention only a few. Professor Puri twice received the Senior U.S. Scientist Award from Germany's Alexander von Humboldt Foundation, and he was honored by the German government in recognition of past achievements in research and teaching. Madan Puri has been named the recipient of the 2008 Gottfried E. Noether Senior Scholar Award (an annual, international prize honoring the outstanding statisticians across the globe), for "outstanding contributions to the methodology and/or theory and teaching of nonparametric statistics that have had substantial, sustained impact on the subject, its practical applications and its pedagogy." For many years Professor Puri has been highly cited researcher in mathematics according to ISI Web of knowledge ISI HighlyCited. com According to For many years Professor Puri has been highly cited researcher in mathematics according to ISI Web of Knowledge ISI HighlyCited.Com Professor Puri, his greatest honor came in 2003 when under the editorship of Professors Peter Hall, Marc Hallin, and George Roussas, the International Science Publishers published "Selected Collected Works of Madan L. Puri," a series of three volumes, each containing about 800 pages.

Cross References

►Fuzzy Logic in Statistical Data Analysis: Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Statistical Methods for Non-Precise Data

References and Further Reading

Klement EP, Puri ML, Ralescu DA (1984) Law of large numbers and central limit theorem for fuzzy random variables. Cybern Syst Anal 2:525–529
Klement EP, Puri ML, Ralescu DA (1986) Limit theorems for fuzzy random variables. Proc R Soc London 407:171–182
Negoita CV, Ralescu DA (1975) Applications of fuzzy sets to system analysis. Wiley, New York
Proske F, Puri ML (2002a) Central limit theorem for Banach space valued fuzzy random variables. Proc Am Math Soc 130: 1493–1501
Proske F, Puri ML (2002b) Strong law of large numbers for Banach space valued fuzzy random variables. J Theor Probab 15:543–552
Puri ML, Ralescu DA (1982) Integration on fuzzy sets. Adv Appl Math 3:430–434
Puri ML, Ralescu DA (1983) Differentials of fuzzy functions. J Math Anal Appl 91:552–558

Puri ML, Ralescu DA (1985) The concept of normality for fuzzy random variables. Ann Probab 13:1373–1379
Puri ML, Ralescu DA (1986) Fuzzy random variables. J Math Anal Appl 114:409–422
Puri ML, Ralescu DA (1991) Limit theorems for fuzzy martingales. J Math Anal Appl 160:107–122
Rozenfeld A (1982) How many are a few? Fuzzy sets, fuzzy numbers, and fuzzy mathematics. Math Intell 4:139–143
Singpurwalla ND, Booker JM (2004) Membership functions and probability measures of fuzzy sets. J Am Stat Assoc 99: 867–889
Zadeh LA (1965) Fuzzy sets. Inform Contr 8:338–353

G

Gamma Distribution

Kimiko O. Bowman[1], L. R. Shenton[2]
[1]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]Professor Emeritus of Statistics
University of Georgia, Athens, GA, USA

Introduction

For the gamma random variable X there is the probability function

$$Pr(X = x; s, a, \rho) = \frac{1}{a\Gamma(\rho)}\left(\frac{x-s}{a}\right)^{\rho-1}\exp\left(\frac{s-x}{a}\right) \quad (s < x < \infty; a, \rho > 0),$$

where a scale, ρ shape, and s location parameter. The gamma distribution is unimodal but may be inverse J-shaped. The moments are

$$\begin{aligned} &\textit{Mean}: \quad \mu_1' = a\rho + s, \quad \textit{variance}: \quad \mu_2 = a^2\rho, \\ &\textit{Skewness}: \quad \sqrt{\beta_1} = \mu_3/\mu_2^{3/2} = 2/\sqrt{\rho}, \\ &\textit{Kurtosis}: \quad \beta_2 = \mu_4/\mu_2^2 = 3 + 6/\rho. \end{aligned}$$

If $a = 1$, and $s = 0$, the distribution becomes the exponential distribution. If $s = 0$, $a = 2$, and $\rho = \nu/2$, the distribution becomes the familiar χ^2_ν distribution with ν degree of freedom.

$$P_{\chi^2_\nu}\chi = \frac{\chi^{\frac{\nu}{2}-1}e^{-\chi/2}}{2^{\nu/2}\Gamma(\nu/2)} \quad \chi \geq 0.$$

Also the gamma distribution is a Type III in the Pearson manifold of distributions.

The Fig. 1 shows distribution functions of $s = 0$, $a = 1$ and $\rho = 1, 2$ and 4.

Estimation

The maximum likelihood estimators $\hat{a}$ and $\hat{\rho}$ for the two parameter gamma distribution ($s = 0$) are the solution to the two equations

$$\begin{aligned} \ln(\hat{\rho}) - \psi(\hat{\rho}) &= \ln(A/G) \quad (\psi(x) = d\ln\Gamma(x)/dx) \\ \hat{\rho}\hat{a} &= A \end{aligned}$$

where $A = \sum X_j/n$, and $G = \sqrt[n]{X_1, X_2, \cdots, X_n}$. For the three parameter case, the equations to be solved are

$$\psi(\hat{\rho}) + \ln(\hat{s}) = n^{-1}\sum_{j=1}^{n}\ln(X_j - \hat{s}),$$

$$n^{-1}\sum_{j=1}^{N}(X_j - \hat{s}) = \hat{a}\hat{\rho},$$

$$n^{-1}\sum_{j=1}^{N}(X_j - \hat{s})^{-1} = (\hat{a}(\hat{\rho} - 1))^{-1}.$$

We must be aware that the moments of maximum likelihood estimators exist if $\rho > 1$ for the mean, $\rho > 2$ for the variance, $\rho > 3$ for skewness and $\rho > 4$ for kurtosis.

The moment estimators in the 2 parameter case are simple in form and are

$$a^*\rho^* = m_1, \quad a^{*2}p^* = m_2,$$

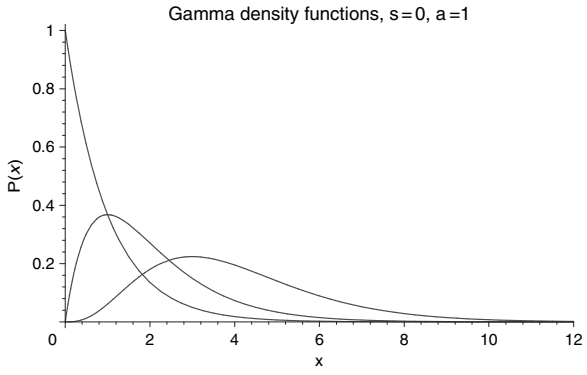
where m_1 is the sample mean, $m_1 = \sum_{j=1}^{n}(X_j)/n = \bar{X}$ and $m_2 = \sum_{j=1}^{n}(X_j - \bar{X})^2/n$ for a random sample $X_1, X_2, \cdots, X_n$. For the three parameter gamma distribution we use $2/\sqrt{\rho^*}$=sample skewness to determine ρ.

Properties of Estimators

Johnson et al. (1994) states "Estimation of parameters of gamma distribution has also received extensive attention in the literature in the last two decades. The contributions of Bowman and Shenton and of A.C. Cohen and his coworkers should be particularly mentioned. Bowman and Shenton (1988) monograph provides detailed analysis of maximum likelihood estimators for two-parameter gamma distribution with emphasis on the shape parameter and presents valuable information on distributions and moments of these estimators, including their joint distributions. Careful discussion of estimation problems associated with the three-parameter gamma density is also presented. ····· The authors also deal with the moments of the moments estimators. The list of references in the monograph covers the development of the authors' work on this area from 1968 onward."

Cohen and Whitten (1988) introduced mixed maximum likelihood approach, including the use of the smallest term of the sample; this approach avoids the difficulty that

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,



Gamma Distribution. Fig. 1 Gamma density functions, $s = 0$, $a = 1$ and $\rho = 1, 2$ and 4

Gamma Distribution. Table 1 Gamma distribution sampling and negative skewness ($\mu_3/\mu_2^{3/2}$)

		$\sqrt{b_1} > 0$			$\sqrt{b_1} < 0$		
ρ	n	S	F	$\bar{S}$	S	F	$\bar{S}$
2	50	20,000	5,420	166	0	5	0
2	500	20,000	0	19	0	0	0
6	50	20,000	2,157	1,245	0	432	0
6	500	20,000	0	1,771	0	0	0
12	50	20,000	6,047	1,706	0	1,862	0
12	500	20,000	272	1,239	0	0	0

S gives the number of solutions subscribing to the tolerance. F gives the number of failures. $\bar{S}$ gives the number of cases when no solution was found after 200 iterates.

the usual maximum likelihood method, which moments only exist if $\rho > 1$ for the mean, $\rho > 2$ for the variance. $\rho > 3$ for skewness and $\rho > 4$ for kurtosis. Balakrishanan and Cohen (1991) introduced estimators with an emphasis on modified estimators and censored samples.

For the details of methods to solve the equations see Bowman and Shenton (1988) and Johnson et al. (1994).

Negative Skewness in Gamma Sampling

With various values of ρ, samples of size 50 and 500 were taken and skewness $\sqrt{b_1}$ analyzed. In Table 1 the analysis is given showing the frequency of negative skewness.

We note: (a) No solutions were found when the sample skewness was negative. The failure rate in this case increases with large ρ and small n. (b) For small n, the failure rate (F) is high for both small and large ρ. (c) For large ρ, the abortive rate ($\bar{S}$) is high, and doubtless would become higher with more stringent tolerance.

A modified model has been given by Cheng and Traylor (1995) for which negative skewness is included in as a possibility. The new modified gamma distribution is

$$g(x;\mu,\sigma,\lambda) = \frac{1}{\sigma\lambda\Gamma(\lambda^2)}\left\{\lambda^{-2}\left[1+\frac{\lambda(\lambda-\mu)}{\sigma}\right]\right\}^{\lambda^{-2}-1} \exp\left\{-\frac{1}{\lambda^2}\left[1+\frac{\lambda(x-\mu)}{\sigma}\right]\right\}.$$
$$(\sigma > 0; \lambda \neq 0 \text{ and } 1+\lambda(x-\mu)/\sigma > 0)$$

In our notation

$$\rho = 1/\lambda^2, \quad a = \sigma\,|\,\lambda\,|, \quad s = \mu - \sigma\lambda^{-1}.$$

Moreover $\sqrt{\beta_1} = 2/\sqrt{\rho}$ = skewness.

Multiple Parameter Distributions and Mixtures

Everett and Hand (1981) study mixtures of Poisson distributions, and binomial distributions, giving, possible solutions. Bowman and Shenton refer to Poisson mixtures (2003, 2004, 2006). A global approach to the subject is due to Karlis and Xekalaki (2005) enlarging the concept of the Poissonian. The generalization of ideas could be applied to continuous univariate distributions.

Other works may be mentioned such as Bowman and Shenton (1998), distribution of ratio of gamma variate, Hirose (1995), three parameter gamma distribution, Revfeim (1991), inverse gamma.

Conclusion

The extensive paper of Karlis and Xekalaki (2005) opens up many new concepts for research. Simulation studies using computer facilities provide a powerful tool. In addition there is the remarkable power of symbolic codes such as Maple, Mathematica, etc.

About the Author

For biographies of both authors *see* the entry ►Omnibus Test for Departures from Normality.

Cross References

►Bivariate Distributions
►Chi-Square Distribution
►Dispersion Models
►Frailty Model
►Multivariate Statistical Distributions
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview

References and Further Reading

Bowman KO, Beauchamp JJ (1975) Pitfalls with some gamma variate simulation routines. J Stat Comput Sim 4:141–154

Bowman KO, Shenton LR (1968) Properties of estimators for the gamma distribution. Report CTC-1, Union Carbide Corp., Oak Ridge, Tennessee

Bowman KO, Shenton LR (1970) Small sample properties of estimators for the gamma distribution. Report CTC-28, UCCND, Oak Ridge, Tennessee

Bowman KO, Shenton LR (1978) Coefficient of variation on sampling from gamma universe. ICQC'78, Tokyo, D1-19–D1-24

Bowman KO, Shenton LR (1982) Properties of estimators for the gamma distribution. Commun Stat Sim Comput 11:377–519

Bowman KO, Shenton LR (1983a) Maximum likelihood estimators for the gamma distribution revisited. Commun Stat Sim Comput 12:697–710

Bowman KO, Shenton LR (1983b) The distribution of the standard deviation and skewness in gamma sampling – a new look at a Craig–Pearon study. ORNL/CSD-109

Bowman KO, Shenton LR (1988) Properties of estimators for the gamma distribution. Marcel Dekker, New York

Bowman KO, Shenton LR (1999) The asymptotic moment profile and maximum likelihood: application to gamma ratio densities. Commun Stat Theor Meth 28(10):2497–2508

Bowman KO, Shenton LR (2002) Problems with maximum likelihood estimators and the 3 parameter gamma distributions. J Stat Comput Sim 72(5):391–401

Bowman KO, Shenton LR (2003) Canonical dichotomized alternant determinants and Poisson mixtures. Far East J Theor Stat 10(2):87–109

Bowman KO, Shenton LR (2004) Poisson mixtures, asymptotic variance, and alternant determinants. Far East J Theor Stat 14(1):79–87

Bowman KO, Shenton LR (2006) Maximum likelihood estimators for normal and gamma mixtures. Far East J Theor Stat 20(2):217–240

Bowman KO, Shenton LR, Karlof C (1995) Estimation problems associated with the three parameter gamma distribution. Commun Stat Theor Meth 24(5):1355–1376

Bowman KO, Shenton LR, Gailey PC (1998) Distribution of ratio of gamma variate. Commun Stat Sim Comput 27(1):1–19

Cheng RCH, Traylor L (1995) Non-regular maximum likelihood problems. J R Stat Soc B 57(1):3–44

Cohen AC (1950) Estimating parameters of Pearson type III populations from truncated samples. JASA 45:411–423

Cohen AC (1951) Estimation of parameters in truncated Pearson frequency distributions. Ann Math Stat 22:256–265

Cohen AC, Whitten BJ (1982) Modified moment and modified maximum likelihood estimators and for parameters of the three-parameter gamma distribution. Commun Stat Sim Comput 11:197–216

Cohen AC, Whitten BJ (1988) Parameter estimation in reliability and life span modals. Marcel Dekker, New York

Everett BS, Hand DJ (1981) Finite mixture distributions. Chapman and Hall, London

Hirose H (1995) Maximum likelihood parameter estimation in the three-parameter Gamma distribution. Comput Stat Data Anal 20:343–354

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol 1, 2nd edn. Wiley, New York, NY

Karlis D, Xekalaki E (2005) Mixed Poisson distributions. Int Stat Rev 73(1):35–58

Pearson K (1948) Karl Pearson's early statistical papers. Cambridge University Press, England

Revfeim KJA (1991) Approximation for the cumulative and inverse gamma distribution. Statistica Neelandica 45:327–331

Shenton LR, Bowman KO (1972) Further remarks on m.l.e. for the gamma distribution. Technometrics 14:725–733

Shenton LR, Bowman KO (1973) Comments on the gamma distribution and uses in the rainfall data. In The Third Conference on Probability and Statistics in Atmospheric Science, 8 pp

Gaussian Processes

Herbert K. H. Lee
Professor
University of California-Santa Cruz, Santa Cruz, CA, USA

A Gaussian process (GP) specifies a distribution for values over a set (typically R^n) which could be an interval of time or a region of space, but could also be more arbitrary, such as the space of a set of explanatory variables. It is often used for modeling quantities that are spatially or temporally correlated, such as rainfall. In this way it differs from standard statistical models where the data are assumed to be independent given the model; here the data are assumed to be correlated and the correlation structure is part of the model as well. Applications are widespread, including time series, geostatistics (where it first appeared as kriging), and general approximation of functions. GPs have also become popular in machine learning, where they are used for regression and classification tasks.

A Gaussian process is a continuously-defined process such that its values at any finite collection of locations jointly have a multivariate Gaussian distribution (e.g., Cressie 1993). This model is highly flexible and can be used for nonparametric modeling. In practice, two common simplifying assumptions are often made: stationarity and isotropy. The idea of stationarity is that the distribution of the process does not depend on location, i.e., all points have the same mean and marginal variance, and the joint distribution of any finite collection of points is invariant to translation. Sometimes the marginal mean of the field is not constant, so the mean is modeled separately (such as with a linear model or a low-order polynomial) and then the de-trended field is treated as a stationary GP. A stationary field can also be isotropic, in that the covariance between any two points depends only on the distance between them (rather than also depending on the orientation).

Examples of Gaussian Processes include the Wiener process and the Ornstein-Uhlenbeck process (Cox and Miller 1965). A Wiener process is a mean-zero process such

that all increments of the process (the difference in values at two points) are independent and normally distributed with mean zero and variance proportional to the length of the increment. A generalization allows for a drift term μ. This process can be represented via a stochastic differential equation:

$$dW(t) = \mu dt + \sigma Z(t)\sqrt{dt},$$

where $W(t)$ is the value of the process at time $t \in R$, σ^2 is a variance term, and $Z(t)$ are independent standard normals. An Ornstein-Uhlenbeck process is related, but it is a mean-reverting process, so the distribution of its increments depends on the current value of the process:

$$dX(t) = -\beta X(t)dt + \sigma Z(t)\sqrt{dt} = -\beta X(t)dt + dW(t),$$

where $W(t)$ is a Wiener process and β controls the rate of mean reversion. An Ornstein-Uhlenbeck process is stationary, while a Wiener process is not.

If we observe a Gaussian process $X(s)$ at a set of locations $s_1, \ldots, s_n \in \mathcal{S}$, we can write its distribution in terms of its mean function $\mu(s)$ and its covariance function $C(s_i, s_j)$ as:

$$X \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $X = (X(s_1), \ldots, X(s_n))$, $\boldsymbol{\mu} = (\mu(s_1), \ldots, \mu(s_n))$ and $\boldsymbol{\Sigma}$ is the variance-covariance matrix with elements $C(s_i, s_j)$. Under the typical assumptions of stationarity and isotropy, $\mu(s_i) = \mu$ for all i, and the elements of the covariance matrix simplify to:

$$C(s_i, s_j) = \theta_1 \rho(d/\theta_2),$$

where θ_1 is the marginal variance, d is the distance between s_i and s_j, $\rho(\cdot)$ is a correlation function, and θ_2 specifies the range of spatial dependence. The correlation function must be nonnegative definite. Common parameterizations of the correlation for spatial and functional applications include:

spherical correlogram	$\rho(d) = \left(1 - \frac{3}{2}d + \frac{1}{2}d^3\right) I_{\{0 \le d \le 1\}}(d)$
exponential correlogram	$\rho(d) = e^{-d}$
Gaussian correlogram	$\rho(d) = e^{-d^2}$
Matérn class	$\rho(d) = \left[(d/2)^{\nu} 2K_{\nu}(d)\right]/\Gamma(\nu)$

where $I_{\{0 \le d \le 1\}}(d)$ is the indicator function which is one when $0 \le d \le 1$ and zero otherwise, K_ν is a modified Bessel function of the second kind and ν is a smoothness parameter. The relationship between the smoothness of the realizations and the behavior of the correlation function near zero is given in (Stein 1999), along with further discussion of theoretical and practical properties of different choices of correlation functions. In the Bayesian paradigm, conjugate priors exist for the simpler GP models (Hjort and Omre 1994), and ▶Markov chain Monte Carlo can be used for the general case.

With the above formulation, the process interpolates the data. It is straightforward to include an additive random noise term in the model. An equivalent formulation arose in geostatistics, by thinking of including additional variability from a separate small-scale process, with the resulting term referred to as the nugget (Cressie 1993).

In practice, a convenient way of obtaining a Gaussian process is by convolving a white noise process with a smoothing kernel (Barry and Ver Hoef 1996; Higdon 2002). For locations $s \in \mathcal{S}$, let W be a Wiener process, and let $k(\cdot; \phi)$ be a kernel, possibly depending on a low dimensional parameter ϕ. Then we can obtain a Gaussian process by:

$$X(s) = \int_S k(u - s; \phi) dW(u). \quad (1)$$

The resulting covariance function depends only on the displacement vector $d_{s,s'} = s - s'$, for $s, s' \in S$, i.e.,

$$\begin{aligned} \mathrm{cov}(X(s), X(s')) &= \int_S k(u - s; \phi) k(u - s'; \phi) du \\ &= \int_S k(u - d_{s,s'}; \phi) k(u; \phi) du. \end{aligned}$$

Under suitable regularity conditions, there is a one to one relationship between the smoothing kernel k and the covariance function, based on the convolution theorem for Fourier transforms. A discrete approximation of Eq. 1 can be obtained by fixing a finite number of well-spaced background points, $s_1, \ldots, s_M$:

$$X(s) \approx \sum_{i=1}^{M} k(s_i - s; \phi) w(s_i),$$

where $w(s)$ is a white noise process.

The GP model can be extended to settings with a spatially-correlated multivariate response by employing methods such as cokriging (Wackernagel 1998). There are a variety of extensions for dealing with nonstationarity via approaches such as deformations (Sampson and Guttorp 1992), evolving convolutions (Higdon et al. 1999), or partitioning (Gramacy 2008).

About the Author

Professor Lee is a Fellow of the American Statistical Association and is currently Editor-in-Chief of the journal *Bayesian Analysis* and an Editor for *Chance*. He has previously been an Associate Editor for the *Journal of the*

American Statistical Association (2003–2009) and *Statistics and Computing* (2007–2009). He is the author of the book *Bayesian Nonparametrics via Neural Networks* (SIAM, 2004), and co-author of *Multiscale Modeling: A Bayesian Perspective* (with Marco A.R. Ferreira, Springer, 2007).

Cross References

►Brownian Motion and Diffusions
►Extremes of Gaussian Processes
►Stochastic Processes
►Stochastic Processes: Classification

References and Further Reading

Barry RP, Ver Hoef JM (1996) Blackbox Kriging: Spatial prediction without specifying variogram models. J Agric Biol Envir S 1: 297–322
Cox DR, Miller HD (1965) The theory of stochastic processes. Wiley, New York
Cressie NAC (1993) Statistics for spatial data, revised edition. Wiley, New York
Gramacy RB, Lee HKH (2008) Bayesian treed Gaussian process models with an application to computer modeling. J Am Stat Assoc 103:1119–1130
Higdon D (2002) Space and space-time modeling using process convolutions. In: Anderson C, Barnett V, Chatwin PC, El-Shaarawi AH (eds) Quantitative methods for current environmental issues. Springer, London, pp 37–56
Higdon DM, Swall J, Kern JC (1999) Non-stationary spatial modeling. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian statistics 6, Oxford University Press, Oxford, pp 761–768
Hjort NL, Omre H (1994) Topics in spatial statistics. Scand J Stat 21:289–357
Sampson PD, Guttorp P (1992) Nonparametric estimation of nonstationary spatial covariance structure. J Am Stat Assoc 87:108–119
Stein ML (1999) Interpolation of spatial data: some theory for Kriging. Springer, New York
Wackernagel H (1998) Multivariate geostatistics. Springer, Berlin

Gauss-Markov Theorem

John S. Chipman
Regents' Professor of Economics Emeritus
University of Minnesota, Minneapolis, MN, USA

The so-called Gauss-Markov theorem states that under certain conditions, ►least-squares estimators are "best linear unbiased estimators" ("BLUE"), "best" meaning having minimum variance in the class of unbiased linear estimators.

The linear regression model (see ►Linear Regression Models) $y_t = \sum_{j=1}^{k} x_{tj}\beta_j + \varepsilon_t$ $(t = 1, 2, \ldots, n)$ may be written as

$$y = X\beta + \varepsilon, \quad E\{\varepsilon\} = 0, \quad E\{\varepsilon\varepsilon'\} = \sigma^2 V \qquad (1)$$

(the prime denoting transposition), where y is an $n \times 1$ vector of observations on a variable of interest, X is an $n \times k$ matrix (generally assumed to have rank k) of n observations on $k < n$ explanatory variables x_{ij} $(i = 1, 2, \ldots, n;$ $j = 1, 2, \ldots, k)$, and ε is a vector of random errors; β is an unknown $k \times 1$ vector and $\sigma^2 > 0$ an unknown scalar; V (assumed known) is generally assumed to be positive-definite. (These rank assumptions will be relaxed below.) In the simplest case of independent observations, one takes $V = I$ (the identity matrix of order n). (Gauss [1823, §§16, 18, 35, 38] took V to be a diagonal matrix of weights, followed by Markov [1924, p. 325; 1912, p. 203]. Aitken [1935] generalized V to be a positive-definite matrix.) It is desired to obtain an estimator of β.

A (generalized) *least-squares* (GLS) estimate b of β is one that minimizes the (weighted) sum of squares $e'V^{-1}e = (y - Xb)'V^{-1}(y - Xb)$ of the residuals $e = y - Xb$. It is not hard to prove that this minimization is accomplished if and only if b is a solution of $X'V^{-1}Xb = X'V^{-1}y$ (the generalized "normal equations"). Such a solution always exists. (It is unique if and only if X has rank k.) This becomes an *ordinary least-squares* (OLS) estimate if $V = I$, and then the above become the ordinary normal equations.

A *Gauss-Markov (GM) estimator* of β in the model (1) is an affine estimator $\hat{\beta} = Ay + c$ which, among all affine estimators satisfying the condition of *unbiasedness*, namely

$$E\{\hat{\beta}\} = AX\beta + c = \beta \text{ for all } \beta, \quad \text{i.e., } AX = I \text{ and } c = 0, \quad (2)$$

has its variance $E\{A\varepsilon\varepsilon'A'\} = \sigma^2 AVA'$ minimized. Now, from the matrix Cauchy-Schwarz inequality (where $M \geq 0$ means that $x'Mx \geqq 0$ for all x) it follows that

$$AVA' \geq AX(X'V^{-1}X)^{-1}X'A', \quad \text{with equality} \iff$$
$$A = AX(X'V^{-1}X)^{-1}X'V^{-1}. \qquad (3)$$

Together with (3) this implies that the minimizing A is precisely $(X'V^{-1}X)^{-1}X'V^{-1}$, which provides the formula for the GM or GLS estimator $\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y = b$.

In the model (1), an unbiased affine estimator of β exists if and only if $\text{rank}(X) = k$. To generalize the above results to the case $\text{rank}(X) < k$, one may employ either a set of imposed linear restrictions on β (see Pringle and Rayner 1971, pp. 90–98), or else a concept due to Bose (1944) of a (linearly) *estimable functional* $\psi = (\psi_1, \psi_2, \ldots, \psi_k)$, defined by the condition that there exists a vector $a = (a_1, a_2, \ldots, a_n)$ such that $E\{ay\} = \psi\beta$ identically in β. Clearly this implies $\psi = aX$, i.e., that the set of linearly

estimable functionals coincides with the row space of X. This leads to the following result: *An affine estimator $\hat{\beta} = Ay + c$ furnishes an unbiased estimator of any estimable function $\psi\beta$ in the regression model (1) if and only if X satisfies the two restrictions $XAX = X$ and $Xc = 0$.* This generalizes condition (3). Likewise, the second condition of (3) generalizes to $XAV = (XAV)'$.

It may be noted that in the special case $V = I$ the latter two conditions are conditions (1) and (3) of Penrose (cf. Pringle and Rayner 1971, pp. 1–2) defining a unique "generalized inverse" $A = X^{\dagger}$ of X, the other two being $AXA = A$ and $AX = (AX)'$. A matrix A satisfying just $XAX = X$ is sometimes called a weak generalized inverse or "g-inverse" of X, denoted $A = X^{-}$. Rao (1971) proposed the development of a "unified theory of linear estimation" generalizing the Gauss-Markov theory to take care of deficient rank of either X or V, or both. Goldman and Zelen (1964, Theorem 4) had already proposed the use of $(X'V^{\dagger}X)^{\dagger}X'V^{\dagger}y$, and Mitra and Rao (1968, p. 286) had proved that a sufficient condition for the formula $(X'V^{-}X)^{-}X'V^{-}y$ to be valid is that the column space of X be contained in that of V, i.e., $\mathcal{C}(X) \subseteq \mathcal{C}(V)$. Independently, the same formula for the GM estimator was introduced by Zyskind and Martin (1969, Theorem 1). (As explained by Zyskind [1967, p. 1092], Zyskind and Martin [1969, pp. 1190–1191], and Pringle and Rayner [1971, pp. 93–98], singular variance matrices arise naturally in the ▶analysis of variance; they also arise in econometrics, where Basmann's (1957) "generalized classical linear (GCL) estimator" has the form $(Z'VZ)^{-1}Z'Vy$ where Z is $n \times v$ of rank v and $V = V^{-}$ is an idempotent matrix of rank r, requiring $r \geqq v$ for "identifiability"). Rao and Mitra (1971, p. 299) later obtained a complete characterization

$$\hat{\beta} = [X'(V + cXX')^{-}X]^{-}X'(V + cXX')^{-}y,\ c > 0, \text{ or } c = 0 \text{ provided } \mathcal{C}(X) \subseteq \mathcal{C}(V). \tag{4}$$

The result (4) thus provides the desired generalization of the GM theorem to the case of singular V. Important additional results were supplied by Mitra (1973) and Rao (1979).

It was pointed out by Rao (1972, p. 370; 1973, pp. 276–277) that the condition $\psi = aX$ is no longer necessary for estimability if V is singular, since there will be linear functions of y "which are zero with probability 1 that can be added to any estimator without altering its value but violating [the above] condition." Assuming V to be of rank $r \leqq n$ he obtained the needed general condition that there exist an $n \times (n-r)$ matrix N such that $N'y = 0$ (with probability 1) and a $1 \times (n-r)$ vector ρ, such that $\psi = (a - \rho N')X$.

Anderson (1948) showed that if both X and V are of full rank, and $P = [P_1, P_2]$ is an $n \times n$ orthogonal matrix diagonalizing V to $P'VP = \Lambda$, where P_1 is $n \times k$, and if $X = P_1K$ for some K, then $(X'V^{-1}X)^{-1}X'V^{-1}y = (X'X)^{-1}X'y$. This is the basic result underlying the ▶Durbin-Watson test. [Intuitively, the condition $X = P_1K$ states that in time-series analysis, the columns of X (for suitable choice of V, and of P_1 including a constant column) are low-frequency sinusoidal functions of time.] Magness and McGuire (1962) independently showed that the condition $X = P_1K$ is both necessary and sufficient for the GM estimator to be OLS. These results were generalized by Zyskind (1967) to the case in which $\text{rank}(X) \leqq k$ and $\text{rank}(V) \leqq n$. Zyskind also proved the necessity and sufficiency of the simpler condition that there exist a matrix W such that $VX = XW$, i.e., that $\mathcal{C}(VX) \subseteq \mathcal{C}(X)$ (see also Kruskal 1968). When X and V have full rank, the condition implies that $\mathcal{C}(VX) = \mathcal{C}(X)$. For important generalizations see Watson (1967, 1972) and Mitra and Moore (1973).

An interesting special case of Zyskind's result was established independently by McElroy (1967). If X has a column of ones (for the constant term), then the set of positive-definite Vs satisfying Zyskind's condition $VX = XW$ for some W is given by $V = \lambda[v_{ij}]$, where $v_{ii} = 1$ and $v_{ij} = \rho$ for $i \neq j$ and $-1/(n-1) < \rho < 1$.

Historical Notes

The terminology "Gauss-Markov theorem" appears to derive from Lehmann (1951). The theorem had previously been referred to by Neyman (1934, pp. 593–597) and David and Neyman (1938) as the "Markoff theorem on least squares," based on Chap. VII of Markov (1924). R. A. Fisher pointed out (Neyman 1934, p. 616) that "this was in essence the system of Gauss." Later, Neyman (1938, p. 130), who referred to Gauss (1887), said that the principle "was developed by Gauss, but not in a very clear way. It was developed and put into practical form by …Markoff …." Considerably later, Plackett (1949, p. 459) concluded that "it is implicit that [Gauss] is seeking unbiased estimates." Seal (1967, pp. 5–6) went further and stated this categorically, but without specific reference to Gauss. However, Sprott (1978, pp. 193–194) has claimed that the concept of unbiasedness of a linear estimator is not to be found in Gauss; rather, that Gauss (1823, §19) based himself on a concept of "error consistency," namely that $\lim_{\varepsilon \to 0} \hat{\beta} = \beta$. In his words (but in the present notation) Gauss said: (1855, p. 28): "if $k < n$, each unknown $\beta_1, \beta_2, \ldots$ can be expressed as a function of $y_1, y_2, \ldots$ in an infinity of ways, and these values will in general be different; they would coincide if, contrary to our hypothesis, the observations were perfectly exact." In the case of the affine estimator $\hat{\beta} = Ay + c = A(X\beta + \varepsilon) + c$,

if this equation is to hold "exactly" (without error, i.e., for $\varepsilon = 0$) this leads directly to the two conditions on the right in (3). Clearly these conditions are equivalent to those for unbiasedness on the left, as also shown by Sprott (p. 198). It is perhaps only in this sense that the latter condition can be said to be "implicit" in Gauss.

Gauss (1823, §20) went on to show (in the case of unit weights $V = I$), as Plackett (1949, p. 459) points out, that the diagonal elements of AA' (the variances of the $\hat{\beta}_i$ divided by σ^2) are minimized when $X'XA = X'$; this result is of course a special case of the matrix Cauchy-Schwarz inequality (3). Gauss (1823, §§37–38) chose $s^2 = \sum_{i=1}^{n} e_i^2/(n-k)$ as an estimator of σ^2 (for $V = I$), showing that he was well aware of the concept of unbiasedness, yet chose not to use it for the estimation of β.

Already in 1809 Gauss had considered maximum-likelihood estimation of β for the special case in which X is a column of ones, hence the least-squares estimator of β (now a scalar) is the sample mean. He concluded (Gauss 1809, §177; 1857, pp. 257–259; 1855, §3, pp. 117–179; 1887, pp. 100–102) that in order for this estimator to maximize the likelihood, the density function of ε_t would have to be of the form $f(\varepsilon_t) = \left(h/\sqrt{\pi}\right) e^{-h^2\varepsilon_t^2}$, where h is the "precision" (today denoted $1/\sigma\sqrt{2}$), i.e., would have to be normal ("Gaussian"); see also Stigler (1986, p. 141), Stewart (Gauss 1995, pp. 215, 220). However, he later expressed a strong preference for his 1823 distribution-free criterion. See Gauss (1995, p. 183), Eisenhart (1978, p. 382), Sheynin (1979, p. 46), Stigler (1986, pp. 140–143), Stewart (Gauss 1995, pp. 214–217), and an 1839 letter of Gauss to Bessel quoted by Markov (1951, p. 247) (Edgeworth's translation of which is reproduced in Plackett 1972, p. 247), describing his earlier ground as a "metaphysic."

We may also discuss Gauss's role in the development of the method of least squares itself. The first published treatment of this method was that of Legendre (1805, Appendix). In his letters, Gauss claimed to have developed and used the method of least squares before Legendre. The validity of this claim, questioned by Stigler (1986, pp. 145–146), has received support from Stewart (Gauss 1995, pp. 210–211) and Sheynin (1999). Plackett (1972, pp. 241–242) reports that Gauss had completed his *Theory of Motion*– his first work on least squares – in 1806, but "had difficulty in finding a publisher" unless he translated it into Latin. Stewart (Gauss 1995, p. 211n) explains that "owing to the conquest of the German states by Napoleon, Gauss's publisher required him to translate his manuscript, which was started in 1805 and finished in 1806, from the original into Latin" (see also Eisenhart 1978, p. 380). Thus Gauss's contribution of the method of least squares itself did not appear until 1809.

Markov opened his Chapter VII on least squares (1924, p. 323n; 1912, p. 201n) by stating that his view concerning the various attempts to justify the method had been set forth in an earlier work (Markov 1899; 1951, p. 246; Sheynin 2004; see also Sheynin 2006). There he had based his approach to estimation – "without assuming any definite law of probability for the errors" – on three propositions: 1. "we consider only such approximate equalities which …do not contain any constant error" – which may be interpreted as following Gauss's error-consistency principle; 2. "to each approximate equality we assign a certain weight" which is "inversely proportional to the expectations of the squares of the errors," i.e., their variance; and 3. "we determine such an equality whose weight is maximal." Then: "To my mind, only this justification of the method of least squares is rational; it was indicated by Gauss. I consider it rational mainly because it does not obscure the conjectural nature of the method. Keeping to this substantiation, we do not ascribe the ability of providing the most probable, or the most plausible results to the method of least squares and only consider it as a general procedure furnishing approximate values of the unknowns along with a hypothetical estimation of the results obtained." This diffident acceptance of the method contrasts with the enthusiastic advocacy of it by his follower Neyman.

Markov's contribution must of course be judged by what he actually did, not on what he set out to do. First he defines what he means by a "constant error" (1924, p. 324; 1912, p. 202). If a scalar parameter a is unknown, if x is a "possible outcome" of an observation on it, and if "approximate values" x_t of it are observed, with probabilities q_t, then the $a - x_t$ are called the errors of observation, and the *constant error* is defined as $E(a-x) = \sum_{t=1}^{n} q_i(a-x_i) = a - \sum_{t=1}^{n} q_t x_t$ $[= a - E(x)]$. Thus, "absence of constant error" (a frequent phrase of Markov's) is actually equivalent to unbiasedness. (Neyman [1938, p. 131] stated: "Markoff was not a statistician, he was a mathematician. What I shall say will be exactly equivalent to what he says but it will have a form more familiar to statisticians." Thus, Markov's [and Gauss's] error-consistency was translated by Neyman into unbiasedness.) Markov then introduces "actual observations" $a', a'', \ldots$ of a, generated by random variables $u', u'', \ldots$, and further introduces the symbol $\doteq$ (similar to the contemporary $\approx$, and not to be confused with the inequality sign $\neq$) such that $a \doteq a'$ means that the symbol on the right represents an observation which approximates the unknown quantity represented by the symbol on the left. It is evident from Markov's analysis that if u' is a random variable generating the observation a', then $a \doteq u'$ implies that $E(u') = a$ (cf. Markov 1924, p. 328; 1912, pp. 205–206).

Markov proceeds (1924, §§43–44, pp. 323–344; 1912, §37–38, pp. 201–218) to take up (in our notation) the special case of a single unknown parameter β, hence $k = 1$ in (1). From his subsequent analysis it becomes apparent that he implicitly assumes that $x_t = 1$ for $t = 1, 2. \ldots, n$, hence his model is that of a weighted sample mean, i.e., $y_t = \beta + \varepsilon_t$ where V in (1) is a diagonal matrix with diagonal elements v_{tt}. His "approximate equality" for the estimator $\hat{\beta}$ (a concept he does not use) is given by $\beta \approx \sum_{t=1}^{n} a_t y_t$; in his notation,

$$a \mp \lambda' a' + \lambda'' a'' + \ldots + \lambda^{(n)} a^{(n)} \tag{5}$$

(Markov 1924, p. 328; 1912, p. 205), where his λ's correspond to our a's and his a's to our y's, but his a to our β. At this point Markov specifically invokes unbiasedness when he states (1924, p. 328; 1912, p. 206) that since (5) must be "free of constant error" – which implies that the random variables $u', u'', \ldots$ corresponding to the observations $a', a'', \ldots$ all have expectation a, which he writes as $a \mp a', a \mp a'', \ldots$ (Markov 1924, p. 327; 1812, p. 205) – "the mathematical expectation of [the expression on the right in (5)] is equal to $(\lambda' + \lambda'' + \ldots + \lambda^{(n)})a$ for arbitrary choice of [the λ's]," hence (1924, p. 328; 1912, p. 206) the sum of the λ's must be equal to 1. In the notation of (3), the condition $AX = I$ now becomes $aX = 1$, or $\sum_{t=1}^{n} a_t = 1$. This covers proposition 1 above.

Now we consider propositions 2 and 3. Referring to the line below (3) above, it is desired to minimize the variance $E\{a\varepsilon\varepsilon' a'\} = \sigma^2 aVa' = \sigma^2 \sum_{t=1}^{n} a_t^2 v_{tt}$. Markov (1924, p. 330; 1912, p. 207) denotes this by $k \sum_{t=1}^{n} \lambda^{(t)} \lambda^{(t)} / p^{(t)} = k/P$, where $k = \sigma^2$ and $P = 1/\sum_{t=1}^{n} a_t^2 v_{tt}$ is the "weight" to be maximized, subject to the unbiasedness condition $\sum_{t=1}^{n} a_t = 1$ (1924, p. 331; 1912, p. 208). He shows there that this maximum is attained when $a_i/a_j = v_{ii}/v_{jj}$. Thus, the desired estimator is $\hat{\beta} = \sum_{t=1}^{n} (y_t/v_{tt}) \big/ \sum_{t=1}^{n} (1/v_{tt})$ (formula (21)) [his notation replacing "$\hat{\beta} =$" by "$\beta \mp$"], and the weight reduces to $P = \sum_{t=1}^{n} 1/v_{tt}$ (formula (22)).

Markov also treats the general case (1924, §§46–47; 1912, §§39–40). The unbiasedness conditions $E\left(\sum_{t=1}^{n} a_{it} y_t\right) = \beta_i$ are displayed explicitly (1924, p. 376; 1912, p. 220), from which he derives the above result (2) in his formula $(*)$ (1924, pp. 376–377; 1912, p. 221). To treat minimum variance, he departs from Gauss's method and employs k Lagrangean multipliers μ_j, using the transformation $X^* = V^{-1/2} X$ to handle the variances of the y_ts (1924, p. 380; 1912, p. 223 (formula [A]), so as to reach his result $(* * *)$.

As we have seen above, while Gauss formulated the least-squares problem in terms of diagonal V (and he presumably used weighted least squares in his astronomical calculations) his theorems were limited to the case $V = I$ of independent observations. Plackett (1949, p. 460) famously stated that "Markoff, who refers to Gauss's work, may perhaps have clarified assumptions implicit there but proved nothing new" – an opinion that has been accepted as authoritative. However, it is clear that Markov extended Gauss's theorems to diagonal V (dependent observations). And while not abandoning Gauss's error consistency, he introduced unbiasedness explicitly when needed to obtain condition (3). And Neyman played a major role in formulating best unbiasedness as a principle of estimation.

About the Author

Professor Chipman, "*the eminence grise of Econometrics*," was born on June 28, 1926, in Montreal, Canada. He was Associate Editor of *Econometrica*; (1956–1969), Co-editor of *Journal of International Economics* (1971–1976) and Editor (1977–1987); and Associate Editor of *Canadian Journal of Statistics* (1980–1983). Among his many awards and professional honors, Professor Chipman was elected a Fellow, American Statistical Association (1974), American Academy of Arts and Sciences (1979), Member, National Academy of Sciences (1993) and Chair of Section 54 (Economic Sciences), (1997–2000), Member, International Statistical Institute (1994), and Distinguished Fellow, American Economic Association (1999). He has received three honorary doctorates. Professor Chipman received the Humboldt Research Award for Senior U.S. Scientists in 1992–1995 and 2003. He has supervised 32 Ph.D. students. The importance of his work and achievements, was also recognized in 1999, in the volume: *Trade, Theory and Econometrics: Essays in Honor of John S. Chipman* (Eds. J.R. Melvin, J.C. Moore, and R. Riezman, Routledge), with contributions of many distinguished economists, including Paul A. Samuelson ((May 15, 1915 – December 13, 2009) the first American to win the Nobel Prize in Economics). His recent publications include following books: *The Theory of International Trade* (Edward Elgar Pub, two volumes, 2008–2009) and *Introduction To Advanced Econometric Theory* (Routledge 2011).

Cross References

►Best Linear Unbiased Estimation in Linear Models
►General Linear Models
►Least Squares
►Linear Regression Models
►Simple Linear Regression

References and Further Reading

Aitken AC (1935) On least squares and linear combinations of obervations. Proc R Soc Edinburgh A 55:42–48

Anderson TW (1948) On the theory of testing serial correlation. Skandinavisk Aktuarietidskrift 31:88–116

Basmann RL (1957) A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica 25(January):77–83

Bose RC (1944) The fundamental theorem of linear estimation. In Proceedings of the Thirty-First Indian Science Congress, Part III, 4–5

David FN, Neyman J (1938) An extension of the Markoff theorem on least squares. In: Neyman J, Pearson ES (eds) Statistical research memoirs, vol II. University of London, London, pp 105–116

Eisenhart C (1978) Gauss CF, Carl Friedrich. In: International encyclopedia of statistics, vol 1. Free Press, New York, pp 378–386

Gauss CF (1809, 1857) Theoria Motus Corporum Coelestium in Sectionibus Conicis solem Ambientum. Hamburg: Perthes und Besser. (Last page of Art. 177 reproduced in Stigler 1986, p. 142.) English translation by Charles Henry Davis, Theory of the motion of the heavenly bodies moving about the sun in conic sections. Little, Brown, Boston

Gauss CF (1823, 1826, 1995) Theoria Combinationis Observationum Erroribus Minimis Obnoxiæ: Pars Prior, Pars Posterior; Supplementum, Göttingen: Dieterische Universitäts-Druckerei (even pp 2–202), and Theory of the combination of observations least subject to errors: Part One, Part Two; Supplement, trans. Stewart GW (odd pp 3–203), followed by the three Anzeigen (Notices) to the above, pp 174–203, and an Afterward by Stewart GW, pp 205–241. Philadelphia: Society for Industrial and Applied Mathematics. (Many authors ascribe to the Pars Prior the date 1821 when the memoir was presented to the Royal Society of Göttingen.)

Gauss CF (1885) Méthode des moindres carrés. Mémoires sur la combination des observations (French translation by Joseph Bertrand of Gauss 1823, 1826, and extracts (§§175–186) from Gauss (1809) and other works.) Paris: Mallet-Bachelier. English translation by Hale F. Trotter, Gauss's work (1803–1826) on the theory of least squares (1957) Technical Report No. 5 from Department of Army Project No. 5B9901-01-004, Statistical Techniques Research Group, Princeton University, Princeton, NJ

Gauss CF (1887) Abhandlungen zur methode der kleinsten quadrate. Edited in German by Börsch A, Simon P (German translation of Gauss 1823, §§122–189 of Gauss 1809, and extracts from other works of Gauss.) Druck und Verlag von Stankiewicz' Buchdruckerei, Berlin. Reprinted, Physica, Würzburg, 1964

Goldman AJ, Zelen M (1964) Weak generalized inverses and minimum variance linear unbiased estimation. J Res Natl Bureau Std B 68B(October–December):151–172

Kruskal W (1968) When are Gauss-Markov and least squares estimators identical? A coordinate-free approach. Ann Math Stat 39(February):70–75

Legendre AM (1805) Nouvelles méthodes pour la détermination des orbites des comètes. Paris: Firmin Didot. (Appendix, pp 72–80, pp 72–75 reproduced in Stigler 1986, p 58.)

Lehmann EL (1951) A general concept of unbiasedness. Ann Math Stat 22(December):587–592

Magness TA, McGuire JB (1962) Comparison of least squares and minimum variance estimates of regression parameters. Ann Math Stat 33(June):462–470

Markov AA (1899, 1951) Zakon bol'shik chisel i sposob naimen'shikh kvadratov ("The Law of Large Numbers and the Method of Least Squares"). In Izbrannye trudy (Selected Works), 231–251. English translation by Oscar Sheynin, Russian papers on the history of probability and statistics, Berlin 2004, www.sheynin.de

Markov AA (1900, 1908, 1913, 1924, 1912) *Ischislenie Veroiatnostei (Calculus of Probability)*. St. Petersburg; Gosudarstvennoe Izdatel'stvo (State Publisher), Moscow. German translation by Heinrich Liebmann of the second edition, Wahrscheinlichkeitsrechnung. Leipzig and Berlin: B. G. Teubner. (A promised English translation of the latter [cf. Neyman 1938, p 131, David and Neyman 1938, p 116] apparently never appeared.)

McElroy FW (1967) A necessary and sufficient condition that least squares estimators be best linear unbiased. J Am Stat Assoc 62(December):1302–1304

Mitra SK (1973) Unified least squares approach to linear estimation in a general Gauss-Markov model. SIAM J Appl Math 25(December):671–680

Mitra SK, Moore BJ (1973) Gauss-Markov estimation with an incorrect dispersion matrix. Sankhyā A 35(June):139–152

Mitra SK, Rao CR (1968) Some results in estimation and tests of linear hypotheses under the Gauss-Markoff model. Sankhyā A 30(September):281–290

Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J R Stat Soc 97:558–606 [Note II: The Markoff Method and the Markoff Theorem on Least Squares, pp 593–597.] Discussion, pp 607–625

Neyman J (1938, 1952) Lectures and conferences on mathematical statistics, 2nd edn. Lectures and conferences on mathematical statistics and probability. The Graduate School of the U.S. Department of Agriculture, Washington, DC

Plackett RL (1949) A historical note on the method of least squares. Biometrika 36(December):458–460

Plackett RL (1972) Studies in the history of probability and statistics. XXIX. The discovery of the method of least squares. Biometrika 59(August):239–251

Pringle RM, Rayner AA (1971) Generalized inverse matrices, with applications to statistics. Charles Griffin, London

Rao CR (1971) Unified theory of linear estimation. *Sankhyā* A 33(December):371–394

Rao CR (1972) Some recent results in linear estimation. Sankhyā B 34(December):369–378

Rao CR (1973) Representations of best linear unbiased estimators in the Gauss-Markoff model with singular dispersion matrix. J Multivariate Anal 3(September):276–292

Rao CR (1979) Estimation of parameters in the singular Gauss-Markoff model. Commun Stat A Theor Meth 8(14):1353–1358

Rao CR, Mitra SK (1971) Further contributions to the theory of generalized inverses of matrices and applications. Sankhyā A 33(September):289–301

Seal HL (1967) Studies in the history of probability and statistics. XV. The historical development of the Gauss linear model. Biometrika 54(June):1–24

Sheynin O (1979) C.F. Gauss and the theory of errors. Arch Hist Exact Sci 20(March):21–72

Sheynin O (1999) The discovery of the principle of least squares. Historia Scientiarum. Int J Hist Sci Soc Japan, 2nd series, 8(March):249–264

Sheynin O (2006) Markov's work on the treatment of observations. Historia Scientiarum, 2nd series, 16(July):80–95

Sprott DA (1978) Gauss's contributions to statistics. Historia Mathematica 5(May):183–203
Stigler SM (1986) The history of statistics. The Belknap Press of Harvard University Press, Cambridge, MA
Watson GS (1967) Linear least squares regression. Annals of mathematical statistics 38(December):1679–1699
Watson GS (1972) Prediction and the efficiency of least squares. Biometrika 59(April):91–98
Zyskind G (1967) On canonical forms, non-negative covariance matrices and best and simple least squares estimators in linear models. Ann Math Stat 38(August):1092–1109
Zyskind G, Martin FB (1969) On best linear estimation and a general Gauss-Markov theorem in linear models with arbitrary nonnegative covariance structure. SIAM J Appl Math 17(November):1190–1202

General Linear Models

Yuehua Wu
Professor
York University, Toronto, ON, Canada

Let y be a random variable such that $E(y) = \mu$, or $y = \mu + \epsilon$, where ϵ is a random error with $E(\epsilon) = 0$. Suppose that $\mu = x_1\beta_1 + \cdots + x_p\beta_p$, where $x_1, \ldots, x_p$ are p variables and $\beta_1, \ldots, \beta_p$ are p unknown parameters. The model

$$y = x_1\beta_1 + \cdots + x_p\beta_p + \epsilon, \tag{1}$$

is the well-known multiple linear regression model (see ►Linear Regression Models). Here y is called dependent (or response) variable, $x_1, \ldots, x_p$ are called independent (or explanatory) variables or regressors, and $\beta_1, \ldots, \beta_p$ are called regression coefficients. Letting

$$(y_1, x_{11}, \ldots, x_{1p}), \ldots, (y_n, x_{n1}, \ldots, x_{np})$$

be a sequence of the observations of Y and $x_1, \ldots, x_p$, we have

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \epsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where $\epsilon_1, \ldots, \epsilon_n$ are the corresponding random errors. Denote $\boldsymbol{y}_n = (y_1, \ldots, y_n)^T$, $X_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ with $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})^T$ for $j = 1, \ldots, p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, and $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)^T$, where $\boldsymbol{a}^T$ denotes the transpose of a vector $\boldsymbol{a}$. (2) can therefore be expressed as

$$\boldsymbol{y}_n = X_n\boldsymbol{\beta} + \boldsymbol{\epsilon}_n. \tag{3}$$

For convenience, X_n is assumed to be nonrandom with rank p through this article (see Rao (1973) for the case that the rank of X_n is less than p). It is noted that if $x_1, \ldots, x_p$ in (1) are random variables, it is usually assumed that $E(y|x_1, \ldots, x_p) = x_1\beta_1 + \cdots + x_p\beta_p$, which replaces the $E(y)$ given above.

A well known method for estimating $\boldsymbol{\beta}$ in (3) is *Least Squares*. Its theoretical foundation was laid by Gauss (1809) and Markov (1900) among others. Assuming that $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated with zero means and constant variance σ^2, the least squares estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}$ can be obtained by minimizing $(\boldsymbol{y}_n - X_n\boldsymbol{b})^T(\boldsymbol{y}_n - X_n\boldsymbol{b})$ among all possible $\boldsymbol{b} \in R^p$. It can be shown that $\hat{\boldsymbol{\beta}}_n = \left(X_n^T X_n\right)^{-1} X_n^T \boldsymbol{y}_n$. It is easy to verify that $E(\hat{\boldsymbol{\beta}}_n) = \boldsymbol{\beta}$ so that $\hat{\boldsymbol{\beta}}_n$ is an unbiased estimator of $\boldsymbol{\beta}$. Define the residual as $\boldsymbol{r}_n = \boldsymbol{y}_n - \hat{\boldsymbol{y}}_n$ with $\hat{\boldsymbol{y}}_n = X_n\hat{\boldsymbol{\beta}}_n$, the fitted (or predicted) vector. Since $E\left(\boldsymbol{r}_n^T\boldsymbol{r}_n\right) = (n-p)\sigma^2$, an unbiased estimator of σ^2 is given by $\hat{\sigma}_n^2 = \boldsymbol{r}_n^T\boldsymbol{r}_n/(n-p)$. As shown in Rao (1973) or Seber and Lee (2003), $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}_n$ is the best linear unbiased estimate of $\boldsymbol{c}^T\boldsymbol{\beta}$ for any constant vector $\boldsymbol{c}$. If in addition $\epsilon_1, \ldots, \epsilon_n$ are independent with common third and fourth moments, $\hat{\sigma}_n^2$ can be shown to be the unique nonnegative quadratic unbiased estimator of σ^2 with minimum variance (Seber and Lee (2003)). If we further assume that $\epsilon_1, \ldots, \epsilon_n$ are independently and identically $N(0, \sigma^2)$ distributed, it can be shown that $\hat{\boldsymbol{\beta}}_n$ is actually the maximum likelihood estimator of $\boldsymbol{\beta}$. It can also be proved that $\hat{\boldsymbol{\beta}}_n$ is distributed as $N\left(\boldsymbol{\beta}, \sigma^2\left(X_n^T X_n\right)^{-1}\right)$, $(n-p)\hat{\sigma}_n^2$ is distributed as $\sigma^2\chi^2_{n-p}$, and, in additions, $\hat{\boldsymbol{\beta}}_n$ is independent of $\hat{\sigma}_n^2$. However the maximum likelihood estimator of σ^2 is $(n-p)\hat{\sigma}^2/n$ instead of $\hat{\sigma}^2$, which is biased for estimating σ^2 but is asymptotically equal to $\hat{\sigma}^2$ as $n \to \infty$ for a fixed p. It is noted that if $X_n^T X_n$ is nearly singular, a ridge estimator may be used to estimate $\boldsymbol{\beta}$. Although ridge estimators are biased but a properly chosen one will have a smaller mean squared error than the least squares estimator in such a case.

We now consider how to test the hypothesis $H_0 : \Gamma\boldsymbol{\beta} = \boldsymbol{c}$ with a known $q \times p$ matrix Γ of rank q and a known $q \times 1$ vector $\boldsymbol{c}$ for the model (3). Assume that $\epsilon_1, \ldots, \epsilon_n$ are independently and identically $N(0, \sigma^2)$ distributed. To test H_0, an F-test can be employed with the test statistic

$$F = \left((\Gamma\hat{\boldsymbol{\beta}}_n - \boldsymbol{c})^T\left[\Gamma\left(X_n^T X_n\right)^{-1}\Gamma^T\right]^{-1}(\Gamma\hat{\boldsymbol{\beta}}_n - \boldsymbol{c})/q\right)/\hat{\sigma}_n^2,$$

which is actually equivalent to the likelihood ratio test statistic (see, e.g., Seber and Lee (2003) and Rao et al. (2008)). It can be shown that $F \sim F_{q,n-p}$ under H_0, where $F_{q,n-p}$ denotes the central F distribution with degrees of freedom q and $n-p$, respectively. For a level of significance α, H_0 is rejected if $F > F_{q,n-p;1-\alpha}$. Here $F_{q,n-p;1-\alpha}$ denotes

the $(1-\alpha)$-quantile of $F_{q,n-p}$. For discussion on more general testing of hypotheses, see Seber and Lee (2003) and Rao et al. (2008) among others, which also explore how to construct confidence regions for parameters. In addition, a number of practical examples can be found in Draper and Smith (1998).

So far it has been assumed that the random errors have constant variance σ^2. We now assume that $E(\boldsymbol{\epsilon}_n) = 0$, $cov(\boldsymbol{\epsilon}_n) = \sigma^2 V$ and V is a known $n \times n$ positive definite matrix. To estimate $\boldsymbol{\beta}$ in (3), the generalized least squares method is used instead and the resulting estimator is $\tilde{\boldsymbol{\beta}}_n = \left(X_n^T V^{-1} X_n\right)^{-1} X_n^T V^{-1} \boldsymbol{y}_n$, which is the same as the least squares estimator of $\boldsymbol{\beta}$ for the model $\tilde{\boldsymbol{y}}_n = \tilde{X}_n \boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}_n$ with $\tilde{\boldsymbol{y}}_n = V^{-1/2} \boldsymbol{y}_n$, $\tilde{X}_n = V^{-1/2} X_n$ and $\tilde{\boldsymbol{\epsilon}}_n = V^{-1/2} \boldsymbol{\epsilon}_n$. In this case, it can be shown that an unbiased estimator of σ^2 is given by $\tilde{\sigma}_n^2 = (\boldsymbol{y}_n - X_n \tilde{\boldsymbol{\beta}}_n)^T V^{-1} (\boldsymbol{y}_n - X_n \tilde{\boldsymbol{\beta}}_n)/(n-p)$. If we further assume that $\boldsymbol{\epsilon}_n \sim N(0, \sigma^2 V)$, it can be verified that $\tilde{\boldsymbol{\beta}}_n$ and $(n-p)\tilde{\sigma}_n^2/n$ are the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2, respectively. The hypothesis testing and interval estimation of parameters can be carried out similar to those previous described.

The model (3) can be generalized to

$$Y_n G_n = X_n B H_n + Z_n \Theta Q_n + E_n, \tag{4}$$

where Y_n is an $n \times m$ random matrix, G_n is an $m \times k$ matrix, X_n, H_n, Z_n and Q_n are respectively $n \times p$, $q \times k$, $n \times u$ and $v \times k$ matrices, B is a $p \times q$ matrix of unknown parameters, Θ is a $u \times v$ random matrix, and E_n is an $n \times k$ matrix of random errors with zero means. This model is linear in parameters and is fairly general. The following models are its special cases: multiple ▶linear regression models, linear random effects models, ▶linear mixed models, analysis of variance (ANOVA) models (see ▶Analysis of Variance), multivariate analysis of variance (MANOVA) models (see ▶Multivariate Analysis of Variance (MANOVA)), ▶analysis of covariance (ANCOVA) models, multivariate analysis of covariance (MANCOVA) models, response surface regression models and growth curve models. Thus this model is named as a general linear model. It is noted that the model (4) may be extended to allow for change points to occur. Such an example can be found in Seber and Lee (2003) among others.

For simplicity, we limit our attention to the case that G_n^{-1} exists. Then (4) can be written as

$$\begin{aligned} Y_n &= X_n B H_n G_n^{-1} + Z_n \Theta Q_n G_n^{-1} + E_n G_n^{-1} \\ &= X_n B \widetilde{H}_n + Z_n \Theta \widetilde{Q}_n + \widetilde{E}_n \end{aligned}$$

with $\widetilde{H}_n = H_n G_n^{-1}$, $\widetilde{Q}_n = Q_n G_n^{-1}$, and $\widetilde{E}_n = E_n G_n^{-1}$. $Y_n = X_n B \widetilde{H}_n + Z_n \Theta \widetilde{Q}_n + \widetilde{E}_n$ is also a commonly used expression for a general linear model. Denote the Kronecker product of matrices A_1 and A_2 by $A_1 \otimes A_2$ and define a $\tau\varsigma$-dimensional vector $vec(A_3)$ of a $\tau \times \varsigma$ matrix A_3 by stacking its column vectors. Then the general linear model (4) can be rewritten as

$$\begin{aligned} \left(I_n \otimes G_n^T\right) vec\left(Y_n^T\right) = &\left(X_n \otimes H_n^T\right) vec\left(B^T\right) \\ &+ \left(Z_n \otimes Q_n^T\right) vec\left(\Theta^T\right) \\ &+ vec\left(E_n^T\right), \end{aligned} \tag{5}$$

or

$$\begin{aligned} \left(G_n^T \otimes I_n\right) vec\left(Y_n\right) = &\left(H_n^T \otimes X_n\right) vec\left(B\right) \\ &+ \left(Q_n^T \otimes Z_n\right) vec\left(\Theta\right) \\ &+ vec\left(E_n\right). \end{aligned} \tag{6}$$

Denote $E_n = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)^T$. If $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ are independently and identically distributed with zero mean vector and covariance matrix Σ, it can be shown that $vec\left(E_n^T\right)$ has zero mean vector and covariance matrix $I_n \otimes \Sigma$, and hence $vec\left(E_n\right)$ has zero mean vector and covariance matrix $\Sigma \otimes I_n$, where I_n is an $n \times n$ identity matrix.

Least squares method (see ▶Least Squares) may be used to estimate the parameters in (4) directly in terms of (5) or (6). For example, if $Y_n = X_n B H_n + E_n$ and $n \geq m + p$, the least squares method can be used to estimate B. Note that the estimation may not be unique if the rank of H_n is not q. A least squares estimator of B is given by $\widehat{vecB}_n = \left[\left(H_n \Sigma^{-1} H_n^T\right)^+ \otimes \left(X_n^T X_n\right)^{-1}\right]\left(H_n \Sigma^{-1} \otimes X_n^T\right) vec\left(Y_n\right)$, where A^+ denotes the Moore-Penrose inverse of a matrix A. If in addition $q = k = m$ and $H_n = I_m$, it follows that $\widehat{vecB}_n = \left[I_n \otimes \left(X_n^T X_n\right)^{-1} X_n^T\right] vec\left(Y_n\right)$, i.e., $\hat{B}_n = \left(X_n^T X_n\right)^{-1} X_n^T Y_n$. The residual is defined as $\Upsilon_n = Y_n - X_n \hat{B}_n$. If $vec\left(E_n\right)$ has zero mean vector and covariance matrix $\Sigma \otimes I_n$, it can be shown that $\hat{B}_n$ and $\hat{\Sigma}_n = \Upsilon_n^T \Upsilon_n/(n-p)$ are unbiased estimators of B and Σ, respectively. Now assume that $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ are independently and identically $N(0, \Sigma)$ distributed. Then it can be proved that $\hat{B}_n$ is the maximum likelihood estimator of B with the distribution $N\left(B, \left(X_n^T X_n\right)^{-1} \otimes \Sigma\right)$, $(n-p)\hat{\Sigma}_n$ is $W_m(n-p, \Sigma)$ distributed, and, moreover, $\hat{B}_n$ is independent of $\hat{\Sigma}_n$, where $W_k(\ell, \Psi)$ denotes the Wishart distribution with ℓ degrees of freedom and $k \times k$ covariance matrix Ψ. Related references include Rao (1973), Mardia et al. (1979), Muirhead (1982) and Fang and Zhang (1990).

The last two references also considered the case that the random errors are elliptically distributed.

When the random errors are not normally distributed but the error distributions are known up to some distribution parameters, maximum likelihood methods may be employed for estimating unknown parameters. If one suspects that there may be some departures from the underlying assumptions on error distributions or there may be some ►outliers in the data, one may consider to use weighted least squares method, M-estimation method, or other robust estimation methods to estimate the parameters (see Seber and Lee (2003) and Rao et al. (2008) among others).

In addition to those mentioned above, some commonly used general linear model techniques and tools include:

- Shrinkage estimation, heterogeneous linear estimation, Stein-rule estimation, multiway classification.
- Bonferroni simultaneous confidence intervals, Turkey's simultaneous comparisons, Scheffé multiple comparisons.
- Multivariate analysis of variance table, union-intersection approach, t-test, F-test, likelihood ratio test, Wilks's lambda statistic, Hotelling's T^2 test, Pillai's trace, Hotelling-Lawley trace, Roy's greatest root, sphericity test, invariant test, profile analysis, goodness-of-fit test.
- Classical prediction, optimal heterogeneous prediction, optimal homogeneous prediction, prediction regions, Stein-rule predictor, partial least squares, principal components regression, Kalman filter.
- Multiple correlation coefficient, partial correlation coefficient, Q-Q plot, high-leverage point, studentized residual, Cook's distance, variance ratio, analysis of residuals, partial regression plots, variable selection procedures including forward selection, backward elimination, stepwise regression, cross-validation, Mallows' C_p, information theoretic criteria, LASSO, SCAD.
- Imputation, Yate's procedure, Bartlett's ANCOVA.
- Canonical form, data centering and scaling, Cholesky decomposition method, orthogonal-triangular decomposition, recursive algorithm.
- Generalized variance, total least squares, orthogonal design matrix, hierarchical design, Bayesian method.

The problems associated with the general linear models include estimability of B and/or Θ, testability of hypotheses on B and/or Θ, estimation of B and/or Θ under some linear constraints, hypothesis testing of hypotheses on B and/or Θ under some linear constraints, construction of confidence regions, simultaneous testing, multiple comparisons, homogeneity of covariance matrices, collinearity, estimation of covariance, asymptotics, variable selection, high-dimensionality problem (e.g., $n << p$ in (2)), underfitting, overfitting, optimal design, censored data, missing observations, outliers, model assumptions, transformation, outside sample predictions, and efficiency.

Standard software pacakages including SAS, R, S-PLUS, SPSS and MATLAB may be used to solve statistical inference problems for general linear models.

About the Author

Yuehua (Amy) Wu received her doctorate in Statistics from University of Pittsburgh in 1989 (her advisor was Professor C.R. Rao). She is Professor of Mathematics and Statistics at York University and recently was Director of the Statistics section in the department of Mathematics and Statistics.

Cross References

►Analysis of Covariance
►Analysis of Variance
►Best Linear Unbiased Estimation in Linear Models
►Least Squares
►Linear Mixed Models
►Linear Regression Models
►Multivariate Analysis of Variance (MANOVA)
►Statistics: An Overview
►Student's *t*-Tests

References and Further Reading

Draper NR, Smith H (1998) Applied regression analysis, 3rd edn. Wiley, New York
Fang KT, Zhang YT (1990) Generalized multivariate analysis. Springer/Science Press, Berlin/Beijing
Gauss CF (1809) Least squares. Werke 4:1–93, Göttingen
Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis. Academic, London
Markov AA (1900) Ischislenie veroyatnostej, SPb
Muirhead RJ (1982) Aspects Of multivariate statistical theory. Wiley, New York
Rao CR (1973) Linear statistical inference and applications. Wiley, New York

Rao CR, Toutenburg H, Shalabh C, Heumann C (2008) Linear models and generalizations. Least squares and alternatives. 3rd edn. Springer, Berlin/Heidelberg

Seber GAF, Lee AJ (2003) Linear regression analysis. 2nd edn. Wiley, New York

Generalized Extreme Value Family of Probability Distributions

Chris P. Tsokos
Distinguished University Professor
University of South Florida, Tampa, FL, USA

In real life phenomenon we experience largest observations (maximum extreme) or smallest observation (minimum extreme), such as "How tall should one design an embankment so that the sea reaches this level only once in 100 years?"; "What is the lowest value the Dow Jones Industrial Average can reach in the next three years?"; "How high a drug concentration in the bloodstream can go before causing toxicity?", among others.

To characterize and understand the behavior of these extremes, we usually use probabilistic extreme value theory. Such theory deals with the stochastic behavior of the minimum and maximum of independent identically distributed random variables. Here, we shall give a brief introduction of the generalized extreme value family of probability distribution that fits the subject observations. This family of probability distribution function (pdf) consists of three famous classical pdfs, namely the Gumbel, the Frechet, and the Weibull.

Extreme value theory has been successfully applied to address problems in floods, wind gusts, hurricanes, reliability, earthquakes, stock market crashes, survival analysis, rain fall, health sciences, drug evaluation, financial systems, among others. Some brief useful references on the theory are Abdelhafez and Thomas (1990), Achcar (1991), Ahsanullah (1990), Campell and Tsokos (1973a), Cheng et al. (1998), Cohen (1986), Daniels (1942), Davidovich (1992), De Haan (1970), Engelhardt and Bain (1973), Engelund and Rackwitz (1992), Frechet (1927), Galambos (1981), Galambos (1987), Gumbel (1935), Gumbel (1958), Gumbel (1962a, b, c), Gumbel and Goldstein (1964), Hassanein (1972), Hosking (1985), Jenkinson (1969), Mann et al. (1973), von Mises (1923), von Mises (1936), Pickands (1986), Pickands (1981), and Weibull (1939a), and applications Ahmad et al. (1988), Aitkin and Clayton (1980), Al-Abbasi and Fahmi (1991), Azuz (1955), Barnett (1990), Beran et al. (1986), Broussard and Booth (1998), Buishand (1985), Buishand (1989), Campbell and Tsokos (1973a, b) Changery (1982), Chowbury et al. (1991), Coles and Pan (1996), Coles and Tawn (1994), Coles and Tawn (1996), De Hann and Resnick (1998), Diebold et al. (1999), Eldredge (1957), Embrechts et al. (1997), Epstein (1948), Fahmi and Al-Abbasi (1991), Frenkel and Kontorova (1943), Fuller (1914), Goka (1993), Greenwood (1946), Greis and Wood (1981), Gumbel (1941), Gumbel (1945), Gumbel (1949), Harris (1970), Henery (1984), Hisel (1994), Hosking and Wallis (1988), Jain and Singh (1987), Joe (1994), Kimball (1955), Longuet-Higgins (1952), Marshall (1983), Nisan (1988), Nordquist (1945), Okubo and Narita (1980), and Weibull (1939b).

Extreme Value Theory (EVT) is the study of probabilistic extremes and focuses primarily on the asymptotic behavior as the sample size approaches infinity. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables having a common cumulative probability distribution F. The model focuses on the statistical behavior of

$$M_n = \max\{X_1, X_2, \ldots, X_n\},$$

where X_i usually represent values of a process measured on a regular time scale. For example, hourly measurements of stock prices or plasma drug concentration over a certain period; so that M_n represents the maximum of the process over n time units of observation. If n is the number of observations in a day, then M_n corresponds to the daily maximum.

The probability distribution of M_n can be derived exactly for all values of n, that is,

$$\begin{aligned}\Pr\{M_n \le z\} &= \Pr\{X_1 \le z, \ldots, X_n \le z\} \\ &= \Pr\{X_1 \le z\} \ldots \Pr\{X_n \le z\} = F^n(z) \quad (1)\end{aligned}$$

The difficulty that arises in practice is the fact that the cumulative probability distribution F is unknown. One possibility is to use standard statistical techniques to estimate F from observed data, and then substitute this estimate into (1). But very small discrepancies in the estimate of F can lead to substantial discrepancies for F^n. This leads to an approach based on asymptotic argument which requires determining what possible limit probability distributions are possible for M_n as $n \to \infty$. The question then is "what are the possible limit distributions in the extreme case?"

We study the limiting probability distributions of $\frac{M_n - b_n}{a_n}$ where a_n and b_n are sequences of normalizing coefficients such that $F^n\left(\frac{M_n - b_n}{a_n}\right)$ leads to a non-degenerate probability distribution as $n \to \infty$. Specifically, we seek $\{a_n > 0\}$ and $\{b_n\}$ such that $\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \to G(z)$ where $G(z)$ does not depend on n.

Extremal Types Theorem: If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that, as $n \to \infty$, $\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \to G(z)$ where G is a non-generate probability distribution function, then G belongs to one of the following families:

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, -\infty < z < +\infty \quad (2)$$

$$G(z) = \begin{cases} 0, z < b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\partial}\right\}, \quad z \ge b \end{cases} \quad (3)$$

$$G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\partial}\right]\right\}, \quad z \le b \\ 1, z > b \end{cases} \quad (4)$$

For parameters $a(\text{scale}) > 0, b(\text{location})$ and, in (3) and (4), $\partial(\text{shape}) > 0$.

Combining the three classes of probability distributions results in what we refer to as the Generalized Extreme Value (GEV) family of probability distribution. The GEV cumulative probability distribution can be written as

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$
$$1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0, -\infty < \xi < +\infty, \sigma > 0 \quad (5)$$

The equation above is the generalized extreme value family of distributions. This was obtained independently by von Mises (1923, 1936), and Jenkinson (1969) and Hosking (1985). Equation 5 can also be written as

$$G(z) = \begin{cases} \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, & -\infty < z \le \mu - \frac{\sigma}{\xi}, \xi < 0; \\ & \mu - \frac{\sigma}{\xi} \le z < +\infty, \xi > 0; \\ \exp\left\{-\exp\left\{-\frac{z-\mu}{\sigma}\right\}\right\}, & -\infty < z < +\infty, \xi = 0. \end{cases} \quad (6)$$

From the GEV Eq. 6, if we take limit $\xi \to 0$, we obtain the Emil Gumbel (1891–1966), cumulative probability distribution (CPD), which is the same as exponential CPD (2). Gumbel developed the subject cdf for studying the extreme observations of climate and hydrology (Gumbel 1935, 1941). The Gumbel pdf is also known as the double exponential and log-Weibull pdf.

For the largest extremes (maximum) the Gumbel pdf is given by

$$f(z) = \frac{1}{\sigma}\exp\left[-\left(\frac{z-\mu}{\sigma}\right) - \exp\left(-\frac{z-\mu}{\sigma}\right)\right], -\infty < z < +\infty \quad (7)$$

where the scale parameter $\sigma > 0$, location parameter μ. For the smallest extremes (minimum) the Gumbel pdf is of the form,

$$f(z) = \frac{1}{\sigma}\exp\left[\left(\frac{z-\mu}{\sigma}\right) - \exp\left(-\frac{z-\mu}{\sigma}\right)\right],$$
$$-\infty < z < +\infty, \sigma > 0 \quad (8)$$

Cumulative Distribution Function

The corresponding Gumbel cdfs for maximum and minimum are given by

$$F(z) = \exp\left\{-\exp\left(\tfrac{z-\mu}{\sigma}\right)\right\}, -\infty < z < +\infty, \sigma > 0,$$

and

$$F(z) = 1 - \exp\left\{-\exp\left(\tfrac{z-\mu}{\sigma}\right)\right\}, -\infty < z < +\infty, \sigma > 0, \quad (9)$$

respectively.

To apply the subject pdfs using real data, we need to solve numerically the following equation to obtain maximum likelihood estimation of the true parameter μ and σ, that is

$$\hat{\mu} = -\hat{\sigma}\log\left[\frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{-z_i}{\hat{\sigma}}\right)\right]$$
$$\hat{\sigma} - \frac{1}{n}\sum_{i=1}^{n} z_i + \frac{\sum_{i=1}^{n} z_i \exp\left(-\frac{z_i}{\hat{\sigma}}\right)}{\sum_{i=1}^{n}\exp\left(-\frac{z_i}{\hat{\sigma}}\right)} = 0. \quad (10)$$

Furthermore, we can use these estimates $\hat{\mu}$ and $\hat{\sigma}$ to obtain a 95% confidence interval for μ and σ, that is

$$\left(\hat{\mu} - 1.96\sqrt{\frac{1.10867\hat{\sigma}^2}{n}}, \hat{\mu} + 1.96\sqrt{\frac{1.10867\hat{\sigma}^2}{n}}\right) \quad (11)$$

and

$$\left(\hat{\sigma} - 1.96\sqrt{\frac{0.60793\hat{\sigma}^2}{n}}, \hat{\sigma} + 1.96\sqrt{\frac{0.60793\hat{\sigma}^2}{n}}\right). \quad (12)$$

The CPD defined in Eq. 6 which can also be obtained for GEV CPD where $\xi > 0$, was initially discovered by Maurice Frechet (1878–1973). It is especially applicable to

heavy-tailed pdf. The Frechet pdf has wide ranging applications in engineering, environmental modeling, finance and other areas. Recent applications include prediction of solar proton peak fluxes and modeling interfacial damage in microelectronic packages and material properties of constituent particles in an aluminum alloy.

The three parameter Frechet pdf is given by

$$f(z) = \frac{\xi}{\sigma}\left(\frac{\sigma}{z-\mu}\right)^{\xi+1} \exp\left\{-\left(\frac{\sigma}{z-\mu}\right)^{\xi}\right\}, \quad \mu > 0, \xi > 0, \sigma > 0. \tag{13}$$

The corresponding cdf of (13) is given by

$$F(z) = \exp\left\{-\left(\frac{\sigma}{z-\mu}\right)^{\xi}\right\}, \ \mu, \xi, \sigma > 0, \tag{14}$$

where μ, ξ and σ are the location, shape and scale parameter, respectively.

To apply the above pdf to real data, we need to obtain the maximum likelihood estimate of the parameters μ, σ and ξ by numerically solving the following system of likelihood equations

$$\frac{n}{\hat{\xi}} + \frac{n\sum_{i=1}^{n}(t_i-\hat{\mu})^{-\hat{\xi}}\log(t_i-\hat{\mu})}{\sum_{i=1}^{n}(t_i-\hat{\mu})^{-\hat{\xi}}} = \sum_{i=1}^{n}\log(t_i-\hat{\mu}) \tag{15}$$

$$\frac{n\hat{\xi}\sum_{i=1}^{n}(t_i-\hat{\mu})^{-(\hat{\xi}+1)}}{\sum_{i=1}^{n}(t_i-\hat{\mu})^{-\hat{\xi}}} = (\hat{\xi}+1)\sum_{i=1}^{n}\frac{1}{t_i-\hat{\mu}} \tag{16}$$

and

$$\hat{\sigma} = \left\{\frac{1}{n}\sum_{i=1}^{n}(t_i-\hat{\mu})^{-\hat{\xi}}\right\}^{-1/\hat{\xi}}. \tag{17}$$

The CDF defined in the family of probability distributions was developed by Waloddi Weibull (1887–1979) and carried his name. It can also be obtained from the GEV for $\xi < 0$. Weibull did significant pioneering work on reliability, providing a statistical treatment of fatigue, strength, and lifetime in engineering design (Ahmad et al. 1988; Weibull 1939b). It is also applicable in environmental modeling, finance and other areas. Recent applications include evaluation the magnitude of future earthquakes in the Pacific, Argentina, Japan and in the Indian subcontinent.

The pdf of the three - parameter Weibull is given by

$$f(z) = \frac{\xi}{\sigma}\left(\frac{\sigma}{z-\mu}\right)^{\xi-1} \exp\left\{-\left(\frac{\sigma}{z-\mu}\right)^{\xi}\right\}, \ z, \mu, \xi, \sigma > 0. \tag{18}$$

Its CDF is given by

$$F(z) = 1 - \exp\left\{-\left(\frac{\sigma}{z-\mu}\right)\right\}, \ z, \mu, \xi, \sigma > 0. \tag{19}$$

Similarly, to apply the subject pdf we need to obtain a maximum likelihood estimate of the three parameters, μ, σ and ξ that are inherent in the Weibull pdf. We can obtain such estimates by solving numerical the following system of likelihood equations,

$$\frac{n}{\hat{\xi}} + \sum_{i=1}^{n}\log(t_i-\hat{\mu}) = \frac{n\sum_{i=1}^{n}(t_i-\hat{\mu})^{\hat{\xi}}\log(t_i-\hat{\mu})}{\sum_{i=1}^{n}(t_i-\hat{\mu})^{-\hat{\xi}}} \tag{20}$$

$$\frac{n\hat{\xi}\sum_{i=1}^{n}(t_i-\hat{\mu})^{-(\hat{\xi}-1)}}{\sum_{i=1}^{n}(t_i-\hat{\mu})^{-\hat{\xi}}} = (\hat{\xi}-1)\sum_{i=1}^{n}\frac{1}{t_i-\hat{\mu}} \tag{21}$$

and

$$\hat{\sigma} = \left\{\frac{1}{n}\sum_{i=1}^{n}(t_i-\hat{\mu})^{-\hat{\xi}}\right\}^{1/\hat{\xi}}. \tag{22}$$

About the Author

For biography *see* the entry ►Mathematical and Statistical Modeling of Global Warming.

Cross References

►Extreme Value Distributions
►Generalized Weibull Distributions
►Multivariate Statistical Distributions
►Statistical Distributions: An Overview
►Statistics of Extremes
►Weibull Distribution

References and Further Reading

Abdelhafez MEM, Thomas DR (1990) Approximate prediction limits for the Weibull and extreme value regression models. Egyptian Stat J 34:408–419

Achcar JA (1991) A useful reparametrization for the extreme value distribution. Comput Stat Quart 6:113–125

Ahmad MI, Sinclair CD, Spurr BD (1988) Assessment of flood frequency models using empirical distribution function statistics. Water Resour Res 24:1323–1328

Ahsanullah M (1990) Inference and prediction problems of the Gumbel distribution based on smallest location parameters. Statistician 38:191–195

Aitkin M, Clayton D (1980) The fitting of exponential, Weibull, and extreme value distributions to complex censored survival data using GLIM. Appl Stat 29:156–163

Al-Abbasi JN, Fahmi KJ (1991) GEMPAK: a Fortran-77 program for calculating Gumbel's first, third and mixture upper earthquake magnitude distribution employing maximum likelihood estimation. Comput Geosci 17:271–290

Azuz PM (1955) Application of the statistical theory of extreme value to the analysis of maximum pit depth data for aluminum. Corrosion 12:35–46

Barnett V (1990) Ranked set sample design for environmental investigations. Environ Ecolog Stat 6:59–74

Beran M, Hosking JRM, Arnell N (1986) Comment on "Two-component extreme value distribution for flood analysis" by Fabio Rossi, Mauro Fiorentino. Pasquale Versace Water Resources Res 22:263–266

Broussard JP, Booth GG (1998) The behavior of extreme values in Germany's stock index futures: An application to intradaily margin setting. Eur J Oper Res 104:393–402

Buishand TA (1985) The effect of seasonal variation and serial correlation on the extreme value distribution of rainfall data. J Climate Appl Meteor 25:154–160

Buishand TA (1989) Statistics of extremes in climatology. Stat Neerl 43:1–30

Campell JW, Tsokos CP (1973a) The asymptotic distribution of maximum in bivariate samples. J Am Stat Assoc 68:734–739

Campbell JW, Tsokos CP (1973b) The asymptotic distribution of maxima in bivariate samples. J Am Stat Assoc 68:734–739

Changery MJ (1982) Historical extreme winds for the United States-Atlantic and Gulf of Mexico coastlines. U.S. Nuclear Regulatory Commission, NUREG/CR-2639

Cheng S, Peng L, Qi Y (1998) Almost sure convergence in extreme value theory. Math Nachr 190:43–50

Chowbury JU, Stedinger JR, Lu LH (1991) Goodness-of-fit tests for regional generalized extreme value flood distributions. Water Resour Res 27:1765–1776

Cohen JP (1986) Large sample theory for fitting an approximating Gumbel model to maxima. Sankhya A 48:372–392

Coles SG, Pan F (1996) The analysis of extreme value pollution levels: A case study. J R Stat 23:333–348

Coles SG, Tawn JA (1994) Statistical methods for multivariate extremes: an application to structural design. Appl Stat 43:1–48

Coles SG, Tawn JA (1996) A Bayesian analysis of extreme stock data. Appl Stat 45:463–478

Daniels HE (1942) A property of the distribution of extremes. Biometrika 32:194–195

Davidovich MI (1992) On convergence of the Weibull-Gnedenko distribution to the extreme value distribution. Vestnik Akad Nauk Belaruss Ser Mat Fiz, No. 1, Minsk, 103–106

De Haan L (1970) On regular variation and its application to the weak convergence of sample extremes. Mathmatical Center Tracts, 32, Mathematisch Centrum. Amsterdam

De Hann L, Resnick SI (1998) Sea and wind: Multivariate extreme at work. Extremes 1:7–46

Diebold FX, Schuermann T, Stroughair JD (1999) Pitfalls and opportunities in the use of extreme value theory in risk management. Draft Report

Eldredge GG (1957) Analysis of corrosion pitting by extreme value statistics and its application to oil well tubing caliper surveys. Corrosion 13:51–76

Embrechts P, Kluppelberg C, Mikosch T (1997) Modeling extremal events for insurance and finance. Spring, Berlin

Engelhardt M, Bain LJ (1973) Some complete and censored results for the Weibull or extreme-value distribution. Technometrics 15:541–549

Engelund S, Rackwitz R (1992) On predictive distribution function for the three asymptotic extreme value distributions. Struct Saf 11:255–258

Epstein B (1948) Application to the theory of extreme values in fracture problems. J Am Stat Assoc 43:403–412

Fahmi KJ, Al-Abbasi JN (1991) Application of a mixture distribution of extreme values to earthquake magnitudes in Iraq and conterminous regions. Geophys J R Astron Soc 107:209–217

Frechet M (1927) Sur la loi de probabilite de l'ecart maximum. Ann Soc Polon Math Cravovie 6:93–116

Frenkel JI, Kontorova TA (1943) A statistical theory of the brittle strength of real crystals. J Phys USSR 7:108–114

Fuller WE (1914) Flood flows. Trans Am Soc Civ Eng 77:564

Galambos J (1981) Extreme value theory in applied probability. Math Scient 6:13–26

Galambos J (1987) The asymptotic theory of extreme order statistics, 2nd edn. Krieger, Malabar

Goka T (1993) Application of extreme-value theory to reliability physics of electronic parts and on-orbit single event phenomena. Paper presented at the Conference on Extreme Value Theory and Its Applications, May 2–7, 1993, National Institute of Standards, Gainthersburg

Greenwood M (1946) The statistical study of infectious diseases. J R Stat Soc A 109:85–109

Greis NP, Wood EF (1981) Regional flood frequency estimation and network design. Water Resources Res 17:1167–1177

Gumbel EJ (1935) Les valeurs extremes des distribution statistuques. Ann l'Inst R Soc London A 221:163–198

Gumbel EJ (1941) The return period of flood flows. Ann Math Statist 12:163–190

Gumbel EJ (1945) Floods estimated by probability methods. Engrg News-Record 134:97–101

Gumbel EJ (1949a) The Statistical Forecast of Floods. Bulletin No. 15, 1–21, Ohio Water Resources Board

Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York

Gumbel EJ (1962a) Statistical estimation of the endurance limit – an application of extreme-value theory. In: Sarhan AE, Greenberg BG (eds) Contributions to order statistic. Wiley, New York, pp 406–431

Gumbel EJ (1962b) Statistical theory of extreme value (main results). In: Sarhan AE, Greenberg BG) Contributions to order statistics, Chapter 6. Wiley, New York

Gumbel EJ (1962c) Multivariate extremal distributions. Proceeedings of Session ISI, vol 39:471–475

Gumbel EJ, Goldstein N (1964) Empirical bivariate extremal distributions. J Am Stat Assoc 59:794–816

Harris B (1970) Order Statistics and their use in testing and estimation, vol 2. Washington

Hassanein KM (1972) Simultaneous estimation of the parameters of the extreme value distribution by sample quantiles. Technometrics 14:63–70

Henery RJ (1984) An extreme-value model for predicting the results of horse reaces. Appl Statist 33:125–133

Hisel KW (ed) (1994) Extreme values: floods and droughts. Proceedings of International Conference on Stochastic and Statistical Methods in Hydrology and Environmental Enginerring, vol 1, 1993, Kluwer

Hosking JRM (1985) Maximum-likelihood estimation of the parameters of the generalized extreme-value distribution. Appl Stat 34:301–310
Hosking JRM, Wallis JR (1988) The effect of intersite dependence on regional flood frequency analysis. Water Resources Res 24: 588–600
Jain D, Singh VP (1987) Estimating parameters of EV1 distribution for flood frequency analysis. Water Resour Res 23:59–71
Jenkinson AF (1969) Statistics of extremes, Technical Note No. 98, World Meteorological Organization, Chapter 5, pp. 183–227
Joe H (1994) Multivariate extreme value distributions with applications to environmental data. Canad J Stat Probab Lett 9:75–81
Kimball BF (1955) Practical applications of the theory of extreme values. J Am Stat Assoc 50:517–528
Longuet-Higgins MS (1952) On the statistical distribution of the heights of sea waves. J Mar Res 9:245–275
Mann NR, Scheduer EM, Fertig KW (1973) A new goodness-of-fit test for the two parameter Weibull or extreme-value distribution with unknown parameters. Comm Stat 2:383–400
Marshall RJ (1983) A spatial-temporal model for storm rainfall. J Hydrol 62:53–62
Nisan E (1988) Extreme value distribution in estimation of insurance premiums. ASA Proceedings of Business and Economic Statistics Section, pp 562–566
Nordquist JM (1945) Theory of largest values, applied to earthquake magnitude. Trans Am Geophys Union 26:29–31
Okubo T, Narita N (1980) On the distribution of extreme winds expected in Japan. National Bureau of Standards Special Publication, 560–561, 12pp
Pickands J (1981) Multivariate extreme value distributions. Proceedings of 43rd Session of the ISI. Buenos Aires, vol 49, pp 859–878
Pickands J (1986) Statistical inference using extreme order statistics. Ann Stat 3:119–131
von Mises R (1923) Uber die Variationsbreite einer Beobachtungsreihe. Sitzungsber Berlin Math Ges 22:3–8
von Mises R (1936) La distribution de las plus grande de *n* valeurs. Rev Math Union Interbalk 1:141–160. Reproduced in Selected Papers of Richard von Mises, II (1954), Am Math Soc 271–294
Weibull W (1939a) A statistical theory of the strength of materials. Ing Vet Akad Handlingar 151
Weibull W (1939b) The phenomenon of rupture in solids. Ing Vet Akad Handlingar 153:2

Generalized Hyperbolic Distributions

Matthias Fischer
University of Erlangen-Nürnberg, Erlangen, Germany

The (univariate) generalized hyperbolic distribution (GHD) family was intensively discussed originally by Barndorff-Nielsen (1977, 1978) and arose as specific normal mean-variance mixture: Assuming that X follows a normal distribution with random mean $\mu + U\beta$ ($\mu, \beta \in \mathbb{R}$) and random variance U, where U in turn is assumed to follow a generalized inverse Gaussian (GIG) distribution (see, e.g., Jørgensen 1982) with parameters $\lambda \in \mathbb{R}$, $\chi \equiv \delta^2$, $\psi \equiv \alpha^2 - \beta^2$ for $\alpha, \delta > 0, |\beta| < \alpha$, the corresponding GH density on $\mathbb{R}$ derives as

$$f(x;\mu,\delta,\alpha,\beta,\lambda) = \left[\left(\sqrt{\alpha^2-\beta^2}\right)^{\lambda}\left(\sqrt{\delta^2+(x-\mu)^2}\right)^{\lambda-\frac{1}{2}} \mathbf{K}_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2+(x-\mu)^2}\right)\right] \Big/ \left[\sqrt{2\pi}\alpha^{\lambda-\frac{1}{2}}\delta^{\lambda}\mathbf{K}_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)e^{-\beta(x-\mu)}\right],$$

where $\mathbf{K}_{\lambda}(x) = \frac{1}{2}\int_0^{\infty} t^{\lambda-1}e^{-\frac{1}{2}x(t+t^{-1})}dt$ denotes the *modified Bessel function* of the third kind (see Abramowitz and Stegun 1965). The GHD is symmetric around the location parameter μ if the skewness parameter β is zero. The parameter δ describes the scale, whereas α and λ govern both peakedness and tail behavior, respectively. Note that the GHD has heavier tails than the normal distribution. The tail behavior is like that of an exponential function times a power of $|x|$ (see, e.g., Barndorff-Nielsen 1978). An included subfamily where one tail has polynomial and the other exponential tail behavior is discussed by Aas and Haff (2006). Despite its heavier tails, the moment-generating function of a GH variable X still exists and is given by

$$\mathcal{M}(u) \equiv \mathbb{E}(e^{uX}) = \exp(u\mu)\cdot\left(\frac{\alpha^2-\beta^2}{\alpha^2-((\beta+u)^2)}\right)^{\frac{\lambda}{2}}\cdot \frac{\mathbf{K}_{\lambda}\left(\delta\sqrt{\alpha^2-(\beta+u)^2}\right)}{\mathbf{K}_{\lambda}\left(\delta\sqrt{\alpha^2-\beta^2}\right)}, \quad |\beta+u|<\alpha.$$

Hence, mean, variance and higher moments can be derived in a straightforward manner (see, Barndorff-Nielsen and Blæsild 1981 and Blæsild 1990). Above that, the GHD is both infinitely divisible (see Barndorff-Nielsen and Halgreen 1977) and self-decomposable (see Halgreen 1979). Random numbers from a GH population can be generated using an efficient algorithm of Atkinson (1979). Maximum likelihood estimation of the unknown parameters based on i.i.d samples from a GHD might be challenging to some extent because of the flatness of the likelihood function in λ (see Prause 1999). Barndorff-Nielsen (1995), for instance, provides an example of two different GH families (with different λ's) whose density are practically identical over the range covering 99.99% of their mass.

This is one of the reasons why two specific subclasses gain special attraction within the GH-family.

Firstly, the hyperbolic (HYP) distribution family ($\lambda = 1$) whose name derives from the fact that for such a distribution the graph of the log-density is a hyperbola. Hyperbolic distributions were originally motivated as a distributional model for particle sizes of a sand sample from aeolian sand deposits (see Bagnold and Barndorff-Nielsen 1980) but were also successfully applied in turbulence (see Barndorff-Nielsen et al. 1989) and finance (see Eberlein and Keller 1995, Bibby and Sørensen 1997 or Küchler et al. 1999). Goodness-of fit tests for the hyperbolic distributions were proposed by Puig and Stephens (2001).

Secondly, the normal inverse Gaussian family (NIG) ($\lambda = -1/2$) which shows similar fit and flexibility but, in contrast to the hyperbolic distribution, has the feature of being closed under convolution. This can be used, for instance, to price different kind of options (see, e.g., Albrecher and Predota 2004). Barndorff-Nielsen (1995) and Rydberg (1997) discuss ►Lévy processes based on NIG distributions as a model for stock returns. Bayesian estimation of NIG distributions can be found in Karlis and Lillestöl (2004). Occasionally, the so-called hyperboloid distributions ($\lambda = 0$) are focussed (see, e.g., Blæsild 1990). Beyond that, both normal distribution ($\delta \to \infty, \delta/\alpha \to \sigma^2$) and Student-$t$ distribution ($\lambda = -\nu/2, \alpha = \beta \to 0, \delta = \sqrt{\nu}$) are included as limiting cases within the GHD family. As in the one-dimensional case, the multivariate generalized hyperbolic distribution results as a specific normal mean-variance mixture distribution and appeared first in Barndoff-Nielsen (1977, 1978) with intensive discussion in Blæsild (1981). Making use of the mixture representation, Protassov (2004) and Hu (2005) discuss the EM-based estimation maximum likelihood parameter estimation. Applications to finance are provided by Prause (1999) and Hu (2005), for example.

Cross References

►Inverse Gaussian Distribution

References and Further Reading

Aas K, Haff IH (2006) The generalized hyperbolic skew Student's t-distribution. J Financ Econom 4(2):275–309

Abramowitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York

Albrecher H, Predota M (2004) On Asian option pricing for NIG Lévy processes. J Comput Appl Math 1(1):153–168

Atkinson AC (1979) Simulation of generalized inverse Gaussian, generalized hyperbolic, Gamma and related random variables. Research report No. 52, Department of Theoretical Statistics, University of Aarhus

Bagnold RA, Barndorff-Nielsen OE (1980) The pattern of natural size distributions. Sedimentology 27:199–207

Barndorff-Nielsen OE (1977) Exponentially decreasing distributions for the logarithm of particle size. Proc R Soc Lond A 353:401–419

Barndorff-Nielsen OE (1978) Hyperbolic distributions and distributions on hyperbolae. Scand J Stat 5:151–157

Barndorff-Nielsen OE (1982) The hyperbolic distribution in statistical physics. Scand J Stat 9:43–46

Barndorff-Nielsen OE (1995) Normal inverse Gaussian processes and the modelling of stock returns. Research Report No. 300, Department of Theoretical Statistics, University of Aarhus

Barndorff-Nielsen OE, Blæsild P (1981) Hyperbolic distributions and ramifications: contributions to theory and application. In: Taillie C et al (eds) Statistical distributions in scientific work, vol 4. D. Reidel, Dordrecht/Holland, pp 45–66

Barndorff-Nielsen OE, Halgreen C (1977) Infinite divisibility of the hyperbolic and generalized inverse Gaussian distribution. Zeitschrift für Wahrscheinlichkeits theorie und verwandte Gebiete 38:309–312

Barndorff-Nielsen OE, Jensen JL, Soerensen M (1989) Wind shear and hyperbolic distributions. Meteorology 46:417–431

Bibby M, Sørensen M (1997) A hyperbolic diffusion model for stock prices. Finance Stoch 1:25–41

Blæsild P (1981) The two-dimensional hyperbolic distributions and related distributions, with an application to Johannsen's bean data. Biometrika 68:251–263

Blæsild P (1990) Hyperbolic distributions: cumulants, skewness and kurtosis. Research Report No. 209, Department of Theoretical Statistics, University of Aarhus

Eberlein E, Keller U (1995) Hyperbolic distributions in finance. Bernoulli 1(3):281–299

Halgreen C (1979) Self-decomposability of the generalized inverse Gaussian and hyperbolic distributions. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 47:13–18

Hu W (2005) Calibration of multivariate generalized hyperbolic distributions using the EM algorithm, with applications in risk management, portfolio optimization and portfolio credit risk. PhD thesis, The Florida State University

Jensen JL (1988) Hyperboloid distributions. Research Report No. 59, Department of Theoretical Statistics, University of Aarhus

Jørgensen B (1982) Statistical properties of the generalized inverse Gaussian distribution. Lecture Notes in Statistics No. 9, Springer, Berlin

Karlis D, Lillestöl J (2004) Bayesian estimation of NIG models via Monte Carlo methods. Appl Stoch Model Bus Ind 20(4): 323–338

Küchler E, Neumann K, Sørensen M, Streller A (1999) Stock returns and hyperbolic distributions. Math Comput Model 29:1–15

Prause K (1999) The generalized hyperbolic model: estimation, financial derivatives and risk measures. PhD thesis, University of Freiburg, Freiburg

Protassov RS (2004) EM-based estimation maximum likelihood parameter estimation for multivariate generalized hyperolic distributions with fixed λ. Stat Comput 14:67–77

Puig P, Stephens MA (2001) Goodness-of-fit tests for the hyperbolic distribution. Canad J Stat 29(2):309–320

Rydberg TH (1997) The NIG Lévy process: simulation and approximation. Commun Stat Stoch Mod 13:887–910

Generalized Linear Models

Joseph M. Hilbe
Emeritus Professor
University of Hawaii, Honolulu, HI, USA
Adjunct Professor of Statistics
Arizona State University, Tempe, AZ, USA
Solar System Ambassador
California Institute of Technology, Pasadena, CA, USA

History

Generalized Linear Models (GLM) is a covering algorithm allowing for the estimation of a number of otherwise distinct statistical regression models within a single framework. First developed by John Nelder and R.W.M. Wedderburn in 1972, the algorithm and overall GLM methodology has proved to be of substantial value to statisticians in terms of the scope of models under its domain as well as the number of accompanying model statistics facilitating an analysis of fit. In the early days of statistical computing - from 1972 to 1990 - the GLM estimation algorithm also provided a substantial savings of computing memory compared to what was required using standard maximum likelihood techniques. Prior to Nelder and Wedderburn's efforts, GLM models were typically estimated using a Newton-Raphson type full maximum likelihood method, with the exception of the Gaussian model. Commonly known as normal or linear regression, the Gaussian model is usually estimated using a least squares algorithm. GLM, as we shall observe, is a generalization of ordinary least squares regression, employing a weighted least squares algorithm that iteratively solves for parameter estimates and standard errors.

In 1974, Nelder coordinated a project to develop a specialized statistical application called GLIM, an acronym for Generalized Linear Interactive Modeling. Sponsored by the Royal Statistical Society and Rothamsted Experimental Station, GLIM provided the means for statisticians to easily estimate GLM models, as well as other more complicated models which could be constructed using the GLM framework. GLIM soon became one of the most used statistical applications worldwide, and was the first major statistical application to fully exploit the PC environment in 1981. However, it was discontinued in 1994. Presently, nearly all leading general purpose statistical packages offer GLM modeling capabilities; e.g., SAS, R, Stata, S-Plus, Genstat, and SPSS.

Theory

Generalized linear models software, as we shall see, allows the user to estimate a variety of models from within a single framework, as well as providing the capability of changing models with minimal effort. GLM software also comes with a host of standard residual and fit statistics, which greatly assist researchers with assessing the comparative worth of models.

Key features of a generalized linear model include (1) having a response, or dependent variable, selected from the single parameter exponential family of probability distributions, (2) having a link function that linearizes the relationship between the fitted value and explanatory predictors, and (3) having the ability to be estimated using an Iteratively Re-weighted Least Squares (IRLS) algorithm.

The exponential family probability function upon which GLMs are based can be expressed as

$$f(y_i;\theta_i,\phi) = \exp\{(y_i\theta_i - b(\theta_i))/\alpha_i(\phi) - c(y_i;\phi)\} \quad (1)$$

where the distribution is a function of the unknown data, y, which may be conditioned by explanatory predictors, for given parameters θ and ϕ. For generalized linear models, the probability distribution is re-parameterized such that the distribution is a function of unknown parameters based on known data. In this form the distribution is termed a likelihood function, the goal of which is to determine the parameters making the data most likely. Statistician's log-transform the likelihood function in order to convert it to an additive rather than the multiplicative scale. Doing so greatly facilitates estimation based on the function. The log-likelihood function is central to all maximum likelihood estimation algorithms. It is also the basis of the deviance function, which was traditionally employed in GLM algorithms as both the basis of convergence and as a goodness-of-fit statistic. The log-likelihood is defined as

$$L(\theta_i;y_i,\phi) = \sum_{i=1}^{n}\{(y_i\theta_i - b(\theta_i))/\alpha_i(\phi) - c(y_i;\phi)\} \quad (2)$$

where θ is the link function, $b(\theta)$ the cumulant, $\alpha_i(\phi)$ the scale, and $c(y;\phi)$ the normalization term, guaranteeing that the distribution sums to one. The first derivative of the cumulant with respect to θ, $b'(\theta)$, is the mean of the

function, μ; the second derivative, $b''(\theta)$, is the variance, $V(\mu)$. The deviance function is given as

$$2\sum_{i=1}^{n}\{L(y_i;y_i) - L(y_i,\mu_i)\} \tag{3}$$

Table 1 presents the standard probability distribution functions (PDF) belonging to the GLM family.

Each of the distributions in Table 1 are members of the exponential family. It should be noted, however, that the three continuous GLM distributions are usually parameterized with two rather than one parameter: Gaussian, gamma, and inverse Gaussian. Within the GLM framework though, the scale parameter is not estimated, although it is possible to point-estimate the scale value from the dispersion statistic, which is typically displayed in GLM model output. Binomial and count models have the scale value set at 1.0. As a consequence, $\alpha(\phi)$ and ϕ are many times excluded when presenting the GLM-based exponential log-likelihood.

Table 2 provides the formulae for the deviance and log-likelihoods of each GLM family. Also provided is the variance for each family function. The first line of each GLM distribution or family shows the deviance, with the next two providing the log-likelihood functions parameterized in terms of μ and $x'\beta$ respectively. The $x'\beta$ parameterization is used when models are estimated using a full maximum likelihood algorithm.

Generalized Linear Models. Table 1 GLM families: canonical

Family	Characteristics
Continuous distributions	
Gaussian	Standard normal or linear regression
Gamma	Positive-only continuous
Inverse gaussian	Positive-only continuous
Count	
Poisson	Equidispersed count
Negative binomial (NB-C)	Count, with the ancillary parameter a constant
Binary - Bernoulli	Binomial distribution with $m = 1$.
Logistic	Binary (1/0) response
Binomial	Proportional (y/m) :
Logistic (grouped)	y = number of 1's m = cases having same covariate pattern

Note that the link and cumulant functions for each of the above GLM log-likelihood functions can easily be abstracted from the equations, which are formatted in terms of the exponential family form as defined in Eq. 1. For example, the link and cumulant of the Bernoulli distribution, upon which ▶logistic regression is based, are respectively $\ln(\mu/(1-\mu))$ and $-\ln(1-\mu)$. With the link function defined in this manner, the linear predictor for the canonical Bernoulli model (logit) is expressed as:

$$\theta_i = x_i'\beta = \ln(\mu_i/(1-\mu_i)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n. \tag{4}$$

In GLM terminology, $x'\beta$ is also referred to as η, and the link as $g(\mu)$. For links directly derived from the GLM family PDF, the following terms are identical:

$$\theta = x_i'\beta = \eta = g(\mu). \tag{5}$$

The link function may be inverted such that μ is defined in terms of η. The resulting function is called the inverse link function, or $g^{-1}(\eta)$. For the above logit link, $\eta = \ln(\mu/(1-\mu))$. μ is therefore defined, for each observation in the logistic model, as

$$\mu_i = 1/(1+\exp(-\eta_i)) = (\exp(\eta_i))/(1+\exp(\eta_i)) \tag{6}$$

or

$$\mu_i = 1/\left(1+\exp\left(-x_i'\beta\right)\right) = \left(\exp\left(x_i'\beta\right)\right)/\left(1+\exp\left(x_i'\beta\right)\right) \tag{7}$$

Another key feature of generalized linear models is the ability to use the GLM algorithm to estimate non-canonical models; i.e., models in which the link function is not directly derived from the underlying pdf, i.e., $x'\beta$ or η is not defined in terms of the value of θ given in the above listing of log-likelihood functions. Theoretically any type of link function can be associated with a GLM log-likelihood, although many might not be appropriate for the given data. A power link is sometimes used for non-binomial models the power, p, in μ^p, is allowed to vary. The statistician employs a value for the power that leads to a minimal value for the deviance. Powers typically range from 2 to −3, with μ^2 being the square link, μ^1 the log, μ^0 the identity, and μ^{-1} and μ^{-2} the inverse and inverse quadratic link functions, respectively. Intermediate links are also used, e.g., $\mu^{.5}$, the square root link. The normal linear model has an identity link, with the linear predictor being identical to the fitted value.

The probit and log-linked negative binomial (NB-2) models are two commonly used non-canonical linked regression models. The probit link is often used with the ▶binomial distribution for probit models. Although the probit link is not directly derived from the binomial PDF, the estimates of the GLM-based probit model are identical

Generalized Linear Models. Table 2 GLM variance, deviance, and log-likelihood functions

Family	Variance, deviance, log-likelihood ($\mu\|x\beta$)
Gaussian	$\sum(y-\mu)^2$
1	$\sum\{(y\mu-\mu^2/2)/\sigma^2-y^2/2\sigma^2-5\ln(2\pi\sigma^2)\}$
	$\sum\{[y(x\beta)-(x\beta)^2/2]/\sigma^2-y^2/2\sigma^2-5\ln(2\pi\sigma^2)\}$
Bernoulli	$2\sum\{y\ln(y/\mu)+(1-y)\ln((1-y)/(1-\mu))\}$
$\mu(1-\mu)$	$\sum\{y\ln(\mu/(1-\mu))+\ln(1-\mu)\}$
	$\sum\{y(x\beta)-\ln(1+\exp(x\beta))\}$
Binomial	$2\sum\{y\ln(y/\mu)+(m-y)\ln((m-y)/(m-\mu))\}$
$\mu(1-\mu/m)$	$\sum\{y\ln(\mu/m)+(m-y)\ln(1-\mu/m)+\ln\Gamma(m+1)-\ln\Gamma(y+1)+\ln\Gamma(m-y+1)\}$
	$\sum\{y\ln((\exp(x\beta))/(1+\exp(x\beta)))-(m-y)\ln(\exp(x\beta)+1)+\ln\Gamma(m+1)-\ln\Gamma(y+1)+\ln\Gamma(m-y+1)\}$
Poisson	$2\sum\{y\ln(y/\mu)-(y-\mu)\}$
μ	$\sum\{y\ln(\mu)-\mu-\ln\Gamma(y+1)\}$
	$\sum\{y(x\beta)-\exp(x\beta)-\ln\Gamma(y+1)\}$
NB2	$2\sum\{y\ln(y/\mu)-(y+1/\alpha)\ln((1+\alpha y)/(1+\alpha\mu))\}$
$\mu+\alpha\mu^2$	$\sum\{y\ln((\alpha\mu)/(1+\alpha\mu))-(1/\alpha)\ln(1+\alpha\mu)+\ln\Gamma(y+1/\alpha)-\ln\Gamma(y+1)-\ln\Gamma(1/\alpha)\}$
	$\sum\{y\ln((\alpha\exp(x\beta))/(1+\alpha\exp(x\beta)))-(\ln(1+\alpha\exp(x\beta)))/\alpha+\ln\Gamma(y+1/\alpha)-\ln\Gamma(y+1)-\ln\Gamma(1/\alpha)\}$
NBC	$2\sum\{y\ln(y/\mu)-(y+1/\alpha)\ln((1+\alpha y)/(1+\alpha\mu))\}$
$\mu+\alpha\mu^2$	$\sum\{y\ln(\alpha\mu/(1+\alpha\mu))-(1/\alpha)\ln(1+\alpha\mu)+\ln\Gamma(y+1/\alpha)-\ln\Gamma(y+1)-\ln\Gamma(1/\alpha)\}$
	$\sum\{y(x\beta)+(1/\alpha)\ln(1-\exp(x\beta))+\ln\Gamma(y+1/\alpha)-\ln\Gamma(y+1)-\ln\Gamma(1/\alpha)\}$
Gamma	$2\sum\{(y-\mu)/\mu-\ln(y/\mu)\}$
μ^2	$\sum\{((y/\mu)+\ln(\mu))/-\phi+\ln(y)(1-\phi)/\phi-\ln(\phi)/\phi-\ln\Gamma(1/\phi)\}$
	$\sum\{(y(x\beta)-\ln(x\beta))/-\phi+\ln(y)(1-\phi)/\phi-\ln(\phi)/\phi-\ln\Gamma(1/\phi)\}$
Inv Gauss	$\sum\{(y-\mu)^2/(y\mu^2)\}$
μ^3	$\sum\{(y/(2\mu^2)-1/\mu)/-\sigma^2+1/(-2y\sigma^2)-5\ln(2\pi y^3\sigma^2)\}$
	$\sum\left\{y/(2x\beta)-\sqrt{x\beta}/-\sigma^2+1/(-2y\sigma^2)-5\ln(2\pi y^3\sigma^2)\right\}$

to those produced using full maximum likelihood methods. The canonical negative binomial (NB-C) is not the traditional negative binomial used to model overdispersed Poisson data. Rather, the use of the log link with the negative binomial (LNB) family duplicates estimates produced by full maximum likelihood NB-2 commands. However, like all non-canonical models, the standard errors of the LNB are slightly different from those of a full maximum likelihood NB-2, unless the traditional GLM algorithm in Table 5 is amended to produce an observed information matrix that is characteristic of full maximum likelihood estimation. The information derived from the algorithm given in Table 5 uses an expected information matrix, upon which standard errors are based. Applications such as Stata's *glm* command, SAS's *Genmod* procedure, and R's *glm*() and *glm.nb*() functions allow the user to select which information is to be used for standard errors.

The negative binomial family was not added to commercial GLM software until 1993 (Stata), and is in fact a member of the GLM family only if its ancillary or heterogeneity, parameter is entered into the algorithm as a constant. Setting the ancillary parameter, α, to a value

that minimizes the Pearson dispersion statistic closely approximates the value of α estimated using a full maximum likelihood command. SAS, Stata, and R provide the capability for a user to estimate α using a maximum likelihood subroutine, placing the value determined into the GLM algorithm as a constant. The resulting estimates and standard errors are identical to a full NB-2 estimation. These applications also provide the capability of allowing the software to do this automatically.

Generalized Linear Models. Table 3 Foremost non-canonical models

Family-link	Function
Continuous distributions	
Lognormal	Positive continuous
Log-gamma	Exponential survival model
Log-inverse gaussian	Steep initial peak; long slow tail
Bernoulli/binomial:	
Probit	Normal
Complementary loglog	Asymmetric distribution: >0.5 elongated
Loglog	Asymmetric distribution: <0.5 elongated
Negative binomial	
Log (NB2)	Overdispersed Poisson

The ability to incorporate non-canonical links into GLM models greatly extends the scope of models which may be estimated using its algorithm. Commonly used non-canonical models are shown in Table 3.

The link, inverse link, and first derivative of the link for the canonical functions of the standard GLM families, as well as the most used non-canonical functions, are given in Table 4.

IRLS Algorithm

Generalized linear models have traditionally been modeled using an Iteratively Re-Weighted Least Squares (IRLS) algorithm. IRLS is a version of maximum likelihood called Fisher Scoring, and can take a variety of forms. A schematic version of the IRLS algorithm is given in Table 5.

Goodness-of-Fit

GLM models are traditionally evaluated as to their fit based on the deviance and Pearson Chi2, or χ^2, statistics. Lower values of these statistics indicate a better fitted model. Recently, statisticians have also employed the Akaike (AIC) and Bayesian (BIC) Information Criterion statistics as measures of fit. Lower values of the AIC and BIC statistics also indicate better fitted models. The Pearson Chi2, AIC, and BIC statistics are defined in Table 5, and are calculated after a model has been estimated.

The Pearson dispersion statistic is used with Poisson, negative binomial, and binomial models as an indicator of excessive correlation in the data. Likelihood based models,

Generalized Linear Models. Table 4 GLM link functions (* canonical)

Link name	Link	Inverse link	1st Derivative
Gaussian			
*Identity	μ	η	1
Binomial (Bernoulli: $m = 1$)			
*Logit	$\ln(\mu/(m-\mu))$	$m/(1+\exp(-\eta))$	$m/(\mu(m-\mu))$
Probit	$\Phi^{-1}(\mu/m)$	$m\Phi(\eta)$	$m/\phi\{\Phi^{-1}(\mu/m)\}$
Cloglog	$\ln(-\ln(1-\mu/m))$	$m(1-\exp(-\exp(\eta)))$	$(m(1-\mu/m)\ln(1-\mu/m))^{-1}$
Poisson			
*Log	$\ln(\mu)$	$\exp(\eta)$	$1/\mu$
Neg Bin			
*NB-C	$\ln(\mu/(\mu+1/\alpha))$	$\exp(\eta)/(\alpha(1-\exp(\eta)))$	$1/(\mu+\alpha\mu^2)$
Log	$\ln(\mu)$	$\exp(\eta)$	$1/\mu$
Gamma			
*Inverse	$1/\mu$	$1/\eta$	$-1/\mu^2$
Inverse Gaussian			
*Inv Quad	$1/\mu^2$	$1/\sqrt{\eta}$	$-1/\mu^3$

Generalized Linear Models. Table 5 Generic GLM estimating algorithm (expected information matrix)

$\mu = (y + \text{mean}(y))/2$	// Initialize μ; non-binomial
$\mu = (y + 0.5)/(n + 1)$	// Initialize μ; binomial
$\eta = g(\mu)$	// Initialize η; link
WHILE (abs(Δ Dev)>tolerance){	// Loop
$w = 1/(Vg'^2)$	// Weight
$z = \eta + (y - \mu)g'$	// Working response
$\beta = (X'wX)^{-1}X'wz$	// Estimation of parameters
$\eta = x'\beta$	// Linear predictor,η
$\mu = g^{-1}(\eta)$	// Fit, μ; inverse link
Dev0 = Dev	
Dev = Deviance function	// Deviance or LL
Δ Dev = Dev-Dev0	// Check for difference
}	
Chi2 = $\sum(y - \mu)^2/V(\mu)$	// Pearson χ^2
AIC = $(-2LL + 2p)/n$	// AIC GOF statistic
BIC = $-2 \cdot LL + p \cdot \ln n$	// BIC GOF statistic
Where p = number of model predictors + intercept n = number of observations in model LL = log-likelihood function V = variance; $g(\mu)$ = link; $g^{-1}(\eta)$ = inverse link; $g' = \partial\eta/\partial\mu$	

being derived from a PDF, assume that observations are independent. When they are not, correlation is observed in the data. Values of the Pearson dispersion greater than 1.0 indicate more correlation in the data than is warranted by the assumptions of the underlying distribution. Some statisticians have used the deviance statistic on which to base the dispersion, but simulation studies have demonstrated that Pearson is the correct statistic. See ►Modeling count data in this volume for additional information.

From the outset, generalized linear models software has offered users a number of useful residuals which can be used to assess the internal structure of the modeled data. Pearson and deviance residuals are the two most recognized GLM residuals associated with GLM software. Both are observation-based statistics, providing the proportionate contribution of an observation to the overall Pearson Chi2 and deviance fit statistics. The two residuals are given, for each observation, as:

$$\text{Pearson} \qquad (y - \mu)/\sqrt{V(\mu)} \tag{8}$$

$$\text{deviance} \qquad \text{sgn}(y - \mu)\sqrt{\text{deviance}} \tag{9}$$

The Pearson Chi2 and deviance fit can also be calculated on the basis of their residuals by taking the square of each of the residuals respectively, and summing them over all observations in the model. However, they are seldom calculated in such a manner.

Both the Pearson and deviance residuals are usually employed in standardized form. The standardized versions of the Pearson and deviance residuals are given by dividing the respective statistic by $\sqrt{1-h}$ where h is the hat matrix diagonal. Standardized Pearson residuals are normalized to a standard deviation of 1.0 and are adjusted to account for the correlation between y and μ. The standardized deviance residuals are the most commonly used residuals for assessing the internal shape of the modeled data.

Another residual now finding widespread use is the Anscombe residual. First implemented into GLM software in 1993, it now enjoys use in many major software applications. The Anscombe residuals are defined specifically for each family, with the intent of normalizing the residuals as much as possible. The general formula for Anscombe residuals is given as

$$\int_y^\mu d\mu V^{-1/3}(\mu) \tag{10}$$

with $V^{-1/3}(\mu)$ as the inverse cube of the variance. The Anscombe residual for the binomial family is displayed as

$$\frac{A(y) - A(\mu)}{\mu(1-\mu)^{-1/6}\sqrt{\frac{1-h}{m}}} \tag{11}$$

with $A()$ equal to $2.05339 \cdot (\text{Incomplete Beta } (2/3, 2/3, z)$, z taking the value of μ or y. A standard use of this statistic is to graph it on either the fitted value, or the linear predictor. Values of the Anscombe residual are close to those of the standardized deviance residuals.

Application

Consider data from the 1912 Titanic disaster. Information was collected on the survival status, gender, age, and ticket class of the various passengers. With *age* (1 = adult; 0 = child) and *sex* (1 = male; 0-female), and *class* (1 = 1st; 2 = 2nd; 3 = 3rd) with 3rd class as the reference, a simple binary ►logistic regression can be run using a GLM command (Stata). The type of model to be estimated is declared using the *family*() and *link*() functions. *eform* indicates that the coefficients are to be exponentiated, resulting in odds ratios for the logistic model. Note the fact that 1st class passengers had a near 6 times greater odds of survival than did 3rd class passengers. The statistics displayed in the model output are fairly typical of that displayed in GLM software applications.

```
.glm survived age sex class1 class2, fam(bin) eform

Generalized linear models                         No. of obs      =      1316
Optimization     : ML                             Residual df     =      1311
                                                  Scale parameter =         1
Deviance         =  1276.200769                   (1/df) Deviance =   .973456
Pearson          =  1356.674662                   (1/df) Pearson  =   1.03484

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  .9773562
Log likelihood   = -638.1003845                   BIC             = -8139.863
------------------------------------------------------------------------------
             |                 OIM
    survived | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3479809   .0844397    -4.35   0.000     .2162749    .5598924
         sex |   .0935308   .0135855   -16.31   0.000     .0703585    .1243347
      class1 |    5.84959   .9986265    10.35   0.000     4.186109    8.174107
      class2 |   2.129343   .3731801     4.31   0.000     1.510315    3.002091
------------------------------------------------------------------------------
```

Using *R*, the same model would be specified by

```
glm(survived ~ age + sex + class1 + class2, family=binomial, link=logit, data=titanic).
```

About the Author

For biography *see* the entry ►Logistic Regression.

Cross References

►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Categorical Data Analysis
►Designs for Generalized Linear Models
►Dispersion Models
►Exponential Family Models
►Logistic Regression
►Model-Based Geostatistics
►Modeling Count Data
►Optimum Experimental Design
►Poisson Regression
►Probit Analysis
►Robust Regression Estimation in Generalized Linear Models
►Statistics: An Overview

References and Further Reading

Collett D (2003) Modeling binary data, 2nd edn. Chapman & Hall/CRC, London

Dobson AJ, Barnett AG (2008) An introduction to generalized linear models, 3rd edn. Chapman & Hall/CRC, Boca Raton

Hardin JW, Hilbe JM (2007) Generalized linear models and extensions, 2nd edn. Stata, College Station

Hilbe JM (1994) Generalized linear models. Am Stat 48:255–265

Hilbe JM (2007) Negative binomial regression. Cambridge University Press, Cambridge

Hilbe JM (2009) Logistic regression models. Chapman & Hall/CRC, Boca Raton

Hoffmann JP (2004) Generalized linear models: an applied approach. Allyn & Bacon, Boston

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall/CRC, London

Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc A 135:370–384

Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61:439–447

Generalized Quasi-Likelihood (GQL) Inferences

Brajendra C. Sutradhar
University Research Professor
Memorial University, St. John's, NL, Canada

QL Estimation for Independent Data

For $i = 1, \ldots, K$, let Y_i denote the response variable for the ith individual, and $x_i = (x_{i1}, \ldots, x_{iv}, \ldots, x_{ip})'$ be the associated $p-$dimensional covariate vector. Also, let β be the $p-$dimensional vector of regression effects of x_i on y_i. Further suppose that the responses are collected from K independent individuals. It is understandable that if the probability distribution of Y_i is not known, then one can not use the well known likelihood approach to estimate the underlying regression parameter β. Next suppose that only two moments of the data, that is, the mean and the variance functions of the response variable Y_i for all $i = 1, \ldots, K$, are known, and for a known functional form $a(\cdot)$, these moments are given by

$$E[Y_i] = a'(\theta_i) \text{ and } \text{var}[Y_i] = a''(\theta_i), \tag{1}$$

where for a link function $h(\cdot)$, $\theta_i = h\left(x_i'\beta\right)$, and $a'(\theta_i)$ and $a''(\theta_i)$ are the first and second order derivatives of $a(\theta_i)$, respectively, with respect to θ_i. For the estimation of the regression parameter vector β under this independence set up, Wedderburn (1974) (see also McCullagh (1983)) proposed to solve the so-called quasi-likelihood (QL) estimating equation given by

$$\sum_{i=1}^{K}\left[\frac{\partial a'(\theta_i)}{\partial\beta}\frac{(y_i - a'(\theta_i))}{a''(\theta_i)}\right] = 0. \tag{2}$$

Let $\hat{\beta}_{QL}$ be the QL estimator of β obtained from (2). It is known that this estimator is consistent and highly efficient. In fact, for Poisson and binary data, for example, $\hat{\beta}_{QL}$ is equivalent to the maximum likelihood (ML) estimator and hence it turns out to be an optimal estimator.

Illustration for the Poisson Case

For the Poisson data, one uses

$$a(\theta_i) = \exp(\theta_i) \tag{3}$$

with identity link function $h(\cdot)$, that is, $\theta_i = x_i'\beta$. This gives the mean and the variance functions as

$$\text{var}(Y_i) = a''(\theta_i) = E(Y_i) = a'(\theta_i) = \mu_i \text{ (say)} = \exp\left(x_i'\beta\right),$$

yielding by (2), the QL estimating equation for β as

$$\sum_{i=1}^{K} x_i(y_i - \mu_i) = 0. \tag{4}$$

Note that as the Poisson density is given by $f(y_i|x_i) = \frac{1}{y_i!}\exp[y_i log(\mu_i) - \mu_i]$, with $\mu_i = \exp(\theta_i) = \exp\left(x_i'\beta\right)$, it follows that the log likelihood function of β has the form $\log L(\beta) = -\sum_{i=1}^{K} log(y_i!) + \sum_{i=1}^{K}[y_i\theta_i - a(\theta_i)]$, yielding the likelihood equation for β as

$$\frac{\partial \log L}{\partial\beta} = \sum_{i=1}^{K}[y_i - a'(\theta_i)]\frac{\partial\theta_i}{\partial\beta} = \sum_{i=1}^{K} x_i(y_i - \mu_i) = 0, \tag{5}$$

which is the same as the QL estimating Eq. 4. Thus, if the likelihood function were known, then the ML estimate of β would be the same as the QL estimate $\hat{\beta}_{QL}$.

Illustration for the Binary Case

For the binary data, one uses

$$a'(\theta_i) = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \mu_i \text{ and } a''(\theta_i) = \mu_i(1 - \mu_i), \tag{6}$$

with $\theta_i = x_i'\beta$. The QL estimating Eq. 2 for the binary data, however, provides the same formula (4) as in the Poisson case, except that now for the binary case $\mu_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$, whereas for the Poisson case $\mu_i = \exp(\theta_i)$.

As far as the ML estimation for the binary case is concerned, one first writes the binary density given by $f(y_i|x_i) = \mu_i^{y_i}(1 - \mu_i)^{1-y_i}$. Next by writing the log likelihood function as $\log L(\beta) = \sum_{i=1}^{K} y_i\mu_i + \sum_{i=1}^{K}(1-y_i)(1-\mu_i)$, one obtains the same likelihood estimating equation as in (5), except that here $\mu_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$, under the binary model. Since the QL estimating Eq. 4 is the same as the ML estimating Eq. 5, it then follows that the ML and QL estimates for β would also be the same for the binary data.

GQL Estimation: A Generalization of the QL Estimation to the Correlated Data

As opposed to the independence set up, we now consider y_i as a vector of T repeated binary or count responses, collected from the i–th individual, for all $i = 1, \ldots, K$. Let $y_i = (y_{i1}, \ldots, y_{it}, \ldots, y_{iT})'$, where y_{it} represents the response recorded at time t for the ith individual. Also, let $x_{it} = (x_{it1}, \ldots, x_{itv}, \ldots, x_{itp})'$ be the p–dimensional covariate vector corresponding to the scalar y_{it}, and β be the p–dimensional regression effects of x_{it} on y_{it} for all $i = 1, \ldots, K$, and all $t = 1, \ldots, T$. Suppose that μ_{it} and σ_{itt} be the mean and the variance of Y_{it}, that is $\mu_{it} = E[Y_{it}]$ and $\text{var}[Y_{it}] = \sigma_{itt}$. Note that both μ_{it} and σ_{itt} are functions of β. But, when the variance is a function of mean, it is sufficient to estimate β involved in the mean function only, by treating β involved in the variance function to be known. Further note that since the T repeated responses of an individual are likely to correlated, the estimate of β to be obtained by ignoring the correlations, that is, the solution of the independence assumption based QL estimating equation

$$\sum_{i=1}^{K}\sum_{t=1}^{T}\left[\frac{\partial \mu_{it}}{\partial \beta}\frac{(y_i - \mu_{it})}{\sigma_{itt}}\right] = 0, \tag{7}$$

for β, will be consistent but inefficient. As a remedy to this inefficient estimation problem, Sutradhar (2003) has proposed a generalization of the QL estimation approach, where β is now obtained by solving the GQL estimating equation given by

$$\sum_{i=1}^{K}\frac{\partial \mu_i'}{\partial \beta}\Sigma_i^{-1}(\rho)(y_i - \mu_i) = 0, \tag{8}$$

where $\mu_i = (\mu_{i1}, \ldots, \mu_{it}, \ldots, \mu_{iT})'$ is the mean vector of Y_i, and $\Sigma_i(\rho)$ is the covariance matrix of Y_i that can be expressed as $\Sigma_i(\rho) = A_i^{\frac{1}{2}} C_i(\rho) A_i^{\frac{1}{2}}$, with $A_i = \text{diag}[\sigma_{i11}, \ldots, \sigma_{itt}, \ldots, \sigma_{iTT}]$ and $C_i(\rho)$ as the correlation matrix of Y_i, ρ being a correlation index parameter.

Note that the use of the GQL estimating Eq. 8 requires the structure of the correlation matrix $C_i(\rho)$ to be known, which is, however, unknown in practice. To overcome this difficulty, Sutradhar (2003) has suggested a general stationary auto-correlation structure given by

$$C_i(\rho) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{T-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \cdots & 1 \end{bmatrix}, \tag{9}$$

(see also Sutradhar and Das (1999, Sect. 3)), for all $i = 1, \ldots, K$, where for $\ell = 1, \ldots, T-1$, ρ_ℓ represents the lag ℓ auto-correlation. As far as the estimation of the lag correlations is concerned, they may be consistently estimated by using the well known method of moments. For $\ell = |u - t|$, $u \neq t$, $u, t = 1, \ldots, T$, the moment estimator for the autocorrelation of lag ℓ, ρ_ℓ, has the formula

$$\hat{\rho}_\ell = \frac{\sum_{i=1}^{K}\sum_{t=1}^{T-\ell}\tilde{y}_{it}\tilde{y}_{i,t+\ell}/K(T-\ell)}{\sum_{i=1}^{K}\sum_{t=1}^{T}\tilde{y}_{it}^2/KT}, \tag{10}$$

(Sutradhar and Kovacevic (2000, Eq. (2.18), Sutradhar (2003)), where $\tilde{y}_{it}$ is the standardized residual, defined as $\tilde{y}_{it} = (y_{it} - \mu_{it})/\{\sigma_{itt}\}^{\frac{1}{2}}$.

The GQL estimating Eq. 8 for β and the moment estimate of ρ_ℓ by (10) are solved iteratively until convergence. The final estimate of β obtained from this iterative process is referred to as the GQL estimate of β, and may be denoted by $\hat{\beta}_{GQL}$. This estimator $\hat{\beta}_{GQL}$ is consistent for β and also highly efficient, the ML estimator being fully efficient which is however impossible or extremely complex to obtain in the correlated data set up.

With regard to the generality of the stationary auto-correlation matrix $C_i(\rho)$ in (9), one may show that this matrix, in fact, represents the correlations of many stationary dynamic such as stationary auto-regressive order 1 (AR(1)), stationary moving average order 1 (MA(1)), and stationary equi-correlations (EQC) models. For example, consider the stationary AR(1) model given by

$$y_{it} = \rho * y_{i,t-1} + d_{it}, \tag{11}$$

(McKenzie (1988), Sutradhar (2003)) where it is assumed that for given $y_{i,t-1}$, $\rho * y_{i,t-1}$ denotes the so-called binomial thinning operation (McKenzie 1988). That is,

$$\rho * y_{i,t-1} = \sum_{j=1}^{y_{i,t-1}} b_j(\rho) = z_{i,t-1}, \text{say}, \tag{12}$$

with $\Pr[b_j(\rho) = 1] = \rho$ and $\Pr[b_j(\rho) = 0] = 1 - \rho$. Furthermore, it is assumed in (11) that y_{i1} follows the Poisson distribution with mean parameter $\mu_{i\cdot}$, that is, $y_{i1} \sim Poi(\mu_{i\cdot})$, where $\mu_{i\cdot} = \exp(x_{i\cdot}'\beta)$ with stationary covariate vector $x_{i\cdot}$ such that $x_{it} = x_{i\cdot}$ for all $t = 1, \ldots, T$. Further, in (11), $d_{it} \sim P(\mu_{i\cdot}(1-\rho))$ and is independent of $z_{i,t-1}$. This model in (11) yields the mean, variance and auto-correlations of the data as shown in Table 1. The Table 1 also contains the MA(1) and EQC models and their basic properties including the correlation structures.

It is clear from Table 1 that the correlation structures for all three processes can be represented by $C_i(\rho)$ in (9). By following Qaqish (2003), one may write similar but different dynamic models for the repeated binary data, with their

Generalized Quasi-Likelihood (GQL) Inferences. Table 1 A class of stationary correlation models for longitudinal count data and basic properties

Model	Dynamic relationship	Mean-variance & correlations
AR(1)	$y_{it} = \rho * y_{i,t-1} + d_{it}, t = 2, \ldots$ $y_{i1} \sim Poi(\mu_{i\cdot})$ $d_{it} \sim P(\mu_{i\cdot}(1-\rho)), t = 2, \ldots$	$E[Y_{it}] = \mu_{i\cdot}$ $\text{var}[Y_{it}] = \mu_{i\cdot}$ $\text{corr}[Y_{it}, Y_{i,t+\ell}] = \rho_\ell$ $= \rho^\ell$
MA(1)	$y_{it} = \rho * d_{i,t-1} + d_{it}, t = 2, \ldots$ $y_{i1} = d_{i1} \sim Poi(\mu_{i\cdot}/(1+\rho))$ $d_{it} \sim P(\mu_{i\cdot}/(1+\rho)), t = 2, \ldots$	$E[Y_{it}] = \mu_{i\cdot}$ $\text{var}[Y_{it}] = \mu_{i\cdot}$ $\text{corr}[Y_{it}, Y_{i,t+\ell}] = \rho_\ell$ $= \begin{cases} \frac{\rho}{1+\rho} & \text{for } \ell = 1 \\ 0 & \text{otherwise,} \end{cases}$
EQC	$y_{it} = \rho * y_{i1} + d_{it}, t = 2, \ldots$ $y_{i1} \sim Poi(\mu_{i\cdot})$ $d_{it} \sim P(\mu_{i\cdot}(1-\rho)), t = 2, \ldots$	$E[Y_{it}] = \mu_{i\cdot}$ $\text{var}[Y_{it}] = \mu_{i\cdot}$ $\text{corr}[Y_{it}, Y_{i,t+\ell}] = \rho_\ell$ $= \rho$

correlation structures represented by $C_i(\rho)$. Thus, if the count or binary data follow this type of auto-correlations model, one may then certainly estimate the regression vector consistently and efficiently by solving the general auto-correlations matrix based GQL estimating Eq. 8, where the lag correlations are estimated by (10) consistently.

About the Author

Brajendra Sutradhar is a University Research Professor at Memorial University in St. John's, Canada. He is an Elected member of the International Statistical Institute and a Fellow of the American Statistical Association. He has published about 110 papers in statistics journals in the area of multivariate analysis, time series analysis including forecasting, sampling, survival analysis for correlated failure times, robust inferences in generalized linear mixed models with outliers, and generalized linear longitudinal mixed models with biostatistical and econometric applications. He has served as an Associate editor for six years for *Canadian Journal of Statistics* and for four years for the *Journal of Environmental and Ecological Statistics*. He has served for 3 years as a member of the advisory committee on statistical methods in Statistics Canada. Professor Sutradhar was awarded 2007 distinguished service award of Statistics Society of Canada for his many years of services to the society including his special services for society's annual meetings.

Cross References

►Likelihood
►Measurement Error Models
►Poisson Regression
►Robust Regression Estimation in Generalized Linear Models

References and Further Reading

McCullagh P (1983) Quasilikelihood functions. Ann Stat 11: 59–67
McKenzie E (1988) Some ARMA models for dependent sequences of Poisson counts. Adv Appl Probab 20:822–835
Qaqish BF (2003) A family of multivariate binary distributions for simulating correlated binary variables with specifed marginal means and correlations. Biometrika 90:455–463
Sutradhar BC (2003) An overview on regression models for discrete longitudinal responses. Stat Sci 18:377–393
Sutradhar BC, Das K (1999) On the efficiency of regression estimators in generalized linear models for longitudinal data. Biometrika 86:459–465
Sutradhar BC, Kovacevic M (2000) Analyzing ordinal longitudinal survey data: generalized estimating equations approach. Biometrika 87:837–848
Wedderburn RWM (1974) Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. Biometrika 61:439–447

Generalized Rayleigh Distribution

Mohammad Z. Raqab[1], Mohamed T. Madi[2]
[1]Professor of Statistics
University of Jordan, Amman, Jordan
[2]Professor
UAE University, Al Ain, United Arab Emirates

The Rayleigh distribution is one of the most popular distributions in analyzing skewed data. The Rayleigh distribution was originally proposed in the fields of acoustics and optics by Lord Rayleigh (or by his less glamorous name J.W. Strutt), way back in 1880, and it became widely known since then in oceanography, and in communication theory for describing instantaneous peak power of received radio

signals. It has received a considerable attention from engineers and physicists for modeling wave propagation, radiation, synthetic aperture radar images, and other related phenomena.

A Rayleigh random variable X has cumulative distribution function (cdf)

$$F(x;\lambda) = 1 - e^{-(\lambda x)^2},\ x \geq 0, \lambda > 0, \tag{1}$$

and probability density function (pdf)

$$f(x;\lambda) = 2\lambda^2 x\, e^{-(\lambda x)^2},\ x \geq 0, \lambda > 0, \tag{2}$$

where λ is an inverse scale parameter. From (1), we obtain immediately the kth raw moment of X to be

$$E(X^r) = \frac{\Gamma\left(\frac{r}{2}+1\right)}{\lambda^r}. \tag{3}$$

Using Eq. 3, we compute the mean, variance, skewness and coefficient of kurtosis, respectively, as:

$$\begin{aligned} &\mu = \sqrt{\frac{\pi}{4}}\ \lambda^{-1} \approx 0.8862\,\lambda^{-1},\ \ \sigma^2 = \left(1-\frac{\pi}{4}\right)\lambda^{-2} \\ &\quad \approx 0.2146\,\lambda^{-2} \\ &\alpha_3 = \frac{2\sqrt{\pi}(\pi-3)}{(4-\pi)^{3/2}} \approx 0.6311,\ \ \alpha_4 = \frac{32-3\pi^2}{(4-\pi)^2} \approx 3.2451. \end{aligned} \tag{4}$$

Since the cdf of the Rayleigh distribution is in closed form, it has been used very effectively for analyzing censored lifetime data. It has a linearly increasing hazard rate given by $h_X(x) = 2\lambda^2 x$. Due to the monotone property of the hazard rate, the Rayleigh distribution has been used as a model for the lifetimes of components that age rapidly with time.

The likelihood function based on a complete sample $X_1, X_2, \ldots, X_n$ of size n from the Rayleigh distribution is

$$L(x_1,\ldots,x_n;\lambda) \propto \lambda^{2n}\left(\prod_{i=1}^{n} x_i\right) e^{-\sum_{i=1}^{n}(\lambda x_i)^2}. \tag{5}$$

Therefore, to obtain the maximum likelihood estimate (MLE) of λ, we can maximize (5) directly with respect to λ and get the MLE of λ as

$$\hat{\lambda} = \left\{\frac{n}{\sum_{i=1}^{n} x_i^2}\right\}^{1/2}.$$

By setting $E(X) = \overline{X}$, we get the method of moment estimator of λ (see (4)) as $\lambda^* = 0.8862/\overline{X}$. From the fact that $2\lambda^2 X_i^2 \sim \chi^2_2$ (▶chi-square distribution with two degrees of freedom), it follows that $2n\lambda^2/\hat{\lambda}^2$ has χ^2_{2n} distribution. This implies that a $100(1-\alpha)\%$ confidence interval for λ is derived to be

$$\left(\sqrt{\frac{\chi^2_{2n}(\alpha/2)\hat{\lambda}^2}{2n}},\ \sqrt{\frac{\chi^2_{2n}(1-\alpha/2)\hat{\lambda}^2}{2n}}\right),$$

where $\chi^2_{2n}(\alpha)$ represents 100αth percentile of the χ^2_{2n} distribution.

Burr (1942) introduced twelve different forms of cumulative distribution functions for modeling data. Among those twelve distribution functions, Burr-Type X and Burr-Type XII received the maximum attention. Recently, Surles and Padgett (2001) considered the two parameter Burr Type X distribution by introducing a shape parameter and correctly named it as the generalized Rayleigh (GR) distribution. If the random variable X has a two parameter GR distribution, then it has the cumulative distribution function (cdf);

$$F(x;\alpha,\lambda) = \left(1-e^{-(\lambda x)^2}\right)^{\alpha};\quad x>0, \alpha>0, \lambda>0,$$

and probability density function (pdf)

$$\begin{aligned} f(x;\alpha,\lambda) = 2\alpha\lambda^2 x e^{-(\lambda x)^2}\left(1-e^{-(\lambda x)^2}\right)^{\alpha-1}; \\ x>0, \alpha>0, \lambda>0, \end{aligned} \tag{6}$$

where α and λ are shape and inverse scale parameters, respectively. We denote the GR distribution with shape parameter α and inverse scale parameter λ as GR(α,λ). Several aspects of the one-parameter ($\lambda = 1$) Burr-Type X distribution were studied by Ahmad et al. (1997), Raqab (1998) and Surles and Padgett (2001). When $\alpha = 1$, the GR distribution reduces to the one-parameter Rayleigh distribution with cdf and pdf given in (1) and (2).

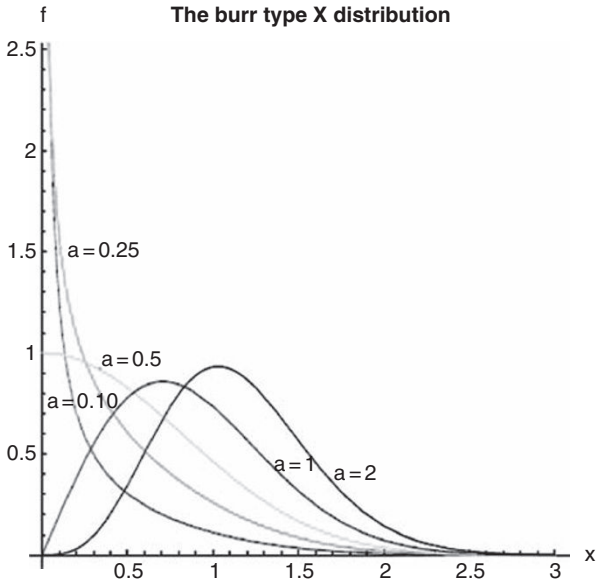
If $\alpha \leq \frac{1}{2}$, the density function in (6) is a decreasing function and for $\alpha > \frac{1}{2}$, it is a right skewed unimodal function. The mode of the density function is equal to $\frac{x_0}{\lambda}$, where x_0 is the solution of the non-linear equation

$$1 - 2x^2 - e^{-x^2}(1-2\alpha x^2) = 0.$$

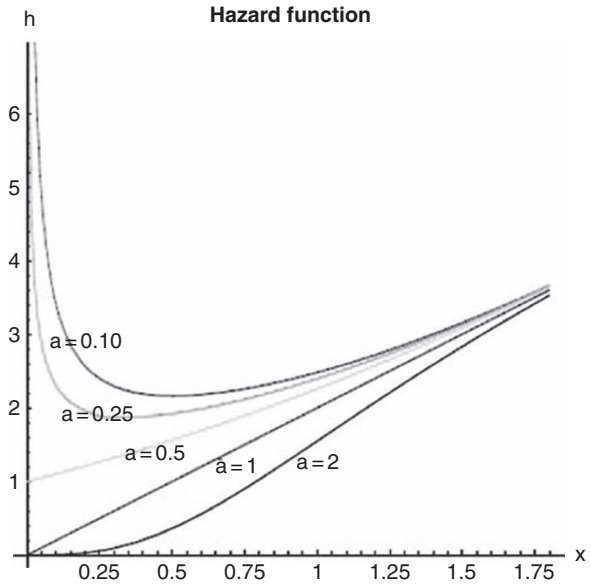
Clearly the mode is a decreasing function of λ as expected and it is an increasing function of α. Different forms of the density functions are presented in Fig. 1. It is clear from Fig. 1 that the GR density functions resemble the gamma and Weibull density functions. The median of a GR random variable occurs at $\left[-\frac{1}{\lambda}\ln\left(1-\left(\frac{1}{2}\right)^{\frac{1}{\alpha}}\right)\right]^{\frac{1}{2}}$ and is also a decreasing function of λ but an increasing function of α.

The hazard function of X is given by

$$h(x;\alpha,\lambda) = \frac{2\alpha\lambda^2 x e^{-(\lambda x)^2}\left(1-e^{-(\lambda x)^2}\right)^{\alpha-1}}{1-\left(1-e^{-(\lambda x)^2}\right)^{\alpha}}.$$



Generalized Rayleigh Distribution. Fig. 1 The density functions of the GR distribution for different shape parameters



Generalized Rayleigh Distribution. Fig. 2 The hazard functions of the GR distribution for different shape parameters

If $\alpha = 1$, the hazard function becomes $2\lambda^2 x$, a linear function of x. From Mudholkar et al. (1995), it follows that if $\alpha \leq \frac{1}{2}$, the hazard function of $GR(\alpha, \lambda)$ is bathtub type and for $\alpha > \frac{1}{2}$, it is increasing.

The hazard functions for different values of α are plotted in Fig. 2. For $\alpha \leq \frac{1}{2}$, it decreases from ∞ to a positive constant and then it increases to ∞. For $\alpha > \frac{1}{2}$, it is an increasing function and it increases from 0 to ∞. It is known that for shape parameter greater than 1, the hazard functions of gamma and Weibull are all increasing functions. The hazard function of ▶gamma distribution increases from 0 to 1. While for ▶Weibull distribution it increases from 0 to ∞. For $\alpha > \frac{1}{2}$, the hazard function of the GR distribution behaves like the hazard function of the Weibull distribution, whose shape parameter is greater than 1. In this respect the GR distribution behaves more like a Weibull distribution than gamma distribution. Therefore, if the data are coming from an environment where the failure rate is gradually increasing without any bound, the GR distribution can also be used instead of a Weibull distribution. As indicated in Fig. 1, the GR density functions are always right skewed and they can be used quite effectively to analyze skewed data sets.

In the context of estimating the model parameters, let $x_1, x_2, \ldots, x_n$ be a random sample of size n from $GR(\alpha, \lambda)$, then the log-likelihood function $L(\alpha, \lambda)$ can be written as:

$$L(\alpha, \lambda) \propto n\, ln\alpha + 2n\, ln\lambda + \sum_{i=1}^{n} ln\, x_i - \lambda^2 \sum_{i=1}^{n} x_i^2 + (\alpha - 1) \sum_{i=1}^{n} ln(1 - e^{-(\lambda x_i)^2}). \quad (7)$$

The normal equations become:

$$\frac{\partial L}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} ln(1 - e^{-(\lambda x_i)^2}) = 0, \quad (8)$$

$$\frac{\partial L}{\partial \lambda} = \frac{n}{\lambda} - \lambda \sum_{i=1}^{n} x_i^2 + \lambda(\alpha - 1) \sum_{i=1}^{n} \frac{x_i^2\, e^{-(\lambda x_i)^2}}{ln(1 - e^{-(\lambda x_i)^2})} = 0. \quad (9)$$

From (8), we obtain the MLE of α as a function of λ, say $\hat{\alpha}(\lambda)$, as

$$\tilde{\alpha}(\lambda) = \frac{n}{\sum_{i=1}^{n} -ln(1 - e^{-(\lambda x_i)^2})} = 0.$$

Substituting $\tilde{\alpha}(\lambda)$ in (7), we obtain the profile function $g(\lambda) = L(\tilde{\alpha}(\lambda), \lambda)$. By setting $\lambda^2 = \mu$, the MLE of λ, say $\tilde{\lambda}$ can be obtained as a fixed point solution of the following equation

$$h(\mu) = \mu, \quad (10)$$

where

$$h(\mu) = \left[\frac{\sum_{i=1}^{n} \frac{x_i^2 e^{-\mu x_i^2}}{1 - e^{-\mu x_i^2}}}{\sum_{i=1}^{n} ln(1 - e^{-\mu x_i^2})} + \frac{1}{n} \sum_{i=1}^{n} x_i^2 + \frac{1}{n} \sum_{i=1}^{n} \frac{x_i^2\, e^{-\mu x_i^2}}{1 - e^{-\mu x_i^2}} \right]^{-1}$$

If $\tilde{\mu}$ is a solution of (10), then $\tilde{\lambda} = \sqrt{\tilde{\mu}}$. Very simple iterative procedure can be used to solve (10).

To obtain the Bayesian estimates of the two parameters, we assume that α and λ are independent and that

$$\alpha \sim G(a_0, b_0) \quad \text{and} \quad \lambda \sim GEP(a_1, b_1), \tag{11}$$

where $G(a, b)$ denotes the gamma distribution with mean $\frac{a}{b}$, and $GEP(a_1, b_1)$ denotes the generalized exponential power distribution with density function

$$\pi(\lambda) \propto \lambda^{2a_1-1} e^{-b_1\lambda^2} I_{(\lambda>0)}$$

where a_0, b_0, a_1, b_1 are chosen to reflect prior knowledge about α and λ.

The likelihood function of α and λ for the given complete sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ can be expressed as:

$$L(\alpha, \lambda|\mathbf{x}) \propto \alpha^n \lambda^{2n} \exp\{n \overline{\ln x} - n\lambda^2 \overline{x^2} - (\alpha - 1)D_n(\lambda)\}, \tag{12}$$

where $D_n(\lambda) = -\sum_{i=1}^n \ln(1 - e^{-(\lambda x_i)^2})$, $\overline{\ln x} = \frac{1}{n}\sum_{i=1}^n \ln x_i$ and $\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$. By combining (11) and (12), we obtain the joint posterior density of α and λ

$$\pi(\alpha, \lambda\,|\mathbf{x}) \propto h_\lambda(a_1 + n, b_1 + n\overline{x^2})\, g_\alpha(a_0 + n, D_n(\lambda) + b_0) \exp\{D_n(\lambda)\},$$

where g_α denotes the gamma density for α and h_λ denotes the generalized exponential power density for λ. The marginal posterior density of λ is given by

$$\pi(\lambda\,|\mathbf{x}) \propto h_\lambda(a_1 + n, b_1 + n\,\overline{x^2})\; W(\lambda), \tag{13}$$

where $W(\lambda) = (D_n(\lambda) + b_0)^{-(a_0+n)} \exp\{D_n(\lambda)\}$. From (13) and using importance sampling, we can express the Bayes estimate of λ as

$$E(\lambda|x) = \frac{E^{(1)}[\lambda\; W(\lambda)]}{E^{(1)}[W(\lambda)]},$$

where $E^{(1)}$ denotes the expectation with respect to $GEP(a_1 + n, b_1 + n\;\overline{x^2})$.

Since the marginal posterior density of α given λ and $\mathbf{x}$ is $g_\alpha(a_0 + n, D_n(\lambda) + b_0)$, the marginal posterior of α is equal to $E_{\lambda|\mathbf{x}}[g_\alpha(a_0 + n,\, D_n(\lambda) + b_0)]$, it follows that

$$\pi(\alpha|x) = \frac{E^{(1)}[W(\lambda)\, g_\alpha(a_0 + n, D_n(\lambda) + b_0)]}{E^{(1)}[W(\lambda)]}.$$

Using the fact that $E(\alpha\,|\,\lambda, \mathbf{x}) = (a_0 + n)/(D_\lambda + b_0)$, we obtain the Bayes estimate of α as

$$E(\alpha|x) = \frac{E^{(1)}[W(\lambda)\,(a_0 + n)/(D_n(\lambda) + b_0)]}{E^{(1)}[W(\lambda)]}.$$

If X follows GR(α, λ), then

$$E(X^2) = \frac{1}{\lambda^2}\left(\psi(\alpha + 1) - \psi(1)\right),$$

and

$$E(X^4) - (E(X^2))^2 = \frac{1}{\lambda^4}\left(\psi'(1) - \psi'(\alpha + 1)\right).$$

Here $\psi(.)$ and $\psi'(.)$ denote the digamma and polygamma functions, respectively. Let us define U and V as follows:

$$U = \frac{1}{n}\sum_{i=1}^n x_i^2, \quad V = \frac{1}{n}\sum_{i=1}^n x_i^4 - U^2.$$

The method of moment's estimator (MME) of α can be obtained as the solution of the following non-linear equation:

$$\frac{V}{U^2} = \frac{\psi'(1) - \psi'(\alpha + 1)}{(\psi(\alpha + 1) - \psi(1))^2}.$$

We denote the estimate of α as $\hat{\alpha}_{MME}$. Once $\hat{\alpha}_{MME}$ is obtained, we obtain the MME of λ, say $\hat{\lambda}_{MME}$ as

$$\hat{\lambda}_{MME} = \sqrt{\frac{\psi(\hat{\alpha}_{MME} + 1) - \psi(1)}{U}}.$$

It is not possible to obtain exact variances of $\hat{\alpha}_{MME}$ and $\hat{\lambda}_{MME}$. The asymptotic variances of $\hat{\alpha}_{MME}$ and $\hat{\lambda}_{MME}$ can be obtained from the normality asymptotic property of these estimates.

For other methods of estimation, one may refer to Kundu and Raqab (2005).

About the Authors

For biography of Mohammad Z. Raqab *see* the entry ►Ordered Statistical Data: Recent Developments.

For biography of Mohamed T. Madi *see* the entry ►Step-Stress Accelerated Life Tests.

Cross References

►Multivariate Statistical Distributions

►Weibull Distribution

References and Further Reading

Ahmad KE, Fakhry ME, Jaheen ZF (1997) Empirical Bayes estimation of P(Y < X) and characterization of Burr-type X model. *J Stat Plan Infer* 64:297–308

Burr IW (1942) Cumulative frequency distribution. Ann Math Stat 13:215–232

Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distribution, vol 1, 2nd edn. Wiley, New York

Kundu D, Raqab MZ (2005) Generalized rayleigh distribution: different methods of estimation. Comput Stat Data Anal 49: 187–200

Mudholkar GS, Srivastava DK, Freimer M (1995) The exponentiated Weibull family: a reanalysis of the bus motor failure data. Technometrics 37:436–445

Raqab MZ (1998) Order statistics from the Burr type X model. Comput Math Appl 36:111–120

Rayleigh JWS (1880) On the resultant of a large number of vibrations of the same pitch and of arbitrary phase. Philos Mag 5th Series 10:73–78

Surles JG, Padgett WJ (2001) Inference for reliability and stress-strength for a scaled Burr Type X distribution. Lifetime Data Anal 7:187–200

Generalized Weibull Distributions

Chin Diew Lai
Professor in Statistics
Massey University, Palmerston North, New Zealand

Introduction

The ▶Weibull distribution has been found very useful in fitting reliability, survival and warranty data and thus it is one of the most important continuous distributions in applications. A drawback of the Weibull distribution as far as lifetime analysis is concerned, is the monotonic behavior of its hazard (failure) rate function. In real life applications, empirical hazard rate curves often exhibit non-monotonic shapes such as a bathtub, upside-down bathtub (unimodal) and others. Thus there is a genuine desire to search for some generalizations or modifications of the Weibull distribution that can provide more flexibility in lifetime modelling.

Let T be the lifetime random variable with $f(t)$, $F(t)$ being its probability density function (pdf) and cumulative distribution function (cdf), respectively.

The hazard rate (failure rate) function is defined as

$$h(t) = \frac{f(t)}{1-F(t)} = \frac{f(t)}{R(t)}, \tag{1}$$

where $R(t) = 1 - F(t)$ is the reliability or survival function of T. The cumulative hazard rate function is defined as

$$H(t) = \int_0^t h(x)dx. \tag{2}$$

It is easy to show that the reliability function can be expressed as

$$R(t) = e^{-H(t)}. \tag{3}$$

It is easy to see that the cumulative hazard rate function completely determines the lifetime distribution and it must satisfies the following three conditions in order to yield a proper lifetime distribution:

(I) $H(t)$ is nondecreasing for all $t \geq 0$
(II) $H(0) = 0$
(III) $\lim_{t\to\infty} H(t) = \infty$.

We will see that (3) provides a convenient and important tool to construct Weibull-type lifetime distributions. Lai and Xie (2006, Chapter 3) gives a comprehensive account on Weibull related distributions.

Standard Weibull Distribution

The standard Weibull distribution is given by

$$R(t) = \exp(-\lambda t^{\alpha}), \quad \lambda, \alpha > 0;\ t \geq 0. \tag{4}$$

It follows from (3) that $H(t) = \lambda t^{\alpha}$. By a simple differentiation, we obtain the hazard rate function $h(t) = \lambda\alpha t^{\alpha-1}$, which is increasing (decreasing) if $\alpha > 1$ ($\alpha < 1$). We now see that despite its many applications, the Weibull distribution lacks flexibility for many reliability applications. For other properties of the Weibull distribution, we refer our readers to Murthy et al. (2003) for details.

Generalizations

There are many ways to generalize a Weibull distribution. Indeed, we now have many such generalizations or extensions in the literature. There is no clear guideline upon which one may identify a distribution as a generalized Weibull distribution.

On the basis of how a lifetime distribution is generated, we now attempt to classify generalized Weibull distributions into seven classes although these classes are not necessarily mutually exclusive.

C1: Distributions arise from a transformation of the Weibull random variable X, such as (a) linear transformation, (b) power transformation, (c) non-linear transformation; (d) log transformation or (e) inverse transformation. For example, $Y = \log X$, then Y is a log Weibull (also known as the type 1 extreme value distribution). If $Y = X^{-1}$, then Y has an inverse Weibull distribution with $R(t) = 1 - \exp\{-\lambda t^{-\alpha}\}$, $t > 0$.

C2: Families of generalized Weibull distributions with two or more parameters that contains (4) as a special case. For example, the modified Weibull of Lai et al. (2003) with

$$R(t) = \exp\{-at^{\alpha}e^{\lambda t}\}, \lambda \geq 0,\ \alpha, a > 0,\ t \geq 0 \tag{5}$$

is such an example. For moments of the above distribution, see Nadarajah (2005).

C3: Families of distributions that converge to a standard Weibull or a generalized Weibull in C2 when one of their parameters tends to zero. For example, The generalized Weibull of Modhalkar et al. (1996) defined by survival function:

$$R(t) = 1 - \left[1 - \left(1 - \lambda\left(\frac{t}{\beta}\right)^{\alpha}\right)^{1/\lambda}\right], \alpha, \beta > 0;\ t \geq 0. \tag{6}$$

As $\lambda \to 0$, $R(t) \to \exp\left\{-\left(\frac{t}{\beta}\right)^{\alpha}\right\}$. Another example, the beta integrated distribution of Lai et al. (1998) with $R(t) = \exp\{-\alpha t^{\gamma}(1-dt)^{c}\}$. Set $c = \lambda/d$ and let $d \to 0$ we obtain the distribution given by (5).

C4: Power transformations of either the cumulative distribution function or the survival function of the Weibull or a generalized Weibull. For example, the exponentiated Weibull of Mudholkar and Srivastava (1993) and the generalized modified Weibull of Carrasco et al. (2008) are the two prime examples.

C5: The survival function of a generalized Weibull is a function of the survival function of the Weibull. For example, the distribution of Marshall and Olkin (1997) and the distribution of Hjorth (1980).

C6: Involving two or more Weibull distributions: (a) finite mixtures; (b) n-fold competing risk (equivalent to independent components being arranged in a series structure); (c) n-fold multiplicative models; and (d) n-fold sectional models. See Murthy et al. (2003) for further details. The class also includes mixtures of two different generalized Weibulls, for example, Bebbington et al. (2007b).

C7: The class of distributions that are expressed as in (3) with the cumulative hazard rate function having a simple form. Of course, this class contains many generalized Weibull distributions that appear in the literature.

Some Important Generalized Weibull Families

We now present several families that have relatively simple survival functions. In addition, they can give rise to non-monotonic hazard rate functions of various shapes such as bathtub, upside-down bathtub (unimodal), or a modified bathtub.

Generalized Modified Weibull Family

The distribution studied by Carrasco et al. (2008) has the survival function given by

$$R(t) = 1-(1-\exp\{-at^{\alpha}e^{\lambda t}\})^{\beta},\ \lambda \geq 0,\ \alpha, a > 0;\ \beta > 0;\ t \geq 0. \tag{7}$$

Clearly, it is a simple extension of the modified Weibull distribution of Lai et al. (2003) since (7) reduces to (5) when $\beta = 1$. In fact, it includes several other distributions such as type 1 extreme value, the exponentiated Weibull of Mudholkar and Srivastava (1993) as given in (11) below, and others. An important feature of this lifetime distribution is its flexibility in providing hazard rates of various shapes.

Generalized Weibull-Gompertz Distribution

Nadarajah (2005) proposed a generalization of Weibull with four parameters having survival function given as below:

$$R(t) = \exp\{-at^{b}(e^{ct^{d}} - 1)\},\ a, d > 0;\ b, c \geq 0;\ t \geq 0. \tag{8}$$

Since it includes the Gompertz (or Gompertz-Makem) as its special case when $b = 0$, we may refer it as the generalized Weibull-Gompertz distribution. Clearly, it contains several distributions listed in Table 1 of Pham and Lai (2007). Again, it can prescribe increasing, decreasing or bathtub shaped hazard rate functions.

Generalized Power Weibull Family

Nikulin and Haghighi (2006) proposed a three-parameter family lifetime distributions

$$R(t) = \exp\{1-(1+(t/\beta)^{\alpha})^{\theta}\}, t \geq 0; \alpha, \beta > 0; \theta > 0. \tag{9}$$

Its hazard rate function $h(t)$ can give rise to increasing, decreasing, bathtub or upside-down bathtub shapes.

Flexible Weibull Distribution

Bebbington et al. (2007a) obtained a generalization of the Weibull having a simple and yet flexible cumulative hazard rate function H:

$$R(t) = \exp\{-(e^{\alpha t-\beta/t})\}; \alpha, \beta > 0; t \geq 0. \tag{10}$$

It was shown that F has an increasing hazard rate if $\alpha\beta > 27/64$ and a modified bathtub (N or roller-coaster shape) hazard rate if $\alpha\beta \leq 27/64$. Note that there are few generalized Weibull distributions that have this shape.

Exponentiated Weibull Family

Mudholkar and Srivastava (1993) proposed a simple generalization of Weibull by raising the cdf of the Weibull to the power of θ giving

$$R(t) = 1-[1-\exp(-t/\beta)^{\alpha}]^{\theta},\ t \geq 0;\ \alpha, \beta > 0,\ \theta \geq 0. \tag{11}$$

The distribution is found to be very flexible for reliability modelling as it can model increasing (decreasing), bathtub (upside-down) shaped hazard rate distributions.

About the Author

Chin-Diew Lai is Professor, Institute of Fundamental Sciences-Statistics, Massey University, Palmerston North, New Zealand. He has published 114 journal articles, 8 book chapters and 4 books, including the text (with

N. Balakrishnan) *Continuous Bivariate Distributions* (2nd ed. Springer, New York, 2009). He is Associate Editor for *Communications in Statistics* (2009–present), and Associate Editor of the *Journal of Applied Mathematics and Decision Sciences* (2007–present).

Cross References

►Extreme Value Distributions
►Generalized Rayleigh Distribution
►Modeling Survival Data
►Multivariate Statistical Distributions
►Statistical Distributions: An Overview
►Statistics of Extremes
►Step-Stress Accelerated Life Tests
►Survival Data
►Weibull Distribution

References and Further Reading

Bebbington M, Lai CD, Zitikis R (2007a) A flexible Weibull extension. Reliab Eng Syst Saf 92:719–726

Bebbington M, Lai CD, Zitikis R (2007b) Modeling human mortality using mixtures of bathtub shaped failure distributions. J Theor Biol 245:528–538

Carrasco JMF, Ortega EMM, Cordeiro GM (2008) A generalized modified Weibull distribution for lifetime modelling. Comput Stat Data Anal 53:450–462

Hjorth U (1980) A reliability distribution with increasing, decreasing, constant and bathtub-shaped failure rate. Technometrics 22:99–107

Lai CD, Xie M (2006) Stochastic ageing and dependence for reliability. Springer, New York

Lai CD, Xie M, Moore T (1998) The beta integrated model. In: Xie M, Murthy DNP (eds) Proceedings of the international workshop on reliability modelling and analysis - from theory to practice, National University of Singapore, pp 153–159

Lai CD, Xie M, Murthy DNP (2003) Modified Weibull model. IEEE Trans Reliab 52:33–37

Marshall AW, Olkin I (1997) A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. Biometrika 84:641–652

Mudholkar G, Srivastava DK (1993) Exponentiated Weibull family for analyzing bathtub failure-rate data. IEEE Trans Reliab 42:299–302

Mudholkar GS, Srivastava DK, Kollia GD (1996) A generalization of the Weibull distribution with application to analysis of survival data. J Am Stat Assoc 91:1575–1583

Murthy DNP, Xie M, Jiang R (2003) Weibull models. Wiley, New York

Nadarajah S (2005) On the moments of the modified Weibull distribution. Reliab Eng Syst Saf 90:114–117

Nikulin M, Haghighi F (2006) A chi-squared test for the generalized power Weibull family for the head-and-neck cancer censored data. J Math Sci 133(3):1333–1341

Pham H, Lai CD (2007) On recent generalizations of the Weibull distribution. IEEE Trans Reliab 56(3):454–459

Geometric and Negative Binomial Distributions

Adrienne W. Kemp
Honorary Senior Lecturer
University of St. Andrews, St. Andrews, UK

Introduction

The values of the probability mass function (pmf) for the *geometric* distribution are in geometric progression:

$$\Pr[X = x] = pq^x, \quad 0 < p < 1,\ q = 1 - p, \quad x = 0, 1, 2, \ldots. \tag{1}$$

It is the waiting time distribution that was studied by Pascal (1679) and Montmort (1713) concerning the number of failures (tails) in a sequence of throws of a coin before obtaining the first success (head) (p is the probability of a head).

The distribution is important in Markov chain models (see ►Markov Chains), e.g., in meteorological models for precipitation amounts and for weather cycles, in the estimation of animal abundance, in the analysis of runs of one botanical species in transects through mixtures of plants, and in surveillance for congenital malformations. Here it is sometimes called the *Furry* distribution.

The sum of $k = 1, 2, 3, \ldots$ geometric random variables, for example the waiting time for k heads in a sequence of throws of a coin, gives a *negative binomial* random variable. This approach enabled Montmort (1713) to solve the Problem of Points which was of intense interest in gambling circles at the time. Meyer (1879) used it to find the probability of x male births in a sequence of births containing a fixed number k of female births, assuming that the probability q of a male birth is constant. The distribution is sometimes called the *Pascal* distribution or *binomial waiting-time* distribution when k is an integer and the distribution is shifted k units from the origin to support $k, k + 1, \ldots$.

The pmf of the negative binomial distribution is

$$\Pr[X = x] = \binom{k + x - 1}{k - 1} p^k q^x, \quad 0 < p < 1,\ q = 1 - p, \quad x = 0, 1, 2, \ldots. \tag{2}$$

Whereas binomial pmf's are given by the terms of the binomial expansion $(1 - \pi + \pi)^n$, negative binomial pmf's are given by the terms of the negative binomial expansion

$p^k(1-q)^{-k} = p^k\left[1+kq+\frac{k(k+1)}{2}q^2+\cdots\right]$. The integer restriction on k is not necessary, provided that $k>0$.

A popular alternative parameterization takes $P=q/p$, i.e., $p=1/(1+P)$, giving the pmf

$$\Pr[X=x]=\binom{k+x-1}{k-1}\left(\frac{1}{1+P}\right)^k\left(\frac{P}{1+P}\right)^x, \quad P>0,\ k>0,\quad x=0,1,2,\ldots. \tag{3}$$

The probability generating function (pgf) is

$$G(z)=\left(\frac{1-q}{1-qz}\right)^k=(1+P-Pz)^{-k} \tag{4}$$

and the mean and variance are

$$\mu=k(1-p)/p=kP, \qquad \mu_2=k(1-p)/p^2=kP(1+P)>\mu. \tag{5}$$

The distribution is overdispersed. Taking $k=1$ gives the corresponding formulae for the geometric distribution.

A parameterization used in the ecological literature is $a=P, m=kP$, giving the pgf $(1+a-az)^{-m/a}$, and $\mu=m$, $\mu_2=m(1+a)$. This is equivalent to Cameron and Trivedi's (1986) NB1 model with $\lambda=kP$, $P=P$ giving $\mu=\lambda$, $\mu_2=\lambda(1+P)$, i.e., the variance is a linear function of the mean. For Cameron and Trivedi's NB2 model, $\lambda=kP$, $k=k$, giving $\mu=\lambda$, $\mu_2=\lambda+\lambda^2/k$, i.e., the variance is a quadratic function of the mean. The NB1 and NB2 models are used by econometricians for negative binomial regression when ▶Poisson regression for large data sets with explanatory variables is inadequate because it cannot handle overdispersion.

Derivations

The negative binomial distribution is both a mixed Poisson and a generalized Poisson distribution. Consider with Greenwood and Yule (1920) a mixture of Poisson distributions whose parameter θ has a ▶gamma distribution with probability density function

$$f(\theta)=\{\beta^\alpha\Gamma(\alpha)\}^{-1}\theta^{\alpha-1}\exp(-\theta/\beta), \quad \theta>0,\alpha>0,\beta>0.$$

Then the outcome is a negative binomial distribution

$$\Pr[X=x]=\{\beta^\alpha\Gamma(\alpha)\}^{-1}\int_0^\infty \theta^{\alpha-1}e^{-\theta/\beta}(\theta^x e^{-\theta}/x!)\,d\theta = \binom{\alpha+x-1}{\alpha-1}\left(\frac{\beta}{\beta+1}\right)^x\left(\frac{1}{\beta+1}\right)^\alpha. \tag{6}$$

The Poisson parameter θ represents the expected number of accidents for an individual; this is assumed to vary between individuals according to a gamma distribution. This "contagion" model has been applied, for example, to car accidents, personal injuries, aircraft failures, medical visits during spells of illness.

In the generalized Poisson derivation of Lüders (1934) and Quenouille (1949) the negative binomial distribution arises from the sum Y of a Poisson (θ) number N of independent random variables $X_1, X_2, \ldots, X_N$, all having the same logarithmic distribution with parameter λ. The pgf of Y is then

$$G(s)=\exp\left[\theta\left\{\frac{\ln(1-\lambda z)}{\ln(1-\lambda)}-1\right\}\right]=\left(\frac{1-\lambda}{1-\lambda z}\right)^{-\theta/\ln(1-\lambda)}. \tag{7}$$

This is called an *Arfwedson* process, also a *Poisson sum* (*Poisson-stopped sum*) of logarithmic rv's. It incorporates heterogeneity and has been used extensively to analyse ecological data, e.g., larval counts, plant densities, migration data.

A limiting form of the Pólya-Eggenberger urn model also yields the negative binomial distribution.

In ▶queueing theory the geometric distribution is the equilibrium queue-length distribution for the M/M/1 queue. The negative binomial is the equilibrium queue-length distribution for the M/M/1 queue given Bhat's (2002) form of balking.

Certain important ▶stochastic processes give rise to the negative binomial distribution, e.g., Yule's (1925) simple birth process with nonzero initial population, Kendall's (1948) simple birth-death-and-immigration process with zero initial population, McKendrick's (1914) time-homogeneous birth-and-immigration process with zero initial population, Lundberg's (1940) nonhomogeneous process with zero initial population known as the *Pólya* process, and Kendall's (1948) nonhomogeneous birth-and-death process with zero death rate.

Because it arises in so many ways the negative binomial distribution is often used to analyse overdispersed data when no particular derivation is agreed upon, e.g., haemocytometer counts, consumer expenditure, product choice, lost sales, lending-library data, family sizes, DNA adduct counts. Given a good empirical fit to data, interpretation of the fit can be problematical.

Properties

The characteristic function (cf) is

$$G(e^{it})=\left(\frac{1-q}{1-qe^{it}}\right)^k=(1+P-Pe^{it})^{-k}. \tag{8}$$

The factorial ▶moment generating function is $(1-Pt)^{-k}$, giving

$$\mu'_{[r]} = \frac{(k+r-1)!}{(k-1)!}P^r, \quad r = 1, 2, \ldots, \tag{9}$$

and the factorial cumulant generating function is $-k\ln(1-Pt)$, giving $\kappa_{[r]} = k(r-1)!P^r$, $r = 1, 2, \ldots$. Replacement of n by $(-k)$ and π by $(-P)$ in the formulae for the moment properties of the binomial distribution gives the corresponding negative binomial formulae.

The uncorrected moment generating function is $p^k(1-qe^t)^{-k}$, the central moment generating function is $e^{-kqt/p}p^k(1-qe^t)^{-k}$, and the cumulant generating function is $k\ln p - k\ln(1-qe^t)$.

In particular

$$\begin{aligned}
\mu &= \kappa_1 = kP, \quad \mu_2 = \kappa_2 = kP(1+P) = \frac{kq}{p^2},\\
\mu_3 &= \kappa_3 = kP(1+P)(1+2P) = \frac{kq(1+q)}{p^3},\\
\mu_4 &= 3k^2P^2(1+P)^2 + kP(1+P)(1+6P+6P^2)\\
&= \frac{3k^2q^2}{p^4} + \frac{kq(p^2+6q)}{p^4},\\
\sqrt{\beta_1} &= \frac{1+2P}{\{kP(1+P)\}^{1/2}} = \frac{1+q}{\sqrt{kq}},\\
\beta_2 &= 3 + \frac{(1+6P+6P^2)}{kP(1+P)} = 3 + \frac{p^2+6q}{kq},
\end{aligned} \tag{10}$$

The index of dispersion is $p^{-1} = 1+P$; the coefficient of variation is $(kq)^{-1/2} = \{(1+P)/(kP)\}^{1/2}$.

From

$$\frac{\Pr[X=x+1]}{\Pr[X=x]} = \frac{(k+x)P}{(x+1)Q}, \quad Q = 1+P \tag{11}$$

it follows that $\Pr[X=x+1] < \Pr[X=x]$ if $x > kP-Q$, and $\Pr[X=x] \geq \Pr[X=x-1]$ if $x \leq kP-P$. Thus if $(k-1)P$ is not an integer, then there is a single mode at X where X is the integer part of $(k-1)P$; when $(k-1)P$ is an integer, then there are two equal modes at $X = (k-1)P$ and $X = kP-Q$. The mode is at $X = 0$ when $kP < Q$, i.e., $kq < 1$.

The generalized Poisson derivation implies that the distribution is infinitely divisible.

If $k<1$, then $p_x p_{x+2}/p_{x+1}^2 > 1$ (where $p_x = \Pr[X=x]$) and the probabilities are log-convex; if $k > 1$, then $p_x p_{x+2}/p_{x+1}^2 < 1$ and the probabilities are log-concave. These log-convexity/log-concavity properties imply that the distribution has a decreasing failure (hazard) rate for $k < 1$ and an increasing failure (hazard) rate for $k > 1$. For $k = 1$ the failure rate is constant. This no-memory (Markovian, non-aging) property characterizes the geometric distribution, making it a discrete analogue of the (continuous) exponential distribution. It is important in reliability theory; also it enables geometric random variables to be computer generated easily. For other characterizations of the geometric distribution and the comparatively few characterizations of the negative binomial distribution see Johnson et al. (2005).

If X_1 and X_2 are independent negative binomial random variables with the same series parameter q and power parameters k_1 and k_2, then $X_1 + X_2$ has a negative binomial distribution with pgf $(1+P-Pz)^{-k_1-k_2} = p^{k_1+k_2}(1-qz)^{-k_1-k_2}$. When $k \to \infty$ and $P \to 0$, with $kP = \theta$ constant, the negative binomial distribution tends to a Poisson distribution with parameter θ. It tends to normality when $kP \to \infty$.

Estimation

The parameter of the geometric distribution is easy to estimate; here the first moment equation is also the maximum likelihood equation and $\hat{P} = (1-\hat{p})/\hat{p} = \bar{x}$.

When both parameters of the negative binomial distribution are unknown there is a choice of estimation methods. For the *Method of Moments* the sample mean $\bar{x}$ and sample variance s^2 are equated to their population values, giving $\tilde{k}\tilde{P} = \bar{x}$ and $\tilde{k}\tilde{P}(1+\tilde{P}) = \sum(x-\bar{x})^2/(n-1) = s^2$, i.e., $\tilde{P} = s^2/\bar{x} - 1$ and $\tilde{k} = \bar{x}^2/(s^2-\bar{x})$.

In the *Method of Mean-and-Zero-Frequency* the observed and expected numbers of zero values are used as well as the first moment equation. Let f_0 be the proportion of zero values. Then the equations are $f_0 = (1+P^\dagger)^{-k^\dagger}$ and $k^\dagger P^\dagger = \bar{x}$, giving $P^\dagger/\ln(1+P^\dagger) = -\bar{x}/\ln f_0$; this can be solved for $P^\dagger$ by iteration.

The *Method of Maximum Likelihood* equations are

$$\hat{k}\hat{P} = \bar{x} \text{ and } \ln(1+\hat{P}) = \sum_{j=1}^{\infty}\left\{(\hat{k}+j-1)^{-1}\sum_{i=j}^{\infty}f_j\right\}; \tag{12}$$

their solution requires iteration but this may be slow if the initial estimates are poor. The importance of the negative binomial distribution has led to much research on its inference in the past 2 decades. This includes rapid estimation methods for the initial estimates, the properties of the maximum likelihood estimators, Bayesian estimation, computer studies of the relative efficiences of different estimation methods, and goodness-of-fit tests; see Johnson et al. (2005).

About the Author

For biography *see* the entry ▶Univariate Discrete Distributions – Overview.

Cross References

►Binomial Distribution
►Distributions of Order K
►Insurance, Statistics in
►Inverse Sampling
►Modeling Count Data
►Poisson Regression
►Random Permutations and Partition Models
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Statistical Methods in Epidemiology
►Univariate Discrete Distributions: An Overview

References and Further Reading

Bhat UN (2002) Elements of applied stachastic processes, 3rd edn. Wiley, New York
Cameron AC, Trivedi PK (1986) Econometric models based on count data: comparisons and applications of some estimators. J Appl Econom 1:29–53
de Montmort PR (1713) Essai d'analyse sur les jeux de hasard, 2nd edn. Quillau, Paris (Reprinted by Chelsea, New York, 1980)
Greenwood M, Yule GU (1920) An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. J R Stat Soc A 83: 255–279
Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, Hoboken
Kendall DG (1948) On some modes of population growth leading to R. A. Fisher's logarithmic series distribution. Biometrika 35:6–15
Kendall DG (1949) Stochastic processes and population growth. J R Stat Soc B 11:230–282
Lüders R (1934) Die statistik der seltenen Ereignisse. Biometrika 26:108–128
Lundberg O (1940) On random processes and their application to sickness and accident statistics. Almqvist & Wicksells, Uppsala (Reprinted 1964)
McKendrick AG (1914) Studies on the theory of continuous probabilities, with special reference to its bearing on natural phenomena of a progressive nature. Proc London Math Soc 13(2): 401–416
McKendrick AG (1926) Applications of mathematics to medical problems. Proc Edinburgh Math Soc 44:98–130
Meyer A (1879) Vorlesungen über Wahrscheinlichkeitsrechnung (trans: Czuber E). Teubner, Leipzig
Pascal B (1679) Varia opera mathematica D. Petri de Fermat, Tolossae
Quenouille MH (1949) A relation between the logarithmic, Poisson, and negative binomial series. Biometrics 5:162–164
Yule GU (1925) A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F.R.S. Philos Trans R Soc Lond B 213:21–87

Geometric Mean

Stevan Stević
Professor, Faculty of Economics, Brčko
University of East Sarajevo, Republic of Srpska, Bosnia and Herzegovina

According to Sir Thomas Thomas Heath (1921, p. 85) in Pythagoras's time, there were three means, the arithmetic, the geometric, and the subcontrary, and the "name of the third ('subcontrary') was changed by Archytas and Hippasus to harmonic." In English, the term geometrical mean can be found as early as in 1695 in the E. Halley's paper (Halley 1695–1697, p. 62). The geometric mean is a measure of central tendency that is "primarily employed within the context of certain types of analysis on business and economics" (Sheskin 2004, p. 8), such as an average of ►index numbers, ratios, and percent changes over time. For example, Fisher "ideal index" is the geometric mean of the Laspeyres index and the Paasche index. Geometric mean is also being used in modern ►portfolio theory and investment analysis (see, for example, Elton et al. 2009, p. 231), and in calculation of compound annual growth rate.

The geometric mean of n positive numbers $x_1, x_2, ..., x_n$ is defined as positive nth root of their product:

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} = \left(\prod_{i=1}^{n} x_i \right)^{1/n} \tag{1}$$

For example, the geometric mean of two numbers, 4 and 9, is the square root of their product, i.e., $\bar{x}_G = \sqrt[2]{4 \cdot 9} = 6$. It is important to emphasize that the calculation of geometric mean is either insoluble or meaningless if a data set contains negative numbers.

The geometric mean given in (1) can be expressed in a logarithmic form as

$$\ln \bar{x}_G = \left(\frac{1}{n} \sum_{i=1}^{n} \ln x_i \right),$$

or

$$\bar{x}_G = \exp\left(\frac{1}{n} \sum_{i=1}^{n} \ln x_i \right).$$

Just like the arithmetic and harmonic means, the geometric mean is under influence of each observation in the data set. It has an advantage over the arithmetic mean in that it is less affected by very small or very large values in skewed data.

The arithmetic–geometric–harmonic means inequality states that for any positive real numbers $x_1, x_2, \ldots, x_n$

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \geq \bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \geq \bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} \quad (2)$$

In (2) equality holds if and only if all the elements of the data set are equal. Interested reader can find 74 different proofs of the above inequality in Bullen (1987).

Cross References

►Aggregation Schemes
►Harmonic Mean
►Index Numbers
►Mean, Median, Mode: An Introduction
►P-Values, Combining of

References and Further Reading

Bullen PS (1987) Handbook of means and their inequalities, 2nd edn. Springer, Heidelberg
Elton EJ, Gruber MJ, Brown SJ, Goetzmann WN (2009) Modern portfolio theory and investment analysis, 8th edn. Wiley, New York
Halley E (1695–1697) A most compendious and facile method for constructing the logarithms, exemplified and demonstrated from the nature of numbers, without any regard to the hyperbola, with a speedy method for finding the number from the logarithm given. Phil Trans R Soc Lond 19:58–67
Heath T (1921) A history of Greek mathematics, vol. 1: from Thales to Euclid. Clarendon, Oxford
Sheskin DJ (2004) Handbook of parametric and nonparametric statistical procedures, 3rd edn. Chapman & Hall/CRC Press, Boca Raton

Geostatistics and Kriging Predictors

Shigeru Mase
Professor
Tokyo Institute of Technology, Tokyo, Japan

What Is Geostatistics?

Spatial statistics deals with spatial data. Geostatistics can be considered as a branch of spatial statistics. Its main objective is to interpolate values of a random field continuously from its sparsely observed data.

In the 1950s, South African gold mining engineer D.G. Krige introduced several statistical methods to predict average gold mine grade. Inspired by his work, French mathematician G. Matheron initiated a regression-based spatial prediction method for random field data in 1960s. He coined the term *kriging* (or, more frequently, *Kriging*) for his method, rewarding Krige's pioneering work. Actually, the kriging is a central tool of geostatistics and has been used almost synonymously.

Started as a mining technology, geostatistics has become now an indispensable statistical tool for various application fields where main objects spread continuously over a space. Examples are environmental sciences, ecology, forestry, epidemiology, oceanology, agriculture, fishery research, meteorology, civil engineering, and so on.

Kriging Predictions

The probabilistic framework of geostatistics is a random field $Z(\mathbf{x})$, $\mathbf{x} \in D \subset \mathbf{R}^d$, the dimension d being two or three typically. The data is a collection of values $Z(\mathbf{x}_1), Z(\mathbf{x}_2), \ldots, Z(\mathbf{x}_n)$ where $\mathbf{x}_i \in D$ may be regularly spaced or not. In Krige's case, $\mathbf{x}_i$ are locations of boring cores and $Z(\mathbf{x}_i)$'s are corresponding gold grades. A kriging predictor $\widehat{Z}(\mathbf{x}_0)$ of the value $Z(\mathbf{x}_0)$ at arbitrary location $\mathbf{x}_0 \in D$ is a liner combination of observed data:

$$\widehat{Z}(\boldsymbol{x}_0) = w_1 Z(\boldsymbol{x}_1) + w_2 Z(\boldsymbol{x}_2) + \cdots + w_n Z(\boldsymbol{x}_n). \quad (1)$$

As a result, one can plot the contour of predicted surfaces $\widehat{Z}(\mathbf{x}_0)$, $\mathbf{x}_0 \in D$.

Actually, it is usual to assume the second-order stationarity of $Z(\mathbf{x})$, that is, (1) the mean $\mu = \mathbf{E}\{Z(\mathbf{x})\}$ is constant, and (2) the covariance function depends only on the location difference, i.e., $C(\mathbf{x} - \mathbf{y}) = Cov(Z(\mathbf{x}), Z(\mathbf{y}))$. Coefficients $\{w_i\}$ of (1) are determined so that the mean squared kriging prediction error $\sigma_E^2(\mathbf{x}_0) = \mathbf{E}|\widehat{Z}(\mathbf{x}_0) - Z(\mathbf{x}_0)|^2$

$$\sigma_E^2(\boldsymbol{x}_0) = C(\mathbf{0}) - 2\sum_{i=1}^{n} w_i C(\boldsymbol{x}_0 - \boldsymbol{x}_i) + \sum_{i,j=1}^{n} w_i w_j C(\boldsymbol{x}_i - \boldsymbol{x}_j) \quad (2)$$

should be minimized under the unbiasedness condition $\mathbf{E}\{\widehat{Z}(\mathbf{x}_0)\} = \mu$, that is,

$$\sum_{i=1}^{n} w_i = 1. \quad (3)$$

A fortiori, the kriging predictor recovers the original data at every $\mathbf{x}_0 = \mathbf{x}_i$, $i = 1, 2, \ldots, n$. By the way, from a historical reason, the word "estimator" instead of "predictor" is preferred in the literature.

In geostatistical literature, the notion of the *intrinsic* stationarity has been preferred to the second-order stationarity. A random field Z is said to be intrinsic stationary if (1) $\mathbf{E}\{Z(\mathbf{x}) - Z(\mathbf{y})\} = 0$ and (2) $\mathbf{E}|Z(\mathbf{x}) - Z(\mathbf{y})|^2$ depends only on the difference $\mathbf{x} - \mathbf{y}$. The function $\gamma(\mathbf{h}) = \mathbf{E}|Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})|^2/2$ is called the *(semi-)variogram* of Z. Note that it is not assumed the existence of $\mathbf{E}\{Z(\mathbf{x})\}$. If Z is second-order stationary, then it is intrinsic stationary and there is the relation

$$\gamma(\boldsymbol{h}) = C(\mathbf{0}) - C(\boldsymbol{h}).$$

$\gamma(\mathbf{x})$ is non-negative, even and $\gamma(0) = 0$. Variogram functions may be unbounded contrary to covariance functions. Its characteristic property is the *conditional non-positive definiteness*: for every $\{\mathbf{x}_i\}$ and $\{w_i\}$ with $\sum_i w_i = 0$

$$\sum_{i,j=1}^{n} w_i w_j \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) \le 0.$$

In terms of variogram functions, the kriging prediction error can be expressed as

$$\sigma_E^2(\boldsymbol{x}_0) = -\gamma(\mathbf{0}) + 2\sum_{i=1}^{n} w_i \gamma(\boldsymbol{x}_i - \boldsymbol{x}) - \sum_{i,j=1}^{n} w_i w_j \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) \quad (4)$$

if the unbiasedness condition (3) is satisfied. The use of variograms instead of covariances intends to eliminate possible linear trends.

The followings are three frequently used covariance function models. They are all isotropic models, that is, depend only on the modulus $h = |\mathbf{h}|$:

spherical model
$$C_{sph}(h) = \begin{cases} b\left(1 - \dfrac{3|h|}{2a} + \dfrac{|h|^3}{2a^3}\right) & (0 \le |h| \le a) \\ 0 & (a < |h|) \end{cases},$$

exponential model
$$C_{exp}(h) = b\exp(-|h|/a) \quad (a, b > 0),$$

Gaussian model
$$C_{gau}(h) = b\exp(-h^2/a) \quad (a, b > 0).$$

The parameter a is called the *range*, from which values of the function are exactly or nearly equal to zero. The parameter b is called the *sill*, the maximal value of the function. Sometimes covariance functions have jumps at 0 showing a *nugget effect*, a pure random noise.

Kriging Equations

There are three types of standard kriging predictors:

1. Simple kriging: The common mean μ is assumed to be known.
2. Ordinary kriging: The common mean μ is assumed to be unknown.
3. Universal kriging: The mean function (trend) $\mu(\mathbf{x})$ is assumed to be a linear combination of given basis functions (say, spatial polynomials).

For example, coefficients $\{w_i\}$ of the ordinary kriging predictor are calculated by solving the following ordinary kriging equation (if expressed in terms of variograms):

$$\begin{pmatrix} \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_1) & \dots & \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(\boldsymbol{x}_n - \boldsymbol{x}_1) & \dots & \gamma(\boldsymbol{x}_n - \boldsymbol{x}_n) & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \lambda \end{pmatrix} = \begin{pmatrix} \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_0) \\ \vdots \\ \gamma(\boldsymbol{x}_n - \boldsymbol{x}_0) \\ 1 \end{pmatrix} \quad (5)$$

where λ is the Lagrange multiplier for the constraint $\sum_i w_i = 1$. The prediction error $\sigma_E^2(\mathbf{x}_0)$ is expressed as:

$$\sigma_E^2(\boldsymbol{x}_0) = -\gamma(\mathbf{0}) + \lambda + \sum_{i=1}^{n} w_i \gamma(\boldsymbol{x}_i - \boldsymbol{x}_0). \quad (6)$$

Block Kriging

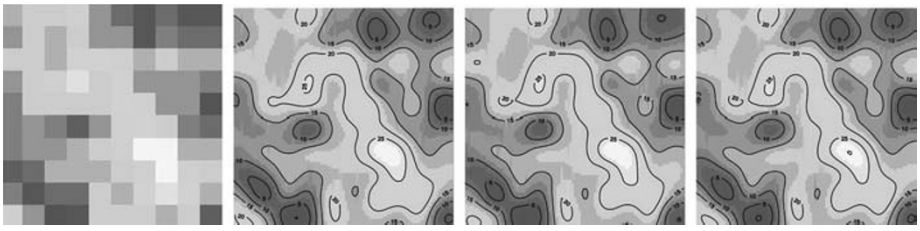
Sometimes, one wants to predict spatial averages

$$Z(D_0) = \frac{1}{|D_0|}\int_{D_0} Z(\boldsymbol{x})d\boldsymbol{x}$$

for a block D_0. The corresponding *block kriging* predictor $\widehat{Z}(D_0)$ is also an unbiased linear combination of point data $\{Z(\mathbf{x}_i)\}$. Note that, if Z has a constant mean μ, then $Z(D_0)$ has also the mean μ. Furthermore,

$$Cov(Z(D_0), Z(\boldsymbol{x})) = \frac{1}{|D_0|}\int_{D_0} Cov(Z(\boldsymbol{x}), Z(\boldsymbol{y}))d\boldsymbol{y}. \quad (7)$$

One can derive the simple or ordinary block kriging equation and its prediction error formula analogously as in (5) and (6).



Geostatistics and Kriging Predictors. Fig. 1 Mesh data of populations of Tokyo metropolitan area (1,000/km^2), block-to-point kriging result using the spherical model, the exponential model, and the Gaussian model from left to right

Actually, one can consider following four situations:

point-to-point kriging: $\widehat{Z}(\boldsymbol{x}_0) = \sum_{i=1}^n w_i Z(\boldsymbol{x}_i)$

point-to-block kriging: $\widehat{Z}(D_0) = \sum_{i=1}^n w_i Z(\boldsymbol{x}_i)$

block-to-point kriging: $\widehat{Z}(\boldsymbol{x}_0) = \sum_{i=1}^n w_i Z(D_i)$

block-to-block kriging: $\widehat{Z}(D_0) = \sum_{i=1}^n w_i Z(D_i)$

The standard procedure to derive kriging predictor coefficients is almost the same, at least formally, except for numerical difficulties. For example, the block-to-point or block-to-block kriging equations contain a lot of block data covariances such as

$$Cov(Z(D_i), Z(D_j)) = \frac{1}{|D_i||D_j|} \iint_{D_i \times D_j} Cov(Z(\boldsymbol{x}), Z(\boldsymbol{y}))d\boldsymbol{x}d\boldsymbol{y}.$$

as well as (7) which should be computed using numerical integrations (see ►Numerical Integration). Figure 1 shows a block-to-point prediction result (Mase et al. (2009)).

Covariance or Variogram Model Fittings

In order to apply the kriging method, one has to estimate the type and the parameters (usually the mean, the range, and the sill) of covariance or variogram models. This model fitting is the most difficult part of the kriging method. The standard method is the least squared fitting of an isotropic variogram model $\gamma_\theta(h)$ to the scatter plot (sample variogram) of $(|\mathbf{x}_i - \mathbf{x}_j|, |Z(\mathbf{x}_i) - Z(\mathbf{x}_j)|^2)$ called the *variogram cloud*. In practice, distances $|\mathbf{x}_i - \mathbf{x}_j|$ are classified into small consecutive intervals and corresponding values $|Z(\mathbf{x}_i) - Z(\mathbf{x}_j)|^2$ are averaged per intervals in order to increase robustness.

Another method is the maximum likelihood fitting of a covariance function model $C_\theta(\mathbf{h})$ assuming the Gaussianity of random field $Z(\mathbf{x})$. This procedure needs no data grouping and can be applied even to non-isotropic models. Let Z be a second-order stationary Gaussian random field with mean μ and covariance function $C_\theta(\mathbf{x})$. Then the data vector $\mathbf{Z} = (Z(\mathbf{x}_1), \cdots, Z(\mathbf{x}_n))^t$ is Gaussian with mean vector $\mu = (\mu, \cdots, \mu)^t$ and covariance matrix $\Sigma_\theta = (C_\theta(\mathbf{x}_i - \mathbf{x}_j))_{ij}$. Therefore, the log-likelihood of $\mathbf{Z}$ is

$$l(\mu, \theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_\theta| - \frac{1}{2}(\boldsymbol{Z} - \boldsymbol{\mu})^t \Sigma_\theta^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}).$$

By maximizing $l(\mu, \theta)$, we can get the maximum likelihood estimators of μ and θ.

Multivariate Kriging Method

Sometimes there may be an auxiliary random field data $\{Z_2(\mathbf{y}_j)\}$ (or even more) in addition to the main data $\{Z_1(\mathbf{x}_i)\}$. Locations $\{\mathbf{x}_i\}$ may or may not be identical to $\{\mathbf{y}_j\}$. If we assume the second-order stationarity of the multivariate random field $(Z_1(\mathbf{x}), Z_2(\mathbf{x}))$, we can construct the *cokriging* predictor $\widehat{Z}_1(\mathbf{x}_0)$ as

$$\widehat{Z}_1(\boldsymbol{x}_0) = \sum_i w_i^{(1)} Z_1(\boldsymbol{x}_i) + \sum_j w_j^{(2)} Z_2(\boldsymbol{y}_j).$$

Two sets of coefficients $\{w_i^{(1)}\}$ and $\{w_j^{(2)}\}$ can be determined as before. If the number of auxiliary data is too large, one can lessen the resulting numerical complexity by restricting $\{Z_2(\mathbf{y}_j)\}$ to those closer to $\mathbf{x}_0$. In particular, the *collocated* kriging uses only $Z_2(\mathbf{x}_0)$ if it is available as data.

About the Author

Dr Shigeru Mase is a Professor, Department of Math. and Comp. Sciences, Tokyo Institute of Technology, Japan. He was the Head of the Department of Mathematics and Computer Sciences (1999, 2005). He is an Associate editor of Annals of the Institute of Statistical Mathematics. Professor Mse wrote 3 books, including Spatial Data Analysis (Kyoritsu Publishing Co., 2001) and R Programing Manual (Science Publishing Co., 2007). He received the Statistical Activity award (Japan Statistical Society, 2009) for his leading activity of introducing the R statistical system into Japan.

Cross References

►Model-Based Geostatistics
►Spatial Statistics

References and Further Reading

Banerjee S, Carlin BP, Gelfand AE (2003) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC, London
Chiles Chiles J-P, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York
Cressie NAC (1991) Statistics for spatial data. Wiley, New York
Mase S et al (2009) Geostatistical predictions based on block data. Submitted to Ann Inst Stat Math
Wackernagel H (2003) Multivariate geostatistics, 3rd edn. Springer, New York

Glivenko-Cantelli Theorems

Olimjon Shukurovich Sharipov
Professor
Uzbek Academy of Sciences, Tashkent, Uzbekistan

The Glivenko-Cantelli Theorem

Let $X_1, \ldots, X_n, \ldots$ be a sequence of independent identically distributed (i.i.d.) random variables with a common distribution function $F(x)$. Introduce the empirical distribution function:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i < x\}$$

where $I\{\cdot\}$ is the indicator function. By the strong law of large numbers (SLLN) for any fixed point $x \in R$, we have

$$F_n(x) \to F(x) \text{ a.s., as } n \to \infty$$

The Glivenko-Cantelli theorem states that this convergence is uniform.

Theorem 1 *(Glivenko-Cantelli): As* $n \to \infty$

$$\sup_{x \in R} |F_n(x) - F(x)| \to 0 \text{ a.s.}$$

Theorem 1 was proved by V.I. Glivenko Birkhoff (1931) in the case of continuous $F(x)$ and by F.P. Cantelli Borovkov (1998) in the general case. In mathematical statistics, the Glivenko-Cantelli theorem can be interpreted as follows based on independent observations $x_1, \ldots, x_n$ of the random variable ξ one can approximate the distribution function $F(x)$ of ξ arbitrarily close by the empirical distribution function:

$$\overline{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{x_i < x\}.$$

This is a very important fact. Because of this fact, the Glivenko-Cantelli theorem is commonly referred to as a central or fundamental result of mathematical statistics.

The proof of the theorem is based on the SLLN (see, for instance Bradley (2007), Cantelli (1933)). The following extension of Theorem 1 is valid.

Theorem 2 *Let* $X_1, \ldots, X_n, \ldots$ *be a stationary and ergodic sequence of random variables with a common distribution function* $F(x)$. *Then as* $n \to \infty$

$$\sup_{x \in R} |F_n(x) - F(x)| \to 0 \text{ a.s.}$$

The proof of Theorem 2 is almost the same as the proof of Theorem 1 except for a single difference: instead of using SLLN in the proof of Theorem 1 one should use Birkhoff's ergodic theorem Dudley (1984). Theorem 2 remains true for all stationary sequences of mixing random variables (see for mixing conditions Dudley et al. (1991)) since such sequences are ergodic.

We can reformulate the Gllivenko-Cantelli theorem using empirical measures. Denote by $P(A)$ a distribution of X_1. The empirical distribution is defined as:

$$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \in A\}. \quad (1)$$

By $\mathcal{A}$ we denote a class of all semi intervals [a,b) with finite or infinite endpoints. The following result takes place.

Theorem 3 *Let* $X_1, \ldots, X_n, \ldots$ *be a sequence of i.i.d. random variables with the common distribution* $P(A)$. *Then as* $n \to \infty$

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \to 0 \text{ a.s.}$$

Theorems 1 and 3 are equivalent since they imply each other.

For some improvements of the Glivenko-Cantelli theorem, such as the law of the iterated logarithm see, for instance Bradley (2007).

Generalizations of Glivenko-Cantelli Theorem

The Glivenko-Cantelli theorem gave a start to investigations on convergence of empirical measures and processes in finite and infinite dimensional spaces.

Let $X_1, \ldots, X_n, \ldots$ be a sequence of i.i.d. random vectors with values in R^k and with a common distribution function $F(t)$. Define the empirical distribution function as:

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\left\{X_i^{(1)} < t_1, \ldots, X_i^{(k)} < t_k\right\}$$

where $X_i = \left(X_i^1, \ldots, X_i^k\right)$ and $t = (t_1, \ldots, t_k)$.

The generalization of Theorem 1 is the following.

Theorem 4 *As* $n \to \infty$

$$\sup_{t \in R^k} |F_n(t) - F(t)| \to 0 \ a.s.$$

Now by $P(B)$ we denote a distribution corresponding to distribution function $F(t)$ and by $\mathcal{B}$ a class of all measurable convex sets in R^k. The empirical distribution $P_n(B)$ is defined as in (1).

Theorem 5 *Let* $X_1, \ldots, X_n, \ldots$ *be a sequence of i.i.d. random variables with values in* R^k *and with the common distribution* $P(B)$*, which is absolutely continuous with respect to the Lebesque measure in* R^k*. Then as* $n \to \infty$

$$\sup_{B \in \mathcal{B}} |P_n(B) - P(B)| \to 0 \ a.s.$$

Theorem 5 is a consequence of results by Ranga Rao Glivenko (1933) (see also Appendix I in Bradley (2007)).

In infinite dimensional spaces the situation is more complicated. Now, assume that $(\mathcal{X}, \mathcal{C})$ is a measurable space and let $X_1, \ldots, X_n, \ldots$ be a sequence of i.i.d. random variables with values in $\mathcal{X}$ with a common distribution $P \in \mathcal{P}$ (here $\mathcal{P}$ is a set of all distributions on $\mathcal{X}$) and again we denote by $P_n(B)$ the empirical distribution defined as in (1).

In separable metric spaces $\mathcal{X}$, Varadarajan Ranga Rao (1962) proved almost surely weak convergence of empirical distributions, i.e., as $n \to \infty$

$$P_n(B) \to P(B) \text{ for all Borel sets } B \in \chi \text{ such that } P(\partial B) = 0 \ a.s.$$

where ∂B is a boundary of B.

In general, in infinite dimensional spaces uniform convergence is not valid even over the class of all half-spaces (see Sazonov (1963)). Generalizations of the Glivenko-Cantelli theorem in linear spaces $\mathcal{X}$ are mostly devoted to studying the almost certain convergence properties of the following two values:

$$\Delta(V) = \sup_{B \in V} |P_n(B) - P(B)|$$

$$\Delta(\mathcal{F}) = \sup_{f \in \mathcal{F}} |P_n(f) - P(f)|$$

where V is a class of sets, $\mathcal{F}$ is a class of functions $f : \mathcal{X} \to R$ and $P(f) = \int f dP$ for any distribution P.

Generalizing the Glivenko-Cantelli theorem several notions such as Glivenko-Cantelli class, universal Glivenko-Cantelli class, and uniform Glivenko-Cantelli class of sets or functions were introduced.

The class of sets V is called a Glivenko-Cantelli class of sets, if as $n \to \infty$ the following holds:

$$\Delta(V) \to 0 \ a.s. \tag{2}$$

The class of sets V is called a universal Glivenko-Cantelli class of sets, if (2) holds for any $P \in \mathcal{P}$.

The class of sets V is called a uniform Glivenko-Cantelli class of sets, if as $n \to \infty$ the following holds:

$$\sup_{P \in \mathcal{P}} \Delta(V) \to 0 \ a.s.$$

The corresponding classes for the sets of functions can be defined similarly.

One of the problems in infinite dimensional spaces is a measurability problem in non-separable spaces. We will give two theorems and in one of them this problem will appear. Before giving results we need to introduce necessary notions and notation.

We denote by $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ the covering number, which is defined as the minimal number of balls $\{g : \|g - f\| < \varepsilon\}$ of radius ε needed to cover $\mathcal{F}$. Given two functions l and u, the bracket $[l, u]$ is a set of all functions f such that $l \le f \le u$. An ε-bracket is a bracket $[l, u]$ with $\|l - u\| < \varepsilon$. The minimum number of ε-brackets needed to cover $\mathcal{F}$ we denote by $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$. Note that $\log N(\varepsilon, \mathcal{F}, \|\cdot\|)$ and $\log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ are called an ▸entropy and an entropy with bracketing, respectively. An envelope function of $\mathcal{F}$ is any function $F(x)$ such that $|f(x)| \le F(x)$ for any x and $f \in \mathcal{F}$. As a norm $\|\cdot\|$ we use a $L_1(P)$-norm i.e.

$$||f||_{L_1(P)} = \int |f| dP.$$

A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to R$ on probability *space*$(\mathcal{X}, \mathcal{C}, \mathcal{P})$ is called P-measurable if the function

$$(X_1, \ldots, X_n) \to \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} e_i f(X_i) \right|$$

is measurable on the completion of $(\mathcal{X}^n, \mathcal{C}^n, P^n)$ for every n and every vector $(e_1, \ldots, e_n) \in R^n$. By P^* we denote the outer probability.

Now we can formulate the following theorems (for the proofs see Varadarajan (1958)).

Theorem 6 *Let* $\mathcal{F}$ *be a class of measurable functions such that* $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ *for every* $\varepsilon > 0$*. Then* $\mathcal{F}$ *is the Glivenko-Cantelli class.*

Theorem 7 *Let* $\mathcal{F}$ *be a P-measurable class of measurable functions with envelope F such that* $P^*(F) < \infty$*. Let* $\mathcal{F}_M$ *be the class of functions* $fI\{F \le M\}$ *when f range over* $\mathcal{F}$*. If* $\log N(\varepsilon, \mathcal{F}_M, L_1(P_n)) = o_{P^*}(n)$ *for every* ε *and* $M > 0$*, then as* $n \to \infty$

$$\Delta(\mathcal{F}) \to 0 \ P^* - a.s.$$

i.e., $\mathcal{F}$ *is the Glivenko-Cantelli class.*

For other results in infinite dimensional spaces we refer to Shorack and Wellner 1986; Dudley 1984; Dudley et al. 1991; Vapnik and Červoninkis 1971; Vapnik and Červoninkis 1981; Talagrand 1987; TØpsoe 1970; van der Vaart and Wellner 1996 (see also references therein).

About the Author

Prof. Dr. Olimjon Shukurovich Sharipov is Chair of Scientific Grant of the Department of Probability Theory and Mathematical Statistics of the Institute of Mathematics and Information Technologies of Uzbek Academy of Sciences, Tashkent, Uzbekistan. He has published more than 40 papers.

Cross References

►Almost Sure Convergence of Random Variables
►Empirical Processes
►Estimation: An Overview
►Foundations of Probability
►Laws of Large Numbers
►Limit Theorems of Probability Theory

References and Further Reading

Birkhoff GD (1931) Proof of the ergodic theorem. Proc Nat Acad Sci USA 17:656–660
Borovkov AA (1998) Mathematical statistics. Gordon & Breach Science, Amsterdam
Bradley RC (2007) Introduction to strong mixing conditions, vol 1–3. Kendrick, Heber City
Cantelli FP (1933) Sulla determinazione empirica della leggi di probabilita. Giorn Ist Ital Attuari 4:421–424
Dudley RM (1984) A course on empirical processes. (École d'Été de probabilités de Saint-Flour, XII-1982) In: Hennequin PL (ed) Lecture notes in mathematics, vol 1097. Springer, New York
Dudley RM, Gine E, Zinn J (1991) Uniform and universal Glivenko-Cantelli classes. J Theor Probab 4(3):485–510
Glivenko VI (1933) Sulla determinazione empirica della leggi di probabilita. Giorn Ist Ital Attuari 4:92–99
Ranga Rao R (1962) Relations between weak and uniform convergence of measures with applications. Ann Math Stat 33:659–680
Sazonov VV (1963) On the theorem of Glivenko and Cantelli. Teor Verojatnost i Primenen (Russian) 8:299–303
Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley series in probability and mathematical statistics. Wiley, New York
Talagrand M (1987) The Glivenko-Cantelli problem. Ann Probab 15:837–870
Tøpsoe F (1970) On the Glivenko-Cantelli theorem. Z Wahrscheinlichkeitstheorie und Verw Gebiete 4:239–250
van der Vaart AW, Wellner JA (1996) Weak convergence and empirical processes: with applications to statistics Springer series in statistics. Springer, New York
Vapnik VN, Červoninkis AY (1971) On uniform convergence of the frequencies of events to their probabilities. Theor Probab Appl 16:264–280
Vapnik VN, Červoninkis AY (1981) Necessary and sufficient conditions for the uniform convergence of means to their expectations. Theor Probab Appl 26:532–553
Varadarajan VS (1958) On the convergence of sample probability distributions. Sankhya 19:23–26

Graphical Analysis of Variance

G. E. P. Box
Professor Emeritus
University of Wisconsin, Madison, WI, USA

Walter Shewhart said:

"Original data should be presented in a way that would preserve the evidence in the original data," (1939, p. 88).

Frank Anscombe said:

"A computer should make both calculation and graphs. Both kinds of output contribute to understanding," (1973, p. 17).

And Yogi Berra said:

"You can see a lot by just looking."

A Simple Comparative Experiment

As an illustration of graphical analysis of variance, Table 1a shows coagulation times for samples of blood drawn from 24 animals randomly allocated to four different diets A, B, C, D. Table 1b shows an ►analysis of variance for the data.

While the standard analysis of variance shown in Table 1b yielding F values and p values is very useful, the additional graphical analysis in Table 1c can lead to a deeper understanding. This is because it puts the brain of the experimenter in direct contact with the data. In the graphical analysis of variance of Table 1c the coagulation data are represented by dots. The residual dots are calculated and plotted in the usual way. However the plots of the deviations of the treatment means from the grand mean are multiplied by a scale factor. If ν_R is the degrees of freedom for the residuals and ν_T is the degrees of freedom for the treatment means, then the scale factor is $\sqrt{\frac{\nu_R}{\nu_T}}$ in this example $\sqrt{\frac{20}{3}} = 2.6$. The scale factor is such that if there were no differences between treatment means, the expected value of *natural* variance for the treatments would be the same as that for the residuals. By the natural variance is meant the sum of squares of the dot deviations divided by the *number* of dots (not the degrees of freedom). This measure of the spread is

appropriate because it is equated to what the eye *actually sees* (The ratio of natural variances of the dot plots produces the usual F-value). The analysis asks the question, "Might the scaled treatment deviations just as well be part of the noise?" In this example the treatment dot deviations −7.8, 5.2, 10.4 and −7.8 are obtained by multiplying the treatment deviations −3, +2, +4, −3 by 2.6. The graphical analysis supports the finding that the differences in the treatments are unlikely due to chance. But it does more. The graphical analysis helps the analyst appreciate the nature of the differences and similarities produced by the treatments, something that the ANOVA table does not do well. It also directs attention to the individual residuals and any large deviations that might call for further study. For this example it makes clear that there is nothing suspicious about the distribution of the residuals. Also that the treatments A and D are almost certainly alike in their effects but C is markedly different. Experimenters sometimes believe that a high level of significance necessarily implies that treatment effects are accurately determined and separated. The graphical analysis discourages overreaction to high significance levels, and reveals "very nearly" significant differences.

Graphical Analysis of Variance. Table 1a Coagulation times for blood drawn from 24 animals randomly allocated to four diets

Diets (Treatments)				
	A	B	C	D
	$62^{(20)}$	$63^{(12)}$	$68^{(16)}$	$56^{(23)}$
	$60^{(2)}$	$67^{(9)}$	$66^{(7)}$	$62^{(3)}$
	$63^{(11)}$	$71^{(15)}$	$71^{(1)}$	$60^{(6)}$
	$59^{(10)}$	$64^{(14)}$	$67^{(17)}$	$61^{(18)}$
	$63^{(5)}$	$65^{(4)}$	$68^{(13)}$	$63^{(22)}$
	$59^{(24)}$	$66^{(8)}$	$68^{(21)}$	$64^{(19)}$
Treatment averages	61	66	68	61
Grand average	64	64	64	64
Difference	−3	+2	+4	−3

Graphical Analysis of Variance. Table 1b Analysis of Variance (ANOVA) Table. Blood Coagulation example

Sources of variation	Sum of squares	Degrees of freedom	Mean square	
Between treatments	$S_T = 228$	$\nu_T = 3$	$m_T = 76.0$	$F_{3,20} = 13.63$ $p < 0.01$
Within treatments	$S_R = 112$	$\nu_R = 20$	$m_R = 5.6$	
Total	$S_D = 340$	$\nu_D = 23$		

A Latin Square

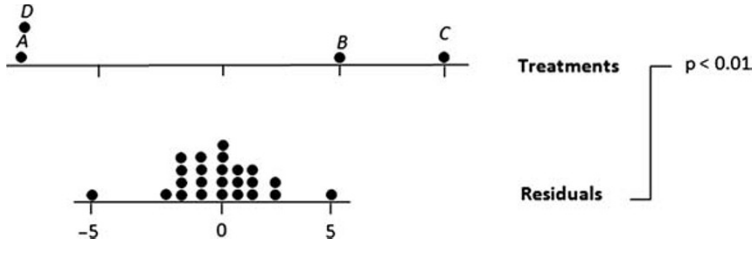
The following Latin square design was run to find out if the amount of carbon monoxide exuded by an automobile could be reduced by adding small amounts of additives A, B, C or D to the gasoline. The additives were tested with four different drivers and four different cars over a difficult fixed course. The design was used to help eliminate from the additive comparisons possible differences produced by drivers and cars. In this arrangement the additives were randomly allocated to the symbols A, B, C, D; the drivers to the symbols I, II, III, IV and the cars to the symbols 1, 2, 3, 4 as in Table 2a.

The standard analysis and the corresponding graphical analysis are shown in Table 2b, 2c. In this particular experiment, apparently there were differences between drivers but not between additives and cars.

A Split Plot Experiment

Two considerations important in choosing any statistical arrangement are convenience and efficiency. In industrial experimentation the split plot design is often convenient and is sometimes the only practical possibility. (The

Graphical Analysis of Variance. Table 1c Graphical Analysis of Variance



nomenclature is from agricultural experimentation where split plots were first used.) In particular this is so whenever there are certain factors that are difficult to change and others that are easy. Table 3a shows the data from an experiment designed to study the corrosion resistance of steel bars subjected to heat treatment with four different coatings C_1, C_2, C_3, C_4 arranged randomly within each of the six heats at furnace temperatures of 360, 370, 380, 380, 370, 360 °C. In this experiment changing the furnace temperature was difficult but re-arranging the positions of the coated bars in the furnace was easy. A fully randomized experiment would have required changing the temperature up to 24 times. This would have been very inconvenient and costly. Notice that there are two error sources - between heats (whole plots) and coatings (sub-plots).

The entries in the analysis of variance in Table 3b may be calculated as if there was no split plotting. Table 3c shows separately the analyzes for whole plots and sub-plots. They are conveniently arranged under two

Graphical Analysis of Variance. Table 2a The 4 × 4 Latin squared: automobile emission data

Drivers	Cars				Averages		
	1	2	3	4	Cars	Drivers	Additives
I	A	B	D	C	1: 19	I: 23	A: 18
	19	24	23	26			
II	D	C	A	B	2: 20	II: 24	B: 22
	23	24	19	30			
III	B	D	C	A	3: 19	III: 15	C: 21
	15	14	15	16			
IV	C	A	B	D	4: 22	IV: 18	D: 19
	19	18	19	16			
					Grand Average: 20		

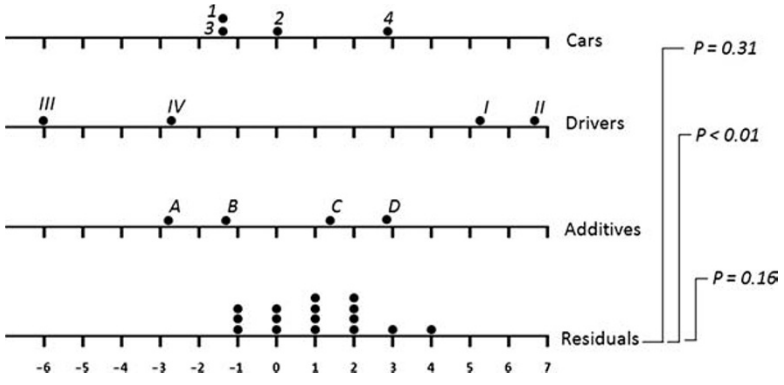
Graphical Analysis of Variance. Table 2b Analysis of variance: Latin square example

Source of variation	Sum of squares	Degrees of freedom	Mean square	F	p
Cars	$S_c = 24$	3	$m_c = 8.00$	$F_{3,6} = 1.5$	0.31
Drivers	$S_D = 216$	3	$m_D = 72.00$	$F_{3,6} = 13.5$	<0.01
Additives	$S_T = 40$	3	$m_T = 13.33$	$F_{3,6} = 2.5$	0.16
Residuals	$S_R = 32$	6	$m_R = 5.33$		
Total	$S_v = 312$	15			

Graphical Analysis of Variance. Table 3a Split plot design to study the corrosion resistance of steel bars treated with four different coatings randomly positioned in a furnace and baked at different temperatures

Corrosion: data rearranged in rows and columns coatings						
Heats	C_1	C_2	C_3	C_4	Averages	
360	67	73	83	89	78.00	56.63
	33	8	46	54	35.25	
370	65	91	87	86	82.25	110.25
	140	142	121	150	138.25	
380	155	127	147	212	160.25	135.59
	108	100	90	153	112.75	
Average	94.67	90.17	95.67	124.00	–	101.125

Graphical Analysis of Variance. Table 2c Graphical ANOVA for the Latin square example

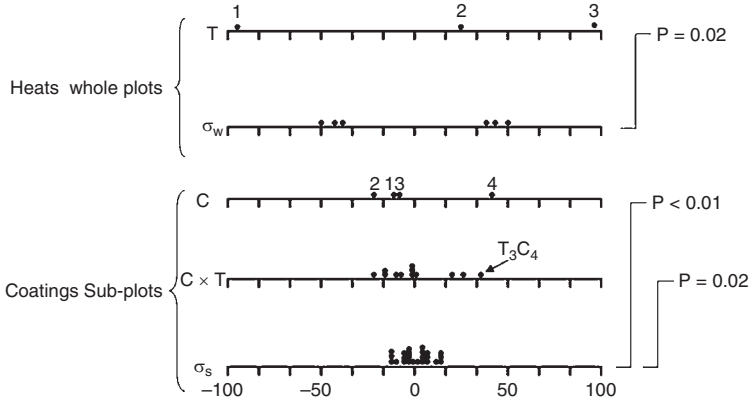


Graphical Analysis of Variance. Table 3b ANOVA for corrosion resistance data. The parallel column layout identifies appropriate errors for whole-plots and subplots effects

Heats (whole plots)					Coatings (subplots)				
Source	df	SS	MS		Source	df	SS	MS	
Average, I	1	245430	245430		C	3	4289	1430	$F_{3,9} = 11.5$
Temperature	2	26519	13260	$F_{2,3} = 2.8$	$T \times C$	6	3270	545	$F_{6,9} = 4.4$
Error E_W	3	14440	4813		Error E_s	9	1121	125	

Note: The convention is used that a single asterisk indicates statistical significance at the 5% level and two asterisks statistical significance at the 1% level.

Graphical Analysis of Variance. Table 3c Graphical ANOVA, corrosion data



headings: Heats (whole plots) and Coatings (sub-plots). Notice the relatively small variances for coatings (sub-plots) as compared with that for heats (whole plots) and the detection of interaction between coatings and temperatures. This clearly shows the advantage of the split plot design since it was the comparison coatings that were of main interest.

About the Author

George Edward Pelham Box (born 18 October, 1919) was in Gravesend, Kent, England) was President of the American Statistical Association in 1978 and President of the Institute of Mathematical Statistics in 1979. He received his Ph.D. from the University College, London in 1953, under the supervision of Egon Pearson. He was employed as a statistician at Imperial Chemical Industries from 1948 to 1956, and from 1957 to 1959 he was Director of the Statistical Techniques Research Group at Princeton University. In 1960, he founded the Department of Statistics at the University of Wisconsin-Madison. In 1971 he was appointed Ronald Aylmer Fisher Professor of Statistics, and in 1980 he became Vilas Research Professor of Mathematics and Statistics, this being the highest honor that Wisconsin could bestow to a member of their faculty. He retired in 1992 and was given the title of Professor Emeritus by the University of Wisconsin. The importance of his fundamental contributions to many areas of statistics has been recognised by receiving many awards, among them the British Empire Medal in 1946, the Shewhart Medal from the American Society for Quality Control in 1968, the Wilks Memorial Award from the American Statistical Association in 1972, the R. A. Fisher Lectureship in 1974, the Guy Medal in Gold from the Royal Statistical Society in 1993 and the Deming medal in 1989, among others. He was elected a member of the American Academy of Arts and Sciences in 1974 and a Fellow of the Royal Society in 1985. Professor Box is very well-known for his work in Experimental Designs, Time Series, and Regression with his name on many important techniques including Box-Cox power transformation, Box-Jenkins time series models, Box-Muller transformation and Box-Behnken designs. He has also authored many well-known texts in time series and stochastic control, Bayesian statistics and experimental design and he invented the concept of evolutionary operation. He is a co-author (with Gwilym Jenkins) of the seminal book *Time Series Analysis: Forecasting and Control* (1970; Holden-Day). Professor Box holds four honorary doctorates.

"George Box, a legend in statistics, is one of the most respected living statisticians. His contributions to statistics are outstanding and he has made fundamental contributions in areas such as design and analysis of experiments, response surface methodology and time series analysis...Professor Box has created a school of industrial statisticians who span industry, academia and government; his followers are in every major corner of the world." (University of Waterloo, News Release, 80th Convocation, 2000, http://newsrelease.uwaterloo.ca/news.php?id=1408).

"All models are wrong, but some are useful. The quotation comes from George Box, one of the great statistical minds of the 20th century" (Ron Wasserstein "George Box: a model statistician", *Significance*, 7(3), 2010).

"George Box is truly one of the towering figures in the history of industrial experimentation" (Geoff Vining, "George's Contributions to Industrial Experimentation", *Quality and Productivity Research Conference*, Madison, 2008).

Cross References

►Analysis of Variance
►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Data Analysis
►Experimental Design: An Introduction
►F Distribution
►Multivariate Analysis of Variance (MANOVA)
►Research Designs

References and Further Reading

Anscombe FJ (1973) Graphs in statistical analysis. Am Stat 27: 17–21
Shewhart WA (1939) Statistical method from the viewpoint of quality control. The Graduate School, The Dept of Agriculture, Washington, DC
Yogi Berra (Attributed)

The examples are from Chaps. 4 and 9 in the book *Statistics for Experimenters* by George Box, J. Stuart Hunter and William G. Hunter, Second edition, 2008, John Wiley & Sons Inc.

Graphical Markov Models

Nanny Wermuth
Professor of Statistics
Chalmers/University of Gothenburg, Gothenburg, Sweden

Graphical Markov models are multivariate statistical models which are currently under vigorous development and which combine two simple but most powerful notions, generating processes in single and joint response variables and conditional independences captured by graphs.

The development of graphical Markov models started in 1975–1980, extending early work in 1920–1930 by geneticist Sewall Wright. Wright used graphs, in which nodes represent variables and arrows capture dependence, to describe hypotheses about stepwise processes in single responses that could have generated his data.

He developed a method, called path analysis, to estimate linear dependences and to judge whether the hypotheses are well compatible with his data that he summarized in terms of simple and partial correlations. With this approach he was far ahead of his time, since corresponding formal statistical methods for estimation and tests of goodness of fit were developed much later and graphs that capture independences much later than tests of goodness of fit.

It remains a primary objective of graphical Markov models to uncover graphical representations that lead to an understanding of data generating processes. Such processes are no longer restricted to linear relations but contain linear dependences as special cases. A probabilistic data generating process is a recursive sequence of conditional distributions in which response variables may be vector variables that contain discrete or continuous components. Thereby, each conditional distribution specifies both the type of dependence of response Y_a, say, on its disjoint explanatory variable vector Y_b and the type of undirected associations of the components of Y_a.

Graphical Markov models generalize sequences in single responses and single explanatory variables that have been named ►Markov chains, after probabilist Andrey A. Markov. Markov recognized in 1900–1910 that seemingly complex joint probability distributions may be radically simplified by using the notion of conditional independence.

In a Markov chain of random variables $Y_1, \ldots, Y_i, \ldots, Y_d$, the joint distribution is built up by starting with the density of f_d of Y_d, by considering with $f_{d-1|d}$ conditional dependence of Y_{d-1} on Y_d, then taking conditional independence of Y_{d-2} from Y_d given Y_{d-1} into account by formulating $f_{d-2|d-1,d} = f_{d-2|d-1}$, by continuing such that, with $f_{i|i+1,\ldots d} = f_{i|i+1}$, response Y_i is conditionally independent of $Y_{i+2}, \ldots, Y_d$ given Y_{i+1} (written compactly in terms of nodes as $i \perp\!\!\!\perp \{i+2, \ldots, d\} | i+1$) and finally having with $f_{1|2,\ldots,d} = f_{1|2}$ just Y_2 as explanatory variable of response Y_1.

The directed graph that captures the independences in such a Markov chain is a single directed path of arrows, with an arrow starting at node d and pointing to node $d-1$ and ending with an arrow starting at node 2 and pointing

to node 1. Thus, for $d = 5$ and node set $N = \{1, 2, 3, 4, 5\}$, the graph of a Markov chain is

$$1 \longleftarrow 2 \longleftarrow 3 \longleftarrow 4 \longleftarrow 5.$$

The graph corresponds to the factorization of the joint density f_N given by

$$f_N = f_{1|2} f_{2|3} f_{3|4} f_{4|5} f_5.$$

The three defining local independence statements implied by the above factorization or by the corresponding path of dependences are $1 \perp\!\!\!\perp \{3, 4, 5\}|2$, $2 \perp\!\!\!\perp \{4, 5\}|3$ and $3 \perp\!\!\!\perp 5|4$. One also says that in the generating process, each response i remembers of its past just the nearest neighbor $i + 1$.

It remains an important secondary objective of graphical Markov models, that some type of graph is to capture the independence structure of interest for f_N, that is a set of all independence statements satisfied by f_V and that is implied by a given graph. In principle, all independence statements that arise from a given set of statements defining a graph, may be derived from basic laws of probability. Thus, the above Markov chain implies for instance

$$1 \perp\!\!\!\perp 4|3, \quad \{1, 2\} \perp\!\!\!\perp \{4, 5\}|3, \quad \text{or} \quad 2 \perp\!\!\!\perp 4|\{1, 3, 5\}.$$

For many variables, methods formulated for graphs alone considerably simplify the task of deciding whether an independence statement is implied or not. These are called separation criteria; see Geiger et al. (1990), Lauritzen et al. (1990) and Marchetti and Wermuth (2009) for criteria on directed acyclic graphs.

Directed acyclic graph are the most direct generalization of Markov chains. They have an ordered sequence of single nodes representing responses that may generate f_N, but each response may remember any subset or all of the variables in its past. Directed acyclic graphs are known as Bayesian networks when the node set consists of discrete random variables that correspond to features of observable units, but it may also include decisions or parameters.

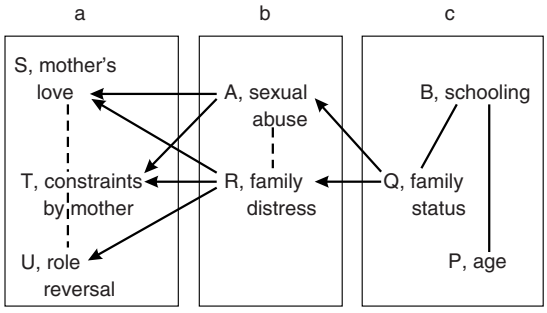
For ordered sequences of vector responses, the graphs are chains of joint responses and the associations and independences of the individual components of each response are represented by undirected edges being present or missing; see Cox and Wermuth (1993, 1996), Lauritzen (1996), Edwards (2000), Drton (2009), Marchetti and Lupparelli (2010), Wermuth and Cox (2004), Whittaker (1990).

The following small example of a well-fitting multivariate regression chain is for a set of data of Jochen Hardt, University of Mainz, on $n = 283$ adult, female patients who agreed to be interviewed at the offices of their general practitioners about different aspects of their childhood. Variables A, B are binary, the others are based on quantitative measurements. Each of Y_a and Y_c have three components and Y_b has two.

The graph is constructed after checking for nonlinear and interactive effects by using the results of a sequence of linear and logistic regressions (see ▶Logistic Regression). These show that the estimated dependencies, not displayed here, are in the direction hypothesized by the researchers. The background variable Y_c does not improve prediction of Y_a given the more specific information about childhood of Y_b.

The resulting factorization is $f_N = f_{a|b} f_{b|c} f_c$. The independences defining the regression chain graph are $S \perp\!\!\!\perp U|\{a, b\}$, $a \perp\!\!\!\perp c|b$, $b \perp\!\!\!\perp BP|Q$ and $Q \perp\!\!\!\perp P|B$, where relations within a are modeled using a covariance graph, those within b using a concentration graph.

For a complete linking of chain graphs and corresponding densities f_N, it is necessary to assure that independence statements satisfied by f_N combine in the same way as for the graphs. This requires special additional properties of the graphs, of the process by which the joint densities f_N are generated or directly of f_V; for discussions of special properties see Dawid (1979), Lauritzen (1996), Studený (2005), Kang and Tian (2009), San Martin et al. (2005), Wermuth (2010).



It is the outstanding feature of many graphical Markov models that consequences of a given model can be derived, for instance regarding implications after marginalizing over some variables, in set M, or after conditioning on others, in set C. In particular, graphs can be obtained for node set $N' = N/\{C, M\}$ which capture precisely the independence structure implied by a generating graph in node set N for $f_{N'|C}$ the density of Y'_N given Y_C, of the distribution of the variables in the reduced node set N'.

Such graphs are named independence-preserving, when they can be used to derive the independence structure that would have resulted from the generating graph by conditioning on a larger node set $\{C, c\}$ or by marginalizing over a larger node set $\{M, m\}$.

Three corresponding types of independence-preserving graphs are known which result from a given generating directed acyclic graph by using the same sets C, M:

graphs of the much larger class of MC-graphs of Koster (2002), maximal ancestral graphs (MAGs) of Richardson and Spirtes (2002) and summary graphs of Wermuth (2010); see Sadeghi (2009) for a proof of Markov equivalence that is for showing that the three corresponding but different types of graph capture the same independence structure.

More importantly, graphical criteria on the summary graph show when a generating conditional dependence of Y_i on Y_k, say, in f_N remains undistorted in $f_{N'|C}$ parametrised, as in a MAG model, in terms of conditional dependences within N' and when it be may be severely distorted; see also Wermuth and Cox (2008). Some of these distortions cannot be avoided for generating processes with randomized allocation of individuals to the levels of Y_k, but possibly by changing C or M. Thus, these results are relevant for controlled clinical trials, for meta analyzes and, more generally, for the planning stage of studies designed to replicate some of the given results of a larger study using a subset of the variables or a subpopulation.

More results on Markov equivalence, on estimation and goodness of fit tests, more direct applications as well as uses of the results concerning distortions and causal interpretations of graphical Markov models are expected in the near future; see also Drton et al. (2009), Cox (2007), Cox and Wermuth (2004). Comparative evaluations will be needed of alternative computational methods that are in use now for very large sets of data; see Balzarini (2007), Edwards et al. (2010), Dobra (2009), Meinshausen and Buhlmann (2006), Wainwright and Jordan (2008).

About the Author

For biography *see* the entry ►Multivariate Statistical Analysis.

Cross References

►Causal Diagrams
►Markov Chains
►Multivariate Statistical Analysis

References and Further Reading

Balzarini M (2007) Improving cluster visualization in self-organizing maps: application in gene expression data analysis. Comput Biol Med 37:1677–1689

Cox DR (2007) Principles of statistical inference. Cambridge University Press, Cambridge

Cox DR, Wermuth N (1993) Linear dependencies represented by chain graphs Stat Sci 8:204–218, 247–277; (with discussion)

Cox DR, Wermuth N (1996) Multivariate dependencies: models, analysis, and interpretation. Chapman & Hall, London

Cox DR, Wermuth N (2004) Causality: a statistical view. Int Stat Rev 72:285–305

Dawid AP (1979) Some misleading arguments involving conditional independence. J R Stat Soc B 41:249–252

Dobra A (2009) Variable selection and dependency networks for genomewide data. Biostatistics 10:621–639

Drton M (2009) Discrete chain graph models. Bernoulli 15:736–753

Drton M, Eichler M, Richardson TS (2009) Computing maximum likelihood estimates in recursive linear models. J Mach Learn Res 10:2329–2348

Edwards D (2000) Introduction to graphical modelling, 2nd edn. Springer, New York

Edwards D, de Abreu GCG, Labouriau R (2010) Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. BMC Bioinformatics 2010:11–18

Geiger D, Verma TS, Pearl J (1990) Identifying Independence in Bayesian Networks. Networks 20:507–534

Kang C, Tian J (2009) Markov properties for linear causal models with correlated errors. J Mach Learn Res 10:41–70

Koster JTA (2002) Marginalising and conditioning in graphical models. Bernoulli 8:817–840

Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford

Lauritzen SL, Dawid AP, Larsen B, Leimer HG (1990) Independence properties of directed Markov fields. Networks 20:491–505

Marchetti GM, Lupparelli M (2010) Chain graph models of multivariate regression type for categorical data. Submitted by: Bernoulli, to appear

Marchetti GM, Wermuth N (2009) Matrix representations and independencies in directed acyclic graphs. Ann Stat 47:961–978

Meinshausen N, Buhlmann P (2006) High dimensional graphs and variable selection with the Lasso. Ann Stat 34:1436–1462

Richardson TS, Spirtes P (2002) Ancestral graph Markov models. Ann Stat 30:962–1030

Sadeghi K (2009) Representing modified independence structures. Transfer thesis, Oxford University

San ME, Mochart M, Rolin JM (2005) Ignorable common information, null sets and Basu's first theorem. Sankhyā 67:674–698

Studený M (2005) Probabilistic conditional independence structures. Springer, London

Wainwright MJ, Jordan MI (2008) Graphical models, exponential families, and variational inference. Found Trends Mach Learn 1:1–305

Wermuth N (2010) Probability distributions with summary graph structure. Submitted by: Bernoulli, to appear. arXiv:1003.3259v1

Wermuth N, Cox DR (1998) On association models defined over independence graphs. Bernoulli 4:477–495

Wermuth N, Cox DR (2004) Joint response graphs and separation induced by triangular systems. J R Stat Soc B Stat Methodol 66:687–717

Wermuth N, Cox DR (2008) Distortions of effects caused by indirect confounding. Biometrika 95:17–33

Wermuth N, Lauritzen SL (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models. J R Stat Soc B 52:21–75, with discusssion

Wermuth N, Marchetti GM, Cox DR (2009) Triangular systems for symmetric binary variables. Electron J Stat 3:932–955

Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester

H

Handling with Missing Observations in Simple Random Sampling and Ranked Set Sampling

Carlos N. Bouza Herrera[1], Dante Covarrubias Melgar[2], Zoila Fernández[3]
[1]Professor
Universidad de La Habana, Habana, Cuba
[2]Universidad Autónoma de Guerrero, Mexico City, Mexico
[3]Professor
Universidad Católica del Norte, Antofagasta, Chile

Nonresponse in Simple Random Sampling

The existence of missing observations is a very important aspect to be considered in the applications of survey sampling; see Rueda and González (2004) for example. In human populations they may be motivated by a refusal of some interviewed persons to give the true value of Y. Hansen and Hurwitz (1946) proposed to select a subsample among the nonrespondents; see Cochran (1977). This feature depends heavily on the proposed subsampling rule. Alternative sampling rules to Hansen–Hurwitz's rule have been proposed; see for example Srinath (1971) and Bouza (1981). Theoretically, we deal with a particular case of double sampling (DS). It is described, when the sampling design is simple random sampling, as follows:

Step 1 Select a sample s from U and evaluate Y among the respondents (determine $\{y_i : i \in s_1 \subset U_1,\ |s_1| = n_1\}$).

Step 2 Determine $n_2' = \theta n_2$, $0 < \theta < 1$; $|s_2| = n_2$ with $s_2 = s \backslash s_1$.

Step 3 Select a subsample s_2' of size n_2' from s_2 and evaluate Y among the units in s_2' $\{y_i : i \in s_2' \subset s_2,\ s_2 \subset U_2\}$.

Step 4 Compute $\bar{y}_1 = \dfrac{\sum_{i=1}^{n_1} y_i}{n_1}$, $\bar{y}_1' = \dfrac{\sum_{i=1}^{n_2'} y_i}{n_2}$, and the estimate of μ $\bar{y} = \dfrac{n_1}{n}\bar{y}_1 + \dfrac{n_2}{n}\bar{y}_2' = w_1\bar{y}_1 + w_2\bar{y}_2'$.

This estimator is unbiased for the population mean. Using the techniques provided by double sampling (see Cochran 1977), the expected error of the estimator is given by $EV(\bar{y}) = \dfrac{\sigma^2}{n} + \dfrac{W_2(1-\theta)\sigma_2{}^2}{\theta}$. Commonly, Hansen–Hurwitz's rule is presented in textbooks (e.g., Cochran 1977), where $\theta = 1/K$ is the subsampling parameter and $EV(\bar{y}) = \dfrac{\sigma^2}{n} + \dfrac{W_2(K-1)\sigma_2{}^2}{n}$.

For other designs the DS procedure is used, and particular alternative estimators, when missing observations are present, must be derived. See, for example, a proposal for product-type estimators in Bouza (2008).

Nonresponse in Ranked Set Sampling

An alternative sampling design is ranked set sampling (RSS). It was first proposed by McIntyre (1952). Volume 12 of the *Handbook of Statistics* dedicated a section to RSS; see Patil et al. (1994). The basic procedure of RSS works as follows:

Step 1 Select m samples of size m using *SRS* with replacement independently.

Step 2 Each unit in the $s_{(t)}$, $t = 1, \ldots, m$, is ranked and the ▶order statistics (OS) $Y_{(1:t)}, \ldots, Y_{(m:m)}$ measured.

Step 3 Repeat the procedure r times and compute $\sum_{j=1}^{r}\sum_{i=1}^{m} Y_{(i:i)j}/rm = \mu_{(s)}$.

Each sampled unit may be ranked without measuring Y using some judgment or an auxiliary variable. See David and Levine (1972). Note that we measure the OS of order t $(t = 1, \ldots, m)$ in each tth sample in each cycle $(j = 1, \ldots, r)$ but the ranks do not intervene in the selection of the sample. As the procedure is repeated r times the sample size is $n = rm$ and the estimator of the mean is unbiased. The error of it is $V[\mu_{(s)}] = \sum_{t=1}^{n}\sigma_{(i)}^2/n^2 = V[\mu_{(s)}] - \sum_{t=1}^{n}\Delta_{(i)}^2/n^2$, where $\mu_{(i)} - \mu = \Delta_{(i)}$. Hence, it is smaller than the error of $\bar{y}$.

In the presence of missing observations and the use of DS for dealing with the nonresponses, a subsample strategy is to select a subsample s_{12}' of size $m(i,2)$ from each s_{i2}, $i = 1, \ldots, r$. The development of the corresponding theory of

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

DS for RSS can be consulted in Al-Saleh and Al-Kadiri (1996). Bouza (2002a) proposed to use

$$\mu_{rss(nr)} = \sum_{k=1}^{r} M(i)/r,$$

where

$$M(i) = w(i,2)\left[\sum_{u=1}^{m(i,2)} Y'(i,u)/m(i,2)\right] + w(i,1)\left[\sum_{u=1}^{m} Y^*(i,u)/m(i,1)\right].$$

Defining $w(i,t) = m(i,t)/m$, $Y'(i,u)$ as the value of Y in the uth unit of s'_{12}, and $Y^*(i,u) = y_{u(u)}$ if the unit with rank u in the u-ranked set responds and zero otherwise. The use of DS showed that the estimator is unbiased and that the expected variance is $EV[\mu_{rss(nr)}] = V + G$, where $V = \frac{\sigma^2}{n} + \frac{W_2(1-\theta)\sigma_2^2}{n\theta}$ and $G = \Delta_1 - \Delta_2$, defining $\Delta_1 = \sum_{j=1}^{m}(\mu_{(j)} - \mu)^2/m$ and $\Delta_2 = \sum_{i=1}^{r} E\left[\sum_{j=1}^{m(i,2)} (\mu_{(j)} - \mu)^2\right]/n$. Hence, the use of RSS is more accurate than the use os srswr also when the nonrespondent sample is subsampled for solving the existence of missing observations in the sample.

Other results in this line are in progress; see for example Bouza (2008).

About the Author

Professor Carlos Narciso Bouza Herrera obtained his doctorate at the University of Belgrade in 1978 (former Yugoslavia, now Serbia). He teaches at the Faculty of Mathematics and Computer Science at the University of Havana. He is the (co)author of over 50 publications that have appeared in Cuba, Mexico, Brazil, Venezuela, Spain, France, Germany, and Pakistan.

Cross References

►Bias Analysis
►Imputation
►Maximum Entropy Method for Estimation of Missing Data
►Multiple Imputation
►Ranked Set Sampling
►Simple Random Sample

References and Further Reading

Al-Saleh MF, Al-Kadiri MAA (1996) Double ranked set sampling. Stat Prob Lett 48:205–214
Bouza CN (1981) Sobre el problema de la fracciòn de muestreo para el caso de las no respuestas. Trabajos de Estadística 21:18–24
Bouza CN (2002a) Estimation of the mean in ranked set sampling with nonresponses. Metrika 56:171–179
Bouza CN (2002b) Ranked set sampling the non-response stratum for estimating the difference of means. Biom J 44:903–915
Bouza CN (2008a) Estimation of the population mean with missing observations using product type estimators. Revista Investigación Operacional 29:207–223
Bouza CN (2008b) Ranked set sampling with missing observations: the estimation of the population mean and the difference of means. Model Assist Stat Appl 3:127–138
Cochran WG (1977) Sampling techniques. Wiley, New York
David HA, Levine DN (1972) Ranked set sampling in the presence of judgement error. Biometrics 28:553–555
Hansen MH, Hurwitz WN (1946) The problem of non responses in survey sampling. J Am Stat Assoc 41:517–523
McIntyre GA (1952) A method of unbiased selective sampling using ranked sets. J Agric Res 3:385–390
Patil GP, Sinha AK, Taillie C (1994) Ranked set sampling. In: Patil GP, Rao CR (eds) Handbook of Statistics, Environmental Statistics, vol 12. North-Holland, Amsterdam
Rueda M, González S (2004) Missing data and auxiliary information in surveys. Comput Stat 19:551–567
Srinath KP (1971) Multi-phase sampling in non-response problems. J Am Stat Assoc 66:583–589

Harmonic Mean

Jasmin Komić
Professor, Faculty of Economics
Banja Luka, University of Banja Luka, Republic of Srpska, Bosnia and Herzegovina

In the time of Pythagoras, there were only three means (Bakker 2003; Brown 1975; Huffman 2005), the arithmetic, the geometric, and third that was called subcontrary, but the "name of which was changed to harmonic by Archytas of Tarentum and Hippasus and their followers, because it manifestly embraced the ratios of what is harmonic and melodic" (Huffman 2005, p. 164). The harmonic mean is a measure of location used mainly in particular circumstances – when the data consists of a set of rates, such as prices ($/kilo), speeds (mph), or productivity (output/manhour). lt is defined as the reciprocal of the arithmetic mean of the reciprocals of the values.

The harmonic mean of n numbers $x_1, x_2, \ldots, x_n$ is calculated in the following way:

$$\bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

As a simple example, the harmonic mean of three numbers, 2, 5, and 10 is equal to

$$\bar{x}_H = \frac{3}{\frac{1}{2} + \frac{1}{5} + \frac{1}{10}} = \frac{3}{\frac{8}{10}} = 3.75.$$

The harmonic mean in this example is less then the arithmetic mean, 5.67. This can be generalized by saying that for any data set that shows variability and does not contain zero value, the harmonic mean will always be smaller than both the arithmetic mean and the geometric mean (for the precise inequality statement see the entry ►Geometric Mean).

Like the arithmetic and geometric means, harmonic mean is based on all observations. If any value of the data set equals zero, the harmonic mean cannot be calculated. Harmonic mean is sensitive to ►outliers when they have much smaller values than the rest of the data, and largely insensitive to outliers that have much larger values than the other data (the reciprocal of a large number is small, and the reciprocal of a small number is large, relatively).

The interested readers are urged to consult Ferger (1931), Francis (2004), Haans (2008), and Hand (1994) about the problems and confusion with the proper usage of the harmonic mean. Probably they will support the following claim given by Ya-Lun Chou (1989) "Because of the absolute necessity of using the harmonic mean in some cases and the confusion between the application of the arithmetic and harmonic averages, the harmonic mean deserves more attention than it receives in most elementary textbooks."

It is important to notice that extreme care must be taken when averages of rates are calculated. Since a rate is always expressed in terms of the ratio of two units (e.g., miles/gallon or price/kg), the criterion for choosing between arithmetic and harmonic means can be stated as follows (Ferger 1931; Francis 2004):

1. *Harmonic* mean is appropriate if the rates are being averaged over *constant numerator* units.
2. *Arithmetic* mean should be used if the rates are being averaged over *constant denominator* units.

Example 1 We want to compare average productivity (in items/day) of two production lines which are both producing the same items. Following abovementioned rule, the arithmetic mean should be used if the productivity for each line is measured over, say, 1 day (i.e., the same time, thus making the denominator units constant for both lines). However, if the productivity for each line is measured over, say, the production of a single item (i.e., the same quantity, thus making the numerator units constant for both lines), the harmonic mean is appropriate.

Suppose that three workers, A, B, and C, in a textile factory can make 6, 5, and 4 T-shirts per hour, respectively. Their productivity can be recorded in either one of the following ways.

Format 1

Worker A: 6 per hour
Worker B: 5 per hour
Worker C: 4 per hour

Format 2

Worker A: 10 min per T-shirt
Worker B: 12 min per T-shirt
Worker C: 15 min per T-shirt

In the first format, it is only appropriate to use the arithmetic mean to find the average productivity since time (denominator in the rate expression items/h) is held as a constant. The question here is: what is the average output per 1 h? Therefore, the average productivity equals 5 (15/3) T-shirts per hour. This means, that if 5 T-shirts are produced, on average, the output of all three workers will be 15 T-shirts in an hour.

However, in the second format, production (nominator in unit/h) is treated as constant and time as a variable. According to the postulated rule, only the harmonic mean will reflect the true average productivity. Now the question is: what is the average time required to complete one unit of product? Thus,

$$\bar{x}_H = \frac{3}{\frac{1}{10} + \frac{1}{12} + \frac{1}{15}} = 12 \text{ min per T-shirt.}$$

Using the suggested rule, it is easy to solve the efficiency paradox introduced by David Hand (1994, see also Haans [2008]), where two groups of engineers are in disagreement about the average fuel efficiency of a set of cars. One group, coming from England, measured efficiency on a miles per gallon scale, the other, coming from France, on a gallons per mile scale. When the arithmetic means are applied to both rates, English and French engineers come to the illogical, opposite conclusions. The paradox disappears if the average of the data in the gallons per mile scale is calculated by the harmonic mean. The reason is that the numerator of the efficiency rate (gallon/mile) has to be treated as a constant, since the question is how many miles can be passed with a single gallon.

Cross References

►Geometric Mean
►Mean, Median, Mode: An Introduction

References and Further Reading

Bakker A (2003) The early history of average values and implications for education. J Stat Educ 11:1
Brown M (1975) Pappus, Plato and the harmonic mean. Phronesis 20(2):173–184
Chou Y (1989) Statistical analysis with business and economic applications. Elsevier, New York
Ferger WF (1931) The nature and use of the harmonic mean. J Am Stat Assoc 26(173):36–40
Francis A (2004) Business mathematics and statistics, 6th edn. Cengage Learning Business Press
Haans A (2008) What does it mean to be average? The miles per gallon versus gallons per mile paradox revisited. Pract Assess Res Eval 13(3)
Hand DJ (1994) Deconstructing statistical questions (with discussion). J R Stat Soc A 157:317–356
Huffman CA (2005) Archytas of Tarentum: Pythagorean, philosopher, and mathematician king. Cambridge University Press, Cambridge

Hazard Ratio Estimator

Per Kragh Andersen
Professor
University of Copenhagen, Copenhagen, Denmark

In survival analysis, statistical models are frequently specified via the hazard function $\alpha(t)$. A simple model for the relation between the hazard functions in two groups (e.g., a treatment group 1 and a control group 0) is the *proportional hazards model* where

$$\alpha_1(t) = \theta\alpha_0(t), \tag{1}$$

and θ is the treatment effect. For a parametrically specified baseline hazard, $\alpha_0(t)$, both the treatment effect and the parameters in the baseline hazard are usually estimated using maximum likelihood. In a semi-parametric model where the baseline hazard is left unspecified several estimators for θ are available: the maximum partial likelihood estimator, cf. Cox (1972), a class of rank estimators, and some ad hoc estimators.

Assume that the available data are $(X_{ij}, D_{ij}; i = 1,\ldots,n_j, j = 0,1)$ where the X_{ij} are the times of observation: a failure time if the corresponding indicator D_{ij} is 1, a right-censoring time if D_{ij} is 0. The Cox estimator, $\widehat{\theta}$, is then the solution to the equation

$$O_1 = E_1(\theta) \tag{2}$$

where, for $j = 0,1$, $O_j = \sum_i D_{ij}$ and

$$E_1(\theta) = \sum_{ij} \frac{Y_1(X_{ij})\theta}{Y_0(X_{ij}) + Y_1(X_{ij})\theta} D_{ij}.$$

Here, $Y_j(t) = \sum_i I(X_{ij} \geq t)$ is the number at risk at time $t-$ in group $j, j = 0,1$. Notice that (2) expresses that for $\theta = \widehat{\theta}$, the observed number, O_1, of failures in group 1 should be equal to a corresponding "expected" number, $E_1(\theta)$ under the proportional hazards assumption.

A class of explicit "rank" estimators, discussed by Andersen et al. (1993, Chap. V) is, for a given *weight process* $L(t)$, given by

$$\widehat{\theta}_L = \frac{\sum_{i=1}^{n_1} L(X_{i1}) \frac{D_{i1}}{Y_1(X_{i1})}}{\sum_{i=1}^{n_0} L(X_{i0}) \frac{D_{i0}}{Y_0(X_{i0})}}. \tag{3}$$

For $L(t) = I(t \leq t^*)$, $\widehat{\theta}_L$ is simply the ratio between the Nelson-Aalen estimators for the cumulative hazards in groups 1 and 0 evaluated at t^*. The Cox estimator, $\widehat{\theta}$ given by (2) is always less dispersed than any $\widehat{\theta}_L$ given by (3). Using an estimator $\widehat{\theta}_L$ and its estimated variance, the hypothesis $\theta = 1$ of no treatment effect may be tested. This gives all the standard linear non-parametric two-sample tests for ▶survival data and, in particular, the weight process given by

$$L(t) = \frac{Y_0(t)Y_1(t)}{Y_0(t) + Y_1(t)},$$

gives the logrank test.

Another explicit ad hoc estimator, discussed by Breslow (1975), is given by

$$\widetilde{\theta} = \frac{O_1/E_1(1)}{O_0/E_0(1)}$$

with

$$E_0(\theta) = \sum_{ij} \frac{Y_0(X_{ij})}{Y_0(X_{ij}) + Y_1(X_{ij})\theta} D_{ij}.$$

The estimator $\widetilde{\theta}$ is generally inconsistent when $\theta \neq 1$ but it has gained some popularity due to its simplicity and close connection to the logrank test which is also based on the observed, O_0 and O_1, and expected, $E_0(1)$ and $E_1(1)$, numbers of failures.

About the Author

For biography *see* the entry ▶Hazard Regression Models.

Cross References

▶Hazard Regression Models
▶Modeling Survival Data
▶Survival Data

References and Further Reading

Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Breslow NE (1975) Analysis of survival data under the proportional hazards model. Int Stat Rev 43:45–58

Cox DR (1972) Regression models and life-tables (with discussion). J Roy Stat Soc B 34:187–220

Hazard Regression Models

Per Kragh Andersen
Professor
University of Copenhagen, Copenhagen, Denmark

In many applications of survival analysis the interest focuses on how *covariates* may affect the outcome. In clinical trials, adjustment of treatment effects for effects of other explanatory variables may be crucial if the randomized groups are unbalanced with respect to important prognostic factors, and in epidemiological cohort studies reliable effects of exposure may be obtained only if some adjustment is made for confounding variables. In these situations, a *regression model* is useful.

Most regression models for ▶survival data are set up via the *hazard function*, e.g., Andersen et al. (1993, Chap. VII), and the most important such model is the Cox (1972) *proportional hazards regression model*.

Cox Regression Model

In its simplest form the Cox model states the hazard function for an individual, i, with covariates $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})'$ to be

$$\alpha_i(t; \boldsymbol{Z}_i) = \alpha_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown regression coefficients and $\alpha_0(t)$, the *baseline hazard*, is the hazard function for individuals with all covariates equal to 0. Thus, the baseline hazard describes the common shape of the survival time distributions for all individuals while the *hazard ratio* function $\exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)$ gives the level of each individual's hazard. The interpretation of the parameter, β_j for a dichotomous $Z_{ij} \in \{0, 1\}$ is that $\exp(\beta_j)$ is the hazard ratio for individuals with $Z_{ij} = 1$ compared to those with $Z_{ij} = 0$ all other covariates being the same for the two individuals. Similar interpretations hold for parameters corresponding to covariates taking more than two values. The model is semi-parametric in the sense that the hazard ratio part is modeled parametrically while the baseline hazard is left unspecified.

Assume that the available data are $(X_i, D_i, \boldsymbol{Z}_i; i = 1, \ldots, n)$ where the X_i are the times of observation: a failure time if the corresponding indicator D_i is 1, a right-censoring time if D_i is 0. The regression coefficients $\boldsymbol{\beta}$ are then estimated by maximizing the *Cox partial likelihood*

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}' \boldsymbol{Z}_i)}{\sum_{j \in R_i} \exp(\boldsymbol{\beta}' \boldsymbol{Z}_j)} \right]^{D_i} \tag{2}$$

where $R_i = \{j : X_j \geq X_i\}$, the *risk set* at time X_i, is the set of individuals still alive and uncensored at that time. Furthermore, the cumulative baseline hazard $A_0(t)$ is estimated by the *Breslow estimator*

$$\widehat{A}_0(t) = \sum_{X_i \leq t} D_i / \sum_{j \in R_i} \exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{Z}_j). \tag{3}$$

In large samples, $\widehat{\boldsymbol{\beta}}$ is approximately normally distributed with the proper mean and with a covariance which is estimated by the information matrix based on (2), see e.g., Andersen et al. (1993, Chap. VII). This means that approximate confidence intervals for the hazard ratio parameters can be calculated and that the usual large sample test statistics based on (2) are available. Also, the asymptotic distribution of the Breslow estimator is normal; however, this estimate is most often used as a tool for estimating *survival probabilities* for individuals with given covariates, $\boldsymbol{Z}_0$. Such an estimate may be obtained by the product integral $\widehat{S}(t; \boldsymbol{Z}_0)$ of $\exp(\widehat{\boldsymbol{\beta}}' \boldsymbol{Z}_0) \widehat{A}_0(t)$. The joint asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ and the Breslow estimator then yields an approximate normal distribution for $\widehat{S}(t; \boldsymbol{Z}_0)$ in large samples.

A number of useful extensions of this simple Cox model are available. Thus, in some cases the covariates are time-dependent, e.g., a covariate might indicate whether or not a given event had occurred by time t, or a time-dependent covariate might consist of repeated recordings of some measurement likely to affect the prognosis. In such cases the regression coefficients $\boldsymbol{\beta}$ are estimated replacing $\exp(\boldsymbol{\beta}' \boldsymbol{Z}_j)$ in (2) by $\exp[\boldsymbol{\beta}' \boldsymbol{Z}_j(X_i)]$. Also a simple extension of the Breslow estimator (3) applies in this case.

Another extension of (1) is the stratified Cox model where individuals are grouped into strata each of which has a separate baseline hazard. This model has important applications for checking the assumptions of (1). The model assumption of proportional hazards may also be tested in a number of ways, the simplest possibility being to add interaction terms of the form $Z_{ij} f(t)$ between Z_{ij} and time where $f(t)$ is some specified function. Also various forms of residuals as for normal linear models may be used for model checking in (1). In (1) it is finally assumed that a quantitative covariate affects the hazard log-linearly. This

assumption may also be checked in several ways and alternative models with other hazard ratio functions $r(\boldsymbol{\beta}'\boldsymbol{Z}_i)$ may be used.

Other Hazard Models

Though the Cox model is the regression model for survival data which is applied most frequently, other hazard regression models, e.g., *parametric* regression models also play important roles in practice. Examples include models with a multiplicative structure, i.e., models like (1) but with a parametric specification, $\alpha_0(t) = \alpha_0(t;\boldsymbol{\theta})$, of the baseline hazard. A multiplicative model with important epidemiological applications is the *Poisson regression* model (see ►Poisson Regression) with a piecewise constant baseline hazard. In large data sets with categorical covariates this model has the advantage that a sufficiency reduction to the number of failures and the amount of person-time at risk in each cell defined by the covariates and the division of time into intervals is possible. This is in contrast to the Cox regression model (1) where each individual data record is needed when fitting the model.

An alternative to the multiplicative structure is provided by *additive hazard models* and the main such example is Aalen's additive model

$$\alpha_i(t;\boldsymbol{Z}_i) = \beta_0(t) + \boldsymbol{\beta}(t)'\boldsymbol{Z}_i. \tag{4}$$

Here, both the baseline hazard, $\beta_0(t)$ and the regression functions $\beta_1(t),\ldots,\beta_p(t)$ are left completely unspecified and estimated non-parametrically much like the Nelson-Aalen estimator. Semi-parametric versions of (4) also exist, that is models where some or all regression functions $\beta_j(t), j = 1,\ldots,p$, are constant. Such models, as well as more general and flexible models containing both (1) and (4) as special cases are discussed by Martinussen and Scheike (2006).

About the Author

Dr. Per Kragh Andersen is Professor of biostatistics at Department of Public Health, Section of Biostatistics, University of Copenhagen, Denmark. He was elected as member of ISI in 1990. He is an author or co-author of more than 300 publications, including the books *Statistical Analysis of Survival Data in Medical Research* (in Danish, 1984, with M. Vaeth), *Statistical Models Based on Counting Processes* (with Ø. Borgan, R.D. Gill and N. Keiding, Springer, 1993), *Survival and Event History Analysis* (with N. Keiding, editors, Wiley, 2006), *Regression with Linear Predictors* (Springer 2010, with L. T. Skovgaard). He is Past President of Danish Society of Theoretical Statistics and has been on the editorial boards of *Scandinavian Journal of Statistics*, *Statistical Methods in Medical Research*, *Statistics in Medicine*, *Biometrics* and *Lifetime Data Analysis*.

Cross References

►Demographic Analysis: A Stochastic Approach
►Hazard Ratio Estimator
►Likelihood
►Misuse of Statistics
►Modeling Survival Data
►Survival Data

References and Further Reading

Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Cox DR (1972) Regression models and life-tables (with discussion). J Roy Stat Soc B 34:187–220

Martinussen T, Scheike TH (2006) Dynamic regression models for survival data. Springer, New York

Heavy-Tailed Distributions

Raoul LePage
Professor
Michigan State University, East Lansing, MI, USA

Throw a ball fast enough toward the horizon and it falls into orbit. Run fast enough to infinity on the positive real axis carrying a bucket leaking probability and you create a heavy-tailed probability distribution (HTD). Sample an HTD and you encounter not one but an infinite supply of instances in which an observation far exceeds all predecessors. Defining this phenomenon mathematically is a matter of deciding how fast to run. Run too fast and you wind up with tails levering their probability so very far out that you wait through extremely many ordinary samples before encountering the next "big one" and its a whopper! The following definition, popular although not pleasing everyone, is slow enough to provide a rich source of HTD but fast enough to capture most of the desired properties.

Definition of HTD

A probability distribution specified through its cumulative distribution function F on the real line is heavy right-tailed if and only if for every $t > 0$ the ratio $\frac{1-F(x)}{e^{-tx}}$ has an infinite limit as x tends to infinity. Such an F has right-tail probabilities $1-F(x)$ decaying to zero ever more slowly than any exponential distribution. A probability distribution is heavy left-tailed if $F(-x)$ is heavy right-tailed.

A probability distribution is heavy-tailed if it is heavy-tailed in at least one direction. An interesting class of HTD has Pareto-like tails.

$$1 - F(x) \sim x^{-p}, \text{ as } x \to \infty \tag{1}$$

for some $p > 0$. Pareto distributions are the case $F(x) = 1 - x^{-p}$ for all $x > 1$. By comparison, for a standard normal distributed random variable Z we have $P(Z > z) \sim \frac{e^{-z^2/2}}{z}$ as z tends to infinity, definitely not heavy-tailed.

HTD in Actual Use

Many processes appear to occasionally but indefinitely produce far larger (or smaller) values than everything seen before, the much wilder market swing, the unusually heavy burst of Internet activity. This may be attributed to specific causes, e.g., growth of large pools of risky investments, but such pools may themselves be cause for market swings to behave as samples from HTD. Now alert to the possibilities we are witness to a great many examples of empirical data looking for all the world like HTD. Most of these examples come from data-rich applications such as financial transactions, communications, spatial data, and the like. Resnick (1997) is a serious undertaking of this kind.

The link between heavy-tailed distributions and the phenomenon of out-sized samples is sufficiently colorful to have inspired some tendency to heavy-tailed this and that. HTD may refer to a probability distribution on any space whose "tail" probabilities (in some sense) are large, or to a random process whose distribution is HTD. The term heavy-tailed is sometimes applied also to particular random processes incorporating time or spatial dependencies but nonetheless exhibiting outsized observations.

This outsized samples phenomenon may be studied from a sequential view (a running account of data arrivals) or instead through distributional results saying that for a large sample the data will likely exhibit outsized values.

Remarks

With HTD we are puzzling over ideas outside the more regular world of normal distributions and their spawn. New thinking is continually required. We see remarkable progress in these matters. There are features of heavy-tailed behavior that might be highlighted, among them the role of conditioning. A good example to raise this point is found in the stationary symmetric a stable (SaS) process defined by the following stochastic integral integral with respect to a symmetric alpha stable ($S\alpha S$) Lévy motion Z on the interval $[-\pi/2, \pi/2]$ with $0 < \alpha < 2$

$$X(t) = \int e^{it\lambda} \mathrm{d}Z(\lambda), \quad t \in \mathbb{R}.$$

Through a *series construction* of this integral one may build the jumps of $\mathrm{d}Z$ out of the consecutive arrivals of a unit rate Poisson process on the half line. If jumps of $\mathrm{d}Z$ are each multiplied by a standard normal deviate, these being independent and identically distributed independently of Z, the distribution of the process X is unchanged except for scale. If we then condition on the jumps of the original $\mathrm{d}Z$ the result is that process X has conditionally a stationary normal distribution (with discrete spectral measure). Utilizing this fact and Bayesian-like calculations it is found for $d > 0$ that the conditional expectations $E(X(t) \mid X(t-d))$ and $E(X^2(t) \mid X(t-d), X(t-2d))$ are both finite almost surely (LePage 1987 includes 1980). Such is the dependency in this particular heavy-tail process that even one or two observations remove heavy-tails from consideration. This shows us that unconditional distributions may not play quite the accustomed role in heavy-tailed processes, for as observations are brought into the picture it is not always just a matter of using them to form estimates to be plugged into unconditional distributions of interest. Returning to $S\alpha S$ in any dimension one has a similar conditionally normal construction. Bayesian-like thinking can produce good estimators and predictors amenable to ►Markov Chain Monte Carlo without the need to solve for stable densities as some practitioners attempt in these problems.

The foregoing example is closely tied with what might be termed "granularity" of heavy tail random phenomena. I refer here to the central role played by individual extremes and in particular their capacity not only shift a trajectory but to radically alter the random experience. An example would be a surge of water great enough to precipitate the opening of a channel, eventually to alter an entire watercourse. Such can be achieved by models rooted in normal distributions and kin, but the necessarily volatile variances needed in that approach may sometimes only prove a distraction as set against direct description via the jumpy Poisson-like moves of heavy tail models. The two approaches do however appear to overlap considerably and we may expect a continuing supply of insights from their continuing interaction.

Worth mentioning also are heavy tail models of small scale events, for if you think about it the whole line (or space) can be compressed into arbitrarily small intervals (or patches) carrying heavy tail phenomena along with them.

We seek ways to model these rarities, how and why they tend to come, whether isolated or clustered, how foretold, their influence on various random behaviors and how we are informed by theory relative to a host of practical questions.

Resources

As mentioned above, HTD are increasingly being studied in connection with applications as diverse as dispersal of contaminants, river flow, large financial movements, bursts of insurance claims, internet traffic surges, and bottlenecks in queuing systems. These mostly arise in large systems ripe for study in the information rich environment of today. What is emerging focuses on probability models whose random constituents have tails sufficiently heavy as to produce effects like those alluded to above. The main objectives cannot be perfectly organized, but permit me to include only a few from the many excellent sources:

(a) Understanding the properties of existing statistical models when random components are heavy tailed, in particular propagation of heavy tail behaviors through systems (Adler, Feldman, Gallagher 1998).
(b) Developing models to match behaviors seen in the applied contexts above, in particular periods of benign behavior punctuated by great excursions (Resnick 1997).
(c) Extending and adapting statistical methods with which to effectively fit and guide statistical inference in heavy-tailed models (Calder and Davis 1998) (LePage, Podgorski, Ryznar 1997).
(d) Learning how a potential for future instabilities might be recognized in a system currently benign but revealing its character through other behaviors (various forms of the Hill estimator, see Resnick 1997).
(e) Developing theories of extreme behavior transcending the perspectives of specialized models Resnick (2007).
(f) Discovering new classes of probability distributions issuing from models or solving mathematical problems by linking their solution to heavy tailed variants of familiar processes such as diffusions (Baeumer, Meeschaert, Nane 2009).

Cross References

►Skewness
►Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts
►Statistical Modeling of Financial Markets

References and Further Reading

Adler R, Feldman R, Gallagher C (1998) Analyzing stable time series. In: Adler R, Feldman R, Taqqu M (eds) A practical guide to heavy tails. Birkhäuser, Boston, pp 133–158
Baeumer B, Meeschaert M, Nane E (2009) Space-time duality for fractional diffusion. J Appl Probab 46(4):1100–1115
Calder M, Davis R (1998) Inference for linear processes with stable noise. In: Adler R, Feldman R, Taqqu M (eds) A practical guide to heavy tails. Birkhäuser, Boston, pp 159–176
Davis R, Mikosch T (2009) Extreme value theory for GARCH processes. In: Andersen TG, Davis RA, Kreiß JP, Mikosch T (eds) Handbook of financial time series. Springer, New York, pp 187–200
LePage R (1987) Conditional moments for stable vectors. In: Cambanis S, Weron A (eds) Probability theory on vector spaces IV, Lecture Notes in Mathematics 1391. Springer-Verlag, New York, pp 148–163
LePage R, Podgorski K, Ryznar M (1997) Strong and conditional invariance principles for samples attracted to stable laws. Probab Theory Rel 108:281–298, Springer
Resnick S (1997) Heavy tail modeling and teletraffic data, Special Invited Paper. Ann Stat 25(5):1805–1869
Resnick S (2007) Heavy-tail phenomena, springer series in operations research and financial engineering, Springer, New York

Heteroscedastic Time Series

Biljana Č. Popović
Head of Department for Mathematical Statistics and Applications, Faculty of Sciences and Mathematics
University of Niš, Niš, Serbia

The trade-off between return and risk plays an important role in many financial models such as ►Portfolio theory and option pricing. So, economic theories in time series usually have implications on the conditional mean dynamics of underlying economic variables as well as on the volatility as the measure of risk.

Many models that are commonly used in empirical finance to describe returns and volatility are linear. There are, however, indications that ►nonlinear models may be more appropriate (as it can be seen, for instance, in Franses and van Dijk 2000). In order to model real data, a nonconstant error variance has to be incorporated.

A model that allows the possibility of nonconstant error variance is called a heteroscedastic model. A model of this kind was defined by Engle (1982). He estimated the means and variances of inflation in the UK. (Engle mentioned heteroscedastic time series previously in his paper Engle (1980).) His approach was as follows.

If the random variable Y_t that describes the state of (economic) space is dependent on the previous state of space, its conditional density function will be $f(y_t|y_{t-1})$. The forecast of today's value based on the past information is $E(Y_t|Y_{t-1})$ and the variance of one-period forecast is $Var(Y_t|Y_{t-1})$. This means that the forecast may be a random variable.

Consider the first-order autoregression

$$Y_t = \gamma Y_{t-1} + \varepsilon_t, \quad t \in D,$$

where D is the set of integers and (ε_t) is a white noise with $Var(\varepsilon_t) = \sigma_\varepsilon^2$ for any $t \in D$. The unconditional mean of Y_t is zero, while the conditional mean is γY_{t-1}. The unconditional variance of Y_t is $\frac{\sigma_\varepsilon^2}{1-\gamma^2}$ and the conditional variance is σ_ε^2, and the more general class of models seems desirable.

A preferable model is

$$Y_t = \varepsilon_t \sigma_t$$
$$\sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2,$$

with $E(\varepsilon_t) = 0$, $Var(\varepsilon_t) = 1$, and ε_t is independent of past realizations of Y_{t-i}, $i = 1, 2, \ldots$. This is an example of an autoregressive conditional heteroscedastic (ARCH) model. Specially, it is the ARCH(1) process. The variance function can be expressed more generally as

$$\sigma_t^2 = h(Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, \boldsymbol{\alpha}),$$

where p is the order of the ARCH process and $\boldsymbol{\alpha}$ is a vector of parameters.

If we set the information set available at time t to be $\mathcal{F}_{t-1}$, i.e., $\mathcal{F}_{t-1}$ denotes the σ-field generated by $\{Y_{t-1}, Y_{t-2}, \ldots\}$, it will be

$$Var(Y_t|\mathcal{F}_{t-1}) = E\left(Y_t^2|\mathcal{F}_{t-1}\right) = \sigma_t^2 = \\ = \alpha_0 + \alpha_1 Y_{t-1}^2 + \alpha_2 Y_{t-2}^2 + \cdots + \alpha_p Y_{t-p}^2.$$

If the normal distribution is specified, we will have the conditional density

$$Y_t|\mathcal{F}_{t-1} \sim \mathcal{N}\left(0, \sigma_t^2\right).$$

The ARCH model has a variety of characteristics that make it attractive for economic applications in particular. For instance, a large error through Y_{t-i}^2 gives rise to the variance, which tends to be followed by another large error. This phenomenon of volatility clustering is common in many financial time series.

When the conditional density is normal (with zero mean), the ARCH model has the following properties.

The pth-order linear ARCH process, with $\alpha_0 > 0$, $\alpha_1, \ldots, \alpha_p \geq 0$, is the second-order stationary (covariance stationary) if and only if the associated characteristic equation has all roots outside the unit circle. The stationary variance is given by $E\left(Y_t^2\right) = \alpha_0 / \left(1 - \sum_{j=1}^p \alpha_j\right)$.

Besides this often used condition, the next necessary and sufficient condition for the second-order stationarity of Y_t is valid.

The same pth-order linear ARCH process is the second-order stationary if and only if

$$\alpha_1 + \cdots + \alpha_p < 1.$$

Also, the last equation is a sufficient condition for strict stationarity and ergodicity of Y_t. (For more details, see, for instance, Li et al. 2002.)

It may be desirable to test whether the ARCH model is an appropriate one. In that case the Lagrange multiplier test can be applied. Generally, for time series data, the presence of heteroscedasticity would be tested at first.

A natural generalization of the ARCH process is to allow for the past conditional variance in the current conditional variance equation. This kind of generalization was done independently by Bollerslev (1986) and Taylor (1986). It is named the generalized autoregressive conditional heteroscedastic time series (GARCH). Volatility of the GARCH(p, q) is defined by

$$\sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2 + \alpha_2 Y_{t-2}^2 + \cdots + \alpha_q Y_{t-q}^2 \\ + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_p \sigma_{t-p}^2,$$

where

$$p \geq 0, \quad q > 0$$
$$\alpha_0 > 0, \quad \alpha_i \geq 0, \quad i = 1, \ldots, q,$$
$$\beta_j \geq 0, \quad j = 1, \ldots, p.$$

The necessary and sufficient condition for the second-order stationarity of the GARCH(p, q) is

$$\sum_{i=1}^q \alpha_i + \sum_{j=1}^p \beta_j < 1.$$

Clearly, when $p = 0$, the model reduces to the ARCH(q).

It is important to note that the regions of strict stationarity of these models are, in general, much larger than those of the second-order stationarity. The necessary and sufficient conditions for the strong stationarity of GARCH also can be displayed.

To estimate the unknown parameters of the GARCH (ARCH), one can use the Gaussian quasi-maximum likelihood and sometimes some other methods like the least square estimator. Note that some numeric procedure is necessary when quasi-maximum likelihood is applied. Statistical properties of the estimators will depend on the model itself.

If

$$\sum_{i=1}^q \alpha_i + \sum_{j=1}^p \beta_j = 1,$$

we deal with the so-called integrated GARCH(p, q) or the IGARCH(p, q) process.

To accommodate the asymmetric relation between many financial variables and their volatility and to relax the restriction on the coefficients in the model, EGARCH (exponential GARCH) was proposed. It was done by

Nelson (1991). The conditional variance of this model satisfies the equation

$$\ln\left(\sigma_t^2\right) = \gamma + \sum_{j=0}^{\infty} \mu_j g(a_{t-1-j}),$$

where γ is a constant, $\mu_0 = 1$, $a_t = \frac{Y_t}{\sigma_t}$, and the function g is chosen to allow for asymmetric changes depending on the sign of a_t. The coefficients μ_j are often assumed to relate to an autoregressive moving average specification of the attached GARCH.

Sometimes the conditional mean $E(Y_t|\mathcal{F}_{t-1})$ is allowed to be a constant, i.e.,

$$Y_t = \mu + \varepsilon_t$$

along with a GARCH(p, q) model for the conditional variance

$$Var\left(Y_t|\mathcal{F}_{t-1}\right) = E\left(\varepsilon_t^2|\mathcal{F}_{t-1}\right) = \sigma_t^2$$
$$= \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-1-j}^2,$$

where $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$. In this model, the tendency for large (small) residuals to be followed by other large (small) residuals but of unpredictable sign is fulfilled.

Many real time series data have the fatter (heavier) tails than are compatible with the normal distribution. One possible way to model these data with GARCH is to use t-distribution.

Bollerslev (1987) set the GARCH allowing for conditionally t-distributed errors.

The normally conditionally distributed errors of the ARCH, and even in the GARCH, make the model leptokurtic. t-distribution makes it fatter tailed. Some mixtures of normal distributions are also used. Fitting the distribution of the error in the GARCH is a widely discussed theme.

To determine the order of the model, one can use the autocorrelation and partial autocorrelation functions of the data, or, better stilt, sample autocorrelation and sample partial autocorrelation functions.

The requirement of heteroscedastic time series is evident not only in economy but also in some other disciplines, for instance, in chemistry (Tsay 1987) and so on.

Nowadays, plenty of heteroscedastic time series can be seen, e.g., one dimensional and multidimensional, which together are usually called GARCH-Zoo.

Cross References

►Heteroscedasticity
►Nonlinear Time Series Analysis
►Statistical Modeling of Financial Markets
►Time Series
►Time Series Regression

References and Further Reading

Bollerslev T (1986) Generalized autoregressive conditional heterosce dasticity. J Econometrics 31:307–327
Bollerslev T (1987) A conditionally Heteroskedstic time series model for speculative prices and rates of return. Rev Econ Stat 69(3):542–547
Engle RF (1980) Estimates of the variance of U.S. Inflation based on the ARCH model, University of California, San Diego Disussion Paper, pp 80–14
Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50(4):987–1007
Franses PH, van Dijk D (2000) Nonlinear time series models in empirical finance. Cambridge University Press, Cambridge
Li WK, Ling S, McAleer M (2002) Recent theoretical results for time series models with GARCH errors. J Econ Surv 16(3):245–269
Nelson DB (1991) Conditional heteroscedasticity in asset return: a new approach. Econometrica 59:347–370
Taylor SJ (1986) Modelling financial time series. Wiley, Chichester
Tsay RS (1987) Conditional heteroscedastic time series models. J Am Stat Assoc 82(398):590–604

Heteroscedasticity

Vesna Bucevska
Associate Professor, Faculty of Economics
University "Ss. Cyril and Methodious", Skopje, Macedonia

Introduction

One of the classical linear regression model assumptions (see ►Linear Regression Models) is that random errors u_i have a common variance σ^2. That does not mean that each error observation has the same size, but simply that each error observation has the same probability distribution with zero mean and constant variance, σ^2. When this assumption is violated, we are talking about heteroscedasticity. It can be defined as a systematic pattern in errors, which means that errors are drawn from different probability distributions with different variances, that is,

$$\text{var}(u_i) = \sigma_i^2 \quad i = 1, 2, \ldots, n. \tag{1}$$

The heteroscedasticity that results as a violation of the above-mentioned assumption of the classical linear regression model is known as a *pure* heteroscedasticity. It occurs when the regression model is correctly specified. Another reason for heteroscedasticity could be the model specification error, especially when a variable is omitted. That kind

of heteroscedasticity is known as *impure* heteroscedasticity. In that case, the corrective measure would be to find and include that variable in the model.

The existence of different variances of random errors or the problem of heteroscedasticity is most common for cross-section data that include data with different sizes. The bigger the difference between sizes of dependent variable observations, the higher the probability that the random errors will have different variances, and therefore will be heteroscedastic.

Heteroscedasticity can take many different forms. The most common one is when the error variance is related to an exogenous variable Z. That variable could be some of the explanatory variables in the model, but it can be also some variable outside the model. In the regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad (2)$$

the variance of the stochastic term u_i can be expressed as:

$$\text{var}(u_i) = \sigma_i^2 Z_i^2. \quad (3)$$

The variable Z is called the proportionality factor, since the value of the error variance is changing proportionally to the squared Z_i. The higher the value of Z for some observation, the higher the variance of error is. Usually Z is a measure of the size of each observation. Note that the heteroscedasticity is not only a characteristic which is strictly valid for cross-section data. For time series data that express trend in the movement of a dependent variable, it is logical to expect that there would be also a trend in the movement of the random error. The heteroscedasticity can also arise if there are big changes in the movement of the dependent variable. One of the reasons for heteroscedasticity could be a dramatic change in the quality of data collection.

Figure 1 illustrates the problem of heteroscedasticity. The probability density function $f(Y_3|X_3)$ at point X_3 shows that there is a high probability that Y_3 will be close to $E(Y_3)$. As we move towards point X_2, the probability density function is more spread $f(Y_2|X_2)$ that is, we are less sure where Y_2 could be found.

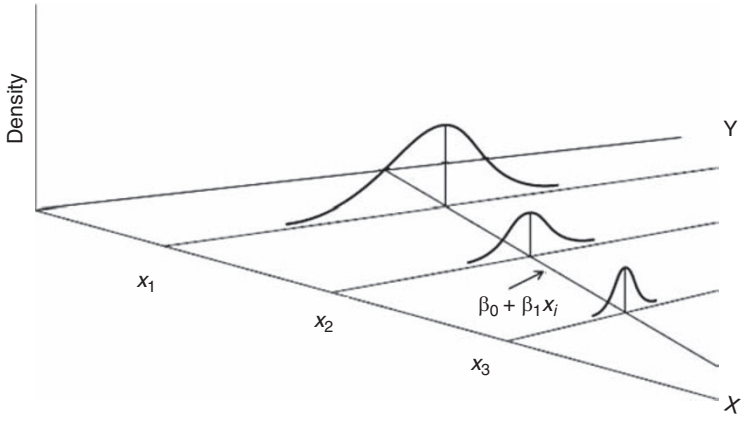
Consequences of Heteroscedasticity

- In the presence of pure heteroscedasticity, OLS estimators $\hat{\beta}$ remain unbiased, which means $E(\hat{\beta}) = \beta$. In other words, if we run regression many times using different data, then the average of all estimated $\hat{\beta}$ will give the real parameter value. However, in the case of impure heteroscedasticity, the consequences for the OLS estimators are more serious and OLS estimators are no longer unbiased.
- OLS estimators do not have the minimum variance anymore (they are not efficient).
- The standard errors of the OLS estimators computed in a usual way are incorrect and biased. This implies that confidence intervals and hypothesis tests that use these standard errors might be misleading.
- If the standard errors are biased, we cannot draw inferences based on *t* statistics or *F* statistics or *LM* statistics.

Detection of Heteroscedasticity

There are several tests that can be used to detect the presence of heteroscedasticity in the data. They can be divided into two groups:

1. Informal (graphical) test. This is a good starting point. The squared residuals are plotted against explanatory variables, or, in the case of the multiple regression model, against the dependent variable. If there is any



Heteroscedasticity. Fig. 1 Heteroscedastic errors

systematic pattern, it can be an indicator of possible heteroscedasticity.

2. Formal tests
   - *Park test*. It is designed for dealing with proportional heteroscedasticity. The Park test proceeds in the following way: the natural logarithm of the squared OLS residuals is regressed on the natural logarithm of the selected proportional factor (Z). With t-test we test the significance of the Z parameter and if it statistically significant, it is an evidence of heteroscedasticity.
   - *White test*. It is more general than the Park test. The test consists of regressing the squared residuals on all explanatory variables and their cross-products. This is an LM test, thus the test statistic is nR^2. The problem arises when the number of explanatory variables is large, thus leading to the problem of ►multicollinearity or even of loss of degrees of freedom, not to be able to estimate the model.
   - *Breusch–Pagan test*. It is very similar to White test, but it overcomes its shortcomings since the squared residuals are regressed on selected explanatory variables, those that cause the problem of heteroscedasticity.
   - *Goldfeld–Quandt test*. It orders the X observations in descending order and divides them into two subgroups, one with potential higher variance and the other one with potential smaller variance. Then the ratio of these two variances is calculated and compared to the critical value of F distribution.

The choice of the most appropriate test for heteroscedasticity is determined by how explicit we want to be about the form of heteroscedasticity.

Remedial Measures

As already stated, the problem with heteroscedasticity is that we cannot rely on the t-statistics, because the standard error estimators are biased. Various solutions are suggested for this problem:

1. The method of weighted least squares (WLS)
2. Obtaining heteroscedasticity-corrected standard errors
3. Redesigning the model

The WLS method can be used in the case when heteroscedasticity is caused by the proportionality factor, Z. It consists of the following steps:

1. Dividing each variable in the original regression model by the proportionality factor Z_i (this dividing will turn residuals into a white noise process u_i) and then, rerun the regression. The second regression model will not suffer from heteroscedasticity.

   For example, suppose we want to estimate the regression model (2) and the variance of the error term takes the form (3). Dividing the regression model by the proportionality factor Z_i, we obtain the following second regression model:

$$Y_i/Z_i = \beta_0/Z_i + \beta_1 X_{1i}/Z_i + \beta_2 X_{2i}/Z_i + u_i^* \qquad (4)$$

   The error term of the transformed regression model, u_i^*, has now a constant variance, and thus the regression model can be estimated by OLS.
2. The major problem, when using this method, consists of defining Z_i and choosing the functional form of Z_i, since different functional forms require different transformations. Wrong choice of the proportionality factor (or weight) can produce biased estimators of the standard errors. If Z_i is not any of the explanatory variables, then we must include a constant term in the above model, otherwise a constant is already included.

   We should be very cautious when interpreting the estimated coefficients of the model (4), since it can be noted that the coefficient β_1, which is a slope coefficient in model (2) now becomes an intercept in model (4). The opposite happens with coefficient β_0.

   Another problem related to WLS is the functional form of the relationship between proportionality factor and error variance. Until now, a direct proportionality of error variance to explanatory variable has been assumed. But, it is possible that error variance can be expressed as a linear relation of explanatory variables. Let's assume that this relation is:

$$\sigma_i^2 = \alpha_0 + \alpha_1 X_{1i}^2 + u_i^*. \qquad (5)$$

   Then, applying OLS we estimate the residuals from Eq. 5 as:

$$\hat{u}_i^{*2} = \alpha_0 + \alpha_1 X_{1i}^2. \qquad (6)$$

   Now, the weight is $\frac{1}{\sqrt{\hat{u}_i^{*2}}}$, and we divide all the variables in the original model with this weight. Finally, we test if the new error term is homoscedastic or not.

Obtaining heteroscedasticity-corrected standard errors is the most popular remedy, which improves the estimation of the standard errors using OLS coefficient estimates. It is very convenient when the form of heteroscedasticity is unknown.

The standard errors, as more accurate, are then used for recalculating the t-statistics using the same means taht remain unchanged. Typically, the corrected standard errors will be larger, thus leading to lower t-statistics. The

approach of corrected standard errors is the most suitable to large samples and is a part of some good statistical software packages. However, it does not work very well with small samples.

Acknowledgments

I would like to thank Professor John Chipman, Professor Badi Baltagi, and the editor for reading a draft of this article and providing me with comments and suggestions which resulted in many improvements.

Cross References

►Astrostatistics
►Heteroscedastic Time Series
►Linear Regression Models
►Tests for Homogeneity of Variance
►Time Series Regression

References and Further Reading

Baltagi BH (2003) A companion to theoretical econometrics. Blackwell, Oxford
Gujarati D (2003) Basic econometrics, 4th edn. McGraw Hill, New York
Kmenta J (1997) Elements of econometrics, 2nd edn. University of Michigan Press, Michigan
Wooldridge J (2009) Introductory econometrics, 4th edn. South-Western, Cincinnati

Hierarchical Clustering

Fionn Murtagh
President of the British Classification Society
Director
Wilton Place, Dublin, Ireland
Professor
University of London, London, UK

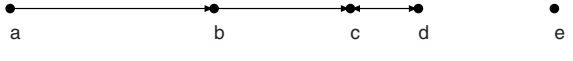
Hierarchical clustering algorithms can be characterized as *greedy* (Horowitz and Sahni 1979). A sequence of irreversible algorithm steps is used to construct the desired data structure. Assume that a pair of clusters, including possibly singletons, is merged or agglomerated at each step of the algorithm. Then the following are equivalent views of the same output structure constructed on n objects: a set of $n - 1$ partitions, starting with the fine partition consisting of n classes and ending with the trivial partition consisting of just one class, the entire object set; a binary tree (one or two child nodes at each non-terminal node) commonly referred to as a dendrogram; a partially ordered set (poset) which is a subset of the power set of the n objects; and an ultrametric topology on the n objects. For background, the reader is referred to Benzécri et al. (1979), Lerman (1981), Murtagh and Heck (1987), Jain and Dubes (1988), Arabie et al. (1996), Mirkin (1996), Gordon (1999), Jain et al. (1999), and Xu and Wunsch (2005).

One could say with justice that Sibson (1973), Rohlf (1982) and Defays (1977) are part of the prehistory of clustering. Their $O(n^2)$ implementations of the single link method and of a (non-unique) complete link method have been widely cited.

In the early 1980s a range of significant improvements were made to the Lance-Williams, or related, dissimilarity update schema (de Rham 1980; Juan 1982), which had been in wide use since the mid-1960s. Murtagh (1983, 1985) presents a survey of these algorithmic improvements. The algorithms, which have the potential for *exactly* replicating results found in the classical but more computationally expensive way, are based on the construction of *nearest neighbor chains* and *reciprocal* or mutual NNs (NN-chains and RNNs).

A NN-chain consists of an arbitrary point (*a* in Fig. 1); followed by its NN (*b* in Fig. 1); followed by the NN from among the remaining points (*c*, *d*, and *e* in Fig. 1) of this second point; and so on until we necessarily have some pair of points which can be termed reciprocal or mutual NNs. (Such a pair of RNNs may be the first two points in the chain; and we have assumed that no two dissimilarities are equal.)

In constructing a NN-chain, irrespective of the starting point, we may agglomerate a pair of RNNs as soon as they are found. What guarantees that we can arrive at the same hierarchy as if we used traditional "stored dissimilarities" or "stored data" algorithms (Anderberg 1973)? Essentially this is the same condition as that under which no inversions or reversals are produced by the clustering method. This would be where s is agglomerated at a lower criterion value (i.e., dissimilarity) than was the case at the previous agglomeration between q and r. Our ambient space has thus contracted because of the agglomeration. This is due to the algorithm used – in particular the agglomeration criterion – and it is something we would normally wish to avoid.



Hierarchical Clustering. Fig. 1 Five points, showing NNs and RNNs

This is formulated as:

Inversion impossible if:

$d(i,j) < d(i,k)$ or $d(j,k) \implies d(i,j) < d(i \cup j,k)$

This is Bruynooghe's *reducibility property* (Bruynooghe 1977; see also Murtagh, 1985, 1992). Using the Lance-Williams dissimilarity update formula, it can be shown that the minimum variance method does not give rise to inversions; neither do the (single, complete, average) linkage methods; but the median and centroid methods cannot be guaranteed not to have inversions.

To return to Fig. 1, if we are dealing with a clustering criterion which precludes inversions, then c and d can justifiably be agglomerated, since no other point (for example, b or e) could have been agglomerated to either of these.

The processing required, following an agglomeration, is to update the NNs of points such as b in Fig. 1 (and on account of such points, this algorithm was dubbed *algorithme des célibataires,* the bachelors' algorithm, in de Rham 1980). The following is a summary of the algorithm:

NN-Chain Algorithm

Step 1 Select a point (i.e., an object in the input data set) arbitrarily.

Step 2 Grow the NN-chain from this point until a pair of RNNs are obtained.

Step 3 Agglomerate these points (replacing with a cluster point, or updating the dissimilarity matrix).

Step 4 From the point which preceded the RNNs (or from any other arbitrary point if the first two points chosen in Steps 1 and 2 constituted a pair of RNNs), return to Step 2 until only one point remains.

In Murtagh (1983, 1985) and Day and Edelsbrunner (1984), one finds discussions of $O(n^2)$ time and $O(n)$ space implementations of Ward's minimum variance (or error sum of squares) method and of the centroid and median methods. The latter two methods are termed the UPGMC and WPGMC criteria (respectively, unweighted and weighted pair-group method using centroids) by Sneath and Sokal (1973). Now, a problem with the cluster criteria used by these latter two methods is that the reducibility property is not satisfied by them. This means that the hierarchy constructed may not be unique as a result of inversions or reversals (non-monotonic variation) in the clustering criterion value determined in the sequence of agglomerations.

Murtagh (1984) describes $O(n^2)$ time and $O(n^2)$ space implementations for the single link method, the complete link method and for the weighted and unweighted group average methods (WPGMA and UPGMA). This approach is quite general vis à vis the dissimilarity used and can also be used for hierarchical clustering methods other than those mentioned.

Day and Edelsbrunner (1984) prove the exact $O(n^2)$ time complexity of the centroid and median methods using an argument related to the combinatorial problem of optimally packing hyperspheres into an m-dimensional volume. They also address the question of metrics: results are valid in a wide class of distances including those associated with the Minkowski metrics.

The construction and maintenance of the nearest neighbor chain as well as the carrying out of agglomerations whenever reciprocal nearest neighbors meet, both offer possibilities for parallelization, and implementation in a distributed fashion. Work in chemoinformatics and information retrieval can be found in Willett (1989), Gillet et al. (1998) and Griffiths et al. (1984). Ward's minimum variance criterion is favored.

For in depth discussion of data encoding and normalization as a preliminary stage of hierarchical clustering, see Murtagh (2005). Finally, as an entry point into the ultrametric view of clustering, and how hierarchical clustering can support constant time, or $O(1)$, proximity search in spaces of arbitrarily high ambient dimensionality, thereby setting aside Bellman's famous curse of dimensionality, see Murtagh (2004).

About the Author

Fionn Murtagh is a member of the Royal Irish Academy, and also Fellow of the International Association for Pattern Recognition, and a Fellow of the British Computer Society. He is President of the British Classification Society (2007 to date) and past President of the Classification Society of North America (2008–2009). He is a Fellow of the Royal Statistical Society and an elected member of the International Statistical Institute. He was Head of the Department of Computer Science, Royal Holloway, University of London (2004–2007). Apart from the University of London, he has held professorial chairs in computing in the past in Queen's University Belfast and the University of Ulster. His longest-serving appointment has been in the space sector, serving with the European Space Agency. He is currently directing Science Foundation Ireland's research funding programs in computing, engineering, telecommunications, mathematics, materials science, renewable energies and other fields. He takes particular pride in his doctoral genealogy, tracing back from his advisor, JP Benzécri, Henri Cartan (of Bourbaki) through Borel, Lebesgue, Poisson, Laplace, Lagrange, Euler, the Bernoullis, Leibniz, and Christiaan

Huygens (the last-mentioned playing a central role in the data analysis methodology of Benzécri.)

Cross References

►Cluster Analysis: An Introduction
►Data Analysis
►Data Mining Time Series Data
►Fuzzy Logic in Statistical Data Analysis
►Multicriteria Clustering
►Multivariate Data Analysis: An Overview

References and Further Reading

Anderberg MR (1973) Cluster analysis for applications. Academic, New York
Arabie P, Hubert LJ, De Soete G (eds) (1996) Clustering and classification. World Scientific, Singapore
Benzécri JP et coll (1979) L'Analyse des Données. I. La Taxinomie, 3rd edn. Dunod, Paris
Bruynooghe M (1977) Méthodes nouvelles en classification automatique des données taxinomiques nombreuses. Statistique et Analyse des Données 3:24–42
Day WHE, Edelsbrunner H (1984) Efficient algorithms for agglomerative hierarchical clustering methods. J Classif 1:7–24
de Rham C (1980) La classification hiérarchique ascendante selon la méthode des voisins réciproques. Les Cahiers de l'Analyse des Données V:135–144
Defays D (1977) An efficient algorithm for a complete link method. Comput J 20:364–366
Gillet VJ, Wild DJ, Willett P, Bradshaw J (1998) Similarity and dissimilarity methods for processing chemical structure databases. Comput J 41:547–558
Gordon AD (1999) Classification, 2nd edn. Champman & Hall, Boca Raton
Griffiths A, Robinson LA, Willett P (1984) Hierarchic agglomerative clustering methods for automatic document classification. J Doc 40:175–205
Horowitz E, Sahni S (1979) Fundamentals of computer algorithms (Chapter 4. The greedy method). Pitman, London
Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs
Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31:264–323
Juan J (1982) Programme de classification hiérarchique par l'algorithme de la recherche en chaîne des voisins réciproques. Les Cahiers de l'Analyse des Données VII:219–225
Lerman IC (1981) Classification et Analyse Ordinale des Données. Dunod, Paris
Mirkin B (1996) Mathematical classification and clustering. Kluwer, Dordrecht
Murtagh F (1983) A survey of recent advances in hierarchical clustering algorithms. Comput J 26:354–359
Murtagh F (1984) Complexities of hierarchic clustering algorithms: state of the art, Computational Statistics Quarterly 1:101–113
Murtagh F (1985) Multidimensional clustering algorithms. Physica-Verlag, Würzburg
Murtagh F (1992) Comments on "Parallel algorithms for hierarchical clustering and cluster validity". IEEE Trans Pattern Anal Mach Intell 14:1056–1057
Murtagh F (2004) On ultrametricity, data coding, and computation. J Classif 21:167–184
Murtagh F (2005) Correspondence analysis and data coding with Java and R. Chapman & Hall, Boca Raton
Murtagh F, Heck A (1987) Multivariate data analysis. Kluwer, Dordrecht
Rohlf FJ (1982) Single link clustering algorithms. In: Krishnaiah PR, Kanal LN (eds) Handbook of statistics, vol 2. North-Holland, Amsterdam, pp 267–284
Sibson R (1973) SLINK: an optimally efficient algorithm for the single link cluster method. Comput J 16:30–34
Sneath PHA, Sokal RR (1973) Numerical taxonomy. W.H. Freeman, San Francisco
Willett P (1989) Efficiency of hierarchic agglomerative clustering using the ICL distributed array processor. J Doc 45:1–45
Xu R, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16:645–678

Hodges-Lehmann Estimators

Scott L. Hershberger
Global Director of Survey Design
Harris Intractive, New York, NY, USA

The *Hodges-Lehmann estimator* provides, in the one-sample case, an estimate of the center of a distribution, and in the two-sample case, an estimate of the difference in the centers of two distributions.

The one-sample estimator is defined as the *median* of the set of $n(n+1)/2$ *Walsh averages*. Each Walsh average is the arithmetic average of two observations. For example, consider the set of 5 observations $(1, 3, 8, 9, 15)$. Table 1 shows the computation of the $5(5+1/2) = 15$ Walsh averages.

The median of the Walsh averages is the one-sample Hodges-Lehmann estimator of the center of the distribution. In this example, the median of the 15 Walsh averages (the Hodges-Lehmann estimator) is 8. Note that in this case, the Hodges-Lehmann estimator is equal to the simple median of the original five observations, which is also 8. Of course, the Hodges-Lehmann estimator does not have to necessarily equal the sample median. While both the median and Hodges-Lehmann estimator are both preferable to the sample mean for nonsymmetric distributions, the Hodges-Lehmann estimator has larger asymptotic relative efficiency with respect to the mean than the median; i.e., 0.96 versus 0.64.

It possible to construct a $(1-\alpha)100\%$ *confidence interval* for the Hodges-Lehmann estimator. For an approximate $(1-\alpha)100\%$ confidence interval first find the value of $W_{\alpha/2}$ as the $100\alpha/2$ percentile of the distribution of the *Wilcoxon test statistic*. Then if $W_{\alpha/2} = K^*$, then the K^*th smallest to the K^*th largest of the Walsh averages

Hodges-Lehmann Estimators. Table 1 Computation of walsh averages

	1	3	8	9	15
1	(1 + 1)/2 = 1	(1 + 3)/2 = 2	(1 + 8)/2 = 4.5	(1 + 9)/2 = 5	(1 + 15)/2 = 8
3		(3 + 3)/2 = 3	(3 + 8)/2 = 5.5	(3 + 9)/2 = 6	(3 + 15)/2 = 9
8			(8 + 8)/2 = 8	(8 + 9)/2 = 8.5	(8 + 15)/2 = 11.5
9				(9 + 9)/2 = 9	(9 + 15)/2 = 12
15					(15 + 15)/2 = 15

Hodges-Lehmann Estimators. Table 2 Computation of pairwise differences

x/y	2	7	10	11	12
1	(1 – 2) = –1	(1 – 7) = –6	(1 – 10) = –9	(1 – 11) = –10	(1 – 12) = –11
3	(3 – 2) = 1	(3 – 7) = –4	(3 – 10) = –7	(3 – 11) = –8	(3 – 12) = –9
8	(8 – 2) = 6	(8 – 7) = 1	(8 – 10) = –2	(8 – 11) = –3	(8 – 12) = –4
9	(9 – 2) = 7	(9 – 7) = 2	(9 – 10) = –1	(9 – 11) = –2	(9 – 12) = –3
15	(15 – 2) = 13	(15 – 7) = 8	(15 – 10) = 5	(15 – 11) = 4	(15 – 12) = 3

determines the $(1-\alpha)$ 100% confidence interval. Values of $W_{\alpha/2}$ for finding approximate 90%, 95%, and 99% confidence intervals are found from readily available tables. For example, for the 15 Walsh averages given above, the 95% confidence interval for the Hodges-Lehmann estimator of 8 has a lower bound of 1 and upper bound of 15. For sample sizes of about 50 or more, K^* can be calculated approximately as

$$K^* = \frac{n(n+1)}{4} - \left(z_{1-\alpha/2} \times \sqrt{\frac{n(n+1)(2n+1)}{24}} \right),$$

rounded up to the next integer value, where $z_{1-\alpha/2}$ is the appropriate value from the standard normal distribution for the $(1-\alpha)$ 100% percentile.

The two-sample Hodges-Lehmann estimator is defined as the *median* of $n_1 \times n_2$ pairwise differences between samples X and Y, $y_j - x_k, j = 1, \ldots, n_1, k = 1, \ldots, n_2$. For example, consider the two samples, $x = (1, 3, 8, 9, 15)$ and $y = (2, 7, 10, 11, 12)$. Table 2 shows the computation of the 5×5 pairwise differences.

The median of the 25 pairwise differences is –2; thus the Hodges-Lehmann estimator of the center location difference between the two samples is –2. By comparison, the mean difference between the two samples is 36/5 – 42/5 = –1.2.

The confidence interval for the Hodges-Lehmann estimator of the center location difference is also based on the $n \times m$ pairwise differences. For an approximate $(1-\alpha)$ 100% confidence interval first calculate

$$K = W_{\alpha/2} - \frac{n_1(n_1+1)}{2},$$

where $W_{\alpha/2}$ is the $100\alpha/2$ percentile of the distribution of the *Mann-Whitney test statistic*. The Kth smallest to the Kth largest of the $n_1 \times n_2$ pairwise differences then determine the $(1-\alpha)$ 100% confidence interval. Values of $W_{\alpha/2}$ for finding approximate 90%, 95%, and 99% confidence intervals are found from readily available tables. For example, for the 25 pairwise differences given above, the 95% confidence interval for the Hodges-Lehmann estimator of –2 has a lower bound of –9 and upper bound of 7.

About the Author

For biography *see* the entry ►Structural Equation Models.

Cross References

►Asymptotic Relative Efficiency in Estimation

References and Further Reading

Lehmann EL (1999) Elements of large sample theory. Springer, New York

Lehmann EL (1998) Nonparametrics: statistical methods based on ranks. Prentice Hall, Upper Saddle River

Sprent P (1998) Data driven statistical methods. Chapman & Hall, London

Walsh JE (1949) Some significance tests for the median which are valid under very general conditions. Ann Math Stat 20:64–81

Horvitz–Thompson Estimator

Tapabrata Maiti
Professor
Michigan State University, East Lansing, MI, USA

Introduction and Definition

The Horvitz–Thompson (H–T) estimator is attributed to D.G. Horvitz and D.J. Thompson to estimate a finite population total when a sample is selected with unequal probabilities without replacement. Let $U = (1, \cdots, i, \cdots, N)$ denote a finite universe of N elements and s be a sample of size n. Let $\pi_i(>0)$ be the inclusion probability, the probability that the i-th unit is in the sample and $\pi_{ij}(>0)$ is the probability that both the units i and j are in the sample. If $p(s)$ is the probability of selecting a sample s, then $\pi_i = \sum_{s \ni i} p(s)$ and $\pi_{ij} = \sum_{s \ni i,j} p(s)$. The inclusion probabilities satisfies the following relations:

$$\sum_{i=1}^{N} \pi_i = n; \sum_{j \neq i}^{N} \pi_{ij} = (n-1)\pi_i; \sum_{i=1}^{N} \sum_{j>i} \pi_{ij} = \frac{1}{2} n(n-1)$$

Let y_i be the response attached to i-th unit $(i = 1, \cdots, N)$ and the target is to estimate the finite population total $T = \sum_{i=1}^{N} y_i$. The Horvitz–Thompson (1952) estimator of the population total is

$$t = \sum_{i \in s} \frac{y_i}{\pi_i}$$

The higher the selection probability π_i of unit i, the less weight $\left(\frac{1}{\pi_i}\right)$ the corresponding value y_i is given. The Hansen and Hurwitz (1943) estimator was constructed similarly for unequal probability sampling with replacement. The H–H estimator uses probability of selection of a unit instead the inclusion probabilities. If $p_1, \cdots, p_N$ are selection probabilities with $\sum_{i=1}^{N} p_i = 1$, and n independent draws are made to select a sample with replacement then the inclusion probability for i-th unit is

$$\pi_i = 1 - (1 - p_i)^n$$

For $n = 1, \pi_i = p_i$, but they are not same in general.

Properties of the H–T Estimator

The H–T estimator t is an unbiased estimator of the population total T, with variance

$$V_1(t) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j$$

Following the relation $\sum_{j \neq i} (\pi_{ij} - \pi_i \pi_j) = -\pi_i(1-\pi_i)$ for a fixed sample size design, the variance can be expressed as

$$V_2(t) = \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

This form of variance is known as Yates–Grundy form (Sen 1953; Yates and Grundy 1953). The variance estimators are

$$\hat{V}_1(t) = \sum_{i=1}^{n} \frac{1-\pi_i}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j$$

for the variance V_1 and

$$\hat{V}_2(t) = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

for the variance V_2. Both of these variance estimators are unbiased.

Example Consider a ▶simple random sample without replacement of size n. The inclusion probabilities are $\pi_i = \frac{n}{N}$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$. The H–T estimator for the population total is $N\bar{y}$ where $\bar{y} = \frac{1}{n} \sum_s y_i$. The variance is $N^2 \frac{1-f}{n} S^2$ where $S^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \frac{T}{N}\right)^2$ and $f = \frac{n}{N}$, known as the sampling fraction. The variance estimator is $N^2 \frac{1-f}{n} s^2$ where $s^2 = \frac{1}{n-1} \sum_s (y_i - \bar{y})^2$. Note that both the variance estimators take same form in this design.

Due to wide variation of $(\pi_i \pi_j - \pi_{ij})$, even sometimes being negative, both the variance estimators $\hat{V}_1$ and $\hat{V}_2$ can be negative for some sampling designs. Rao and Singh (1973) compared the coefficient of variation for these two variance estimators using Brewer's sample selection method (Brewer 1963) and found that the Yates–Grundy form of variance estimator is more stable than the other. Consequently, applying H–T estimator one should be carefully selecting a sampling design where the inclusion probabilities are chosen cautiously in relation to response variable; otherwise may end up with silly outcome as the famous "elephant" example of Basu (1971).

Hajek (1964) provided the necessary and sufficient condition for the ▶asymptotic normality of the H–T estimator under rejective sampling. Berger (1998) extended the asymptotic normality result for general sampling design and also provided the rate of convergence. Berger (1998) used the asymptotic framework as follows: Let $\{n_1, \cdots, n_k, \cdots\}$ and $\{N_1, \cdots, N_k, \cdots\}$ be sequences of sample size and population size respectively, where both n_k and N_k increase as $k \to \infty$.

Särnal et al. (1992) provided a very generalized treatment to the H–T estimator and called as π estimator. This general form of estimator also applies to unequal probability sampling design with replacement. The generalized concept has further been illustrated by Overton

and Stehman (1995). The H–T estimator has been adapted in model based inference by Kott (1988) and Särnal et al. (1992). Kott (1988) particularly derived the finite population correction for the H–T estimator under usual super population model. On the other hand, Särnal et al. (1992) adopted H–T estimator for predicting the population total under the super population model and then used the design perspective for statistical inference. The estimator is popularly known as GREG (generalized regression) predictor and the approach is known as model-assisted, as opposed to model-based approach. For a super population model ζ with p covariate $\boldsymbol{x} = (x_1, \cdots, x_p)^T$,

$$E_\zeta(y_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$
$$V_\zeta(y_i) = \sigma_i^2$$

the GREG predictor for population total T is

$$t_{GREG} = \sum_s g_{is} \frac{y_i}{\pi_i}$$

where

$$g_{is} = 1 + \left(\sum_{i=1}^N \boldsymbol{x}_i - \sum_s \frac{\boldsymbol{x}_i}{\pi_i}\right)^T \left(\sum_s \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\pi_i \sigma_i^2}\right)^{-1} \frac{\boldsymbol{x}_i}{\sigma_i^2}$$

The approximate variance of GREG predictor is

$$V(t_{GREG}) \doteq -\sum_{i=1}^N \sum_{j=1}^N (\pi_i \pi_j - \pi_{ij}) \frac{E_i}{\pi_i} \frac{E_j}{\pi_j}$$

where $E_i = y_i - \boldsymbol{x}_i^T \boldsymbol{B}, \boldsymbol{B} = \left(\sum_{i=1}^N \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2}\right)^{-1} \sum_{i=1}^N \frac{\boldsymbol{x}_i y_i}{\sigma_i^2}$.
An approximated variance estimator is

$$\hat{V}(t_{GREG}) \doteq -\sum_s \sum_s \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \frac{g_{is} e_i}{\pi_i} \frac{g_{js} e_j}{\pi_j}$$

where $e_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{B}}, \hat{\boldsymbol{B}} = \left(\sum_s \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{\sigma_i^2 \pi_i}\right)^{-1} \sum_s \frac{\boldsymbol{x}_i y_i}{\sigma_i^2 \pi_i}$.

The formulae get simplified with variance structure $\sigma_i^2 = \boldsymbol{\lambda}^T \boldsymbol{x}_i$ for known $\boldsymbol{\lambda}$.

Acknowledgments

The work was partially supported by the National Science Foundation Grant SES 0904055.

About the Author

Tapabrata Maiti is currently a Professor and the graduate director of Statistics and Probability, Michigan State University, USA. He is an elected fellow of the American Statistical Association. He is also Associate editor of *Journal of the American Statistical Association*, *Journal of Agricultural, Biological and Environmental Statistics*, *Sankhyā*, the *Indian Journal of Statistics* and the *Test* (the Spanish Journal of Statistics). He has (co-)authored about 50 papers. Professor Maiti was awarded the Bose-Nandi award for the best applied paper in *Calcutta Statistical Association Buletin* (2005), and Raja Rao Memorial Prize for the best published research work in Survey Sampling done in India for the year 1995–1996.

Cross References

►Adaptive Sampling
►Balanced Sampling
►Pareto Sampling
►Sampling Algorithms
►Sampling From Finite Populations

References and Further Reading

Basu D (1971) An essay on the logical foundations of survey sampling, Part I. In: Godambe VP and Sprott DA (eds) Foundation of statistical inference. Holt, Rinehart and Winston, Toronto, pp 203–242
Berger YG (1998) Rate of convergence to normal distribution for the Horvitz-Thompson estimator. J Stat Plann Infer 67:209–226
Brewer KRW (1963) A model of systematic sampling with unequal probabilities. Aust J Stat 5:5–13
Hansen MH, Hurwitz WN (1943) On the theory of sampling from finite population. Ann Math Stat 14:517–529
Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685
Hajek J (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. Ann Math Statist 35: 1491–1523
Kott PS (1988) Model-based finite population correction for the Horvitz–Thompson estimator. Biometrika 75:797–799
Overton WS, Stehman SV (1995) The Horvitz–Thompson theorem as unifying perspective for probability sampling: with examples from natural resource sampling. Am Stat 49:261–268
Rao JNK, Singh MP (1973) On the choice of estimator in survey sampling. Aust J Stat 15:95–104
Särnal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, New York
Sen AR (1953) On the estimate of variance in sampling with varying probabilities. J Indian Soc Agr Stat 5:119–127
Yates F, Grundy PM (1953) Selection without replacement from within strata with probability proportional to size. J R Stat Soc 15:253–261

Hotelling's T^2 Statistic

Robert L. Mason [1], John C. Young [2]
[1]Southwest Research Institute, San Antonio, TX, USA
[2]Lake Charles, LA, USA

The univariate t statistic is well-known to most data analysts. If a random sample of n observations is taken from a normal distribution with mean μ and variance σ^2, the form

of the statistic is given by

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \tag{1}$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation. This statistic has a Student t-distribution with $(n-1)$ degrees of freedom. Squaring the statistic, we obtain

$$t^2 = n(\bar{x} - \mu)'(s^2)^{-1}(\bar{x} - \mu).$$

This value can be defined as the squared Euclidean distance between $\bar{x}$ and μ.

A multivariate analogue to the t statistic can easily be constructed. Suppose a random sample of n observation vectors, given by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})'$, is taken from a p-variate multivariate normal distribution (see ▶Multivariate Normal Distributions) with mean vector $\mathbf{u}$ and covariance matrix, Σ. Then the multivariate version of the t statistic in (1) is given by

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}), \tag{2}$$

where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ is the sample mean and $\mathbf{S} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'/(n-1)$ is the sample covariance matrix.

The statistic in (2) and its sampling distribution was first developed by Harold Hotelling (1931), and it is commonly referred to as Hotelling's T^2 statistic. It was first introduced for usage in testing the null hypothesis that the mean vector, $\boldsymbol{\mu}$, of a multivariate normal distribution is equal to a specific vector $\boldsymbol{\mu}_0$, i.e., $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, against the alternative hypothesis that it is not equal to this vector, i.e., $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. Thus, it can be used as the statistic for a multivariate one-sample test of hypothesis.

The null distribution of the T^2 statistic in (2) is given by a central F distribution and the distribution under the alternative hypothesis is a non-central F distribution. For the above one-sample hypothesis the null distribution is given by

$$T^2 \sim \frac{(n-1)p}{n-p}F_{(p,n-p)}, \tag{3}$$

where $F_{(p,n-p)}$ is an F-distribution with p and $(n-p)$ degrees of freedom. Note that the limiting form of this F-distribution is a ▶chi-square distribution with p degrees of freedom. This is the form of the asymptotic distribution of the T^2 statistic.

The T^2 statistic has many interesting properties. First, it is invariant under all nonsingular linear transformations of the data given by

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{b},$$

where $\mathbf{x}$ is a sampled multivariate normal observation, $\mathbf{b}$ is a known vector of constants, and $\mathbf{A}$ is a non-singular matrix of known values. This invariance property means that the T^2 results are independent of scale and origin changes to the data so that the T^2 statistic based on $\mathbf{z}$ is identical to the T^2 statistic based on $\mathbf{x}$. From this one can conclude that testing the hypothesis $H_0 : \boldsymbol{\mu}_\mathbf{z} = \boldsymbol{\mu}_{\mathbf{z}_0}$ is equivalent to testing the hypothesis $H_0 : \boldsymbol{\mu}_\mathbf{x} = \boldsymbol{\mu}_{\mathbf{x}_0}$, where $\boldsymbol{\mu}_\mathbf{z}$ is the population mean of the $\mathbf{z}$ data and $\boldsymbol{\mu}_\mathbf{x}$ is the population mean of the $\mathbf{x}$ data. Second, the hypothesis test based on the T^2 statistic is the uniformly most powerful invariant test. This means that in the class of all invariant tests there is no test that has greater ability to detect a true null hypothesis than the T^2 statistic. Third, the T^2 statistic is equivalent to the likelihood ratio statistic for this hypothesis. Thus, the optimal and distributional properties of the likelihood ratio test hold for the T^2 statistic.

When two different multivariate normal populations, with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, have equal covariance matrices, the T^2 statistic can be used to test a two-sample null hypothesis of the form $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. With independent samples of sizes n_1 and n_2, respectively, from each population, the T^2 statistic for the test procedure is based on the difference of the sample mean vectors, $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, and the pooled estimate of the common covariance matrix. The form of the statistic is given by

$$T^2 = \frac{n_1 n_2}{(n_1 + n_2)}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{4}$$

where n_1 and n_2 are the sizes of the two independent samples, $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the corresponding sample means and $\mathbf{S}_p = [(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2]/(n_1 + n_2 - 2)$ is the pooled sample covariance matrix corresponding to the sample covariance matrices, $\mathbf{S}_1$ and $\mathbf{S}_2$. Similarly, the null distribution for this T^2 statistic is given by

$$T^2 \sim \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p}F_{(p,n_1+n_2-p-1)}. \tag{5}$$

The quadratic form of (4), given by

$$\mathbf{D}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_p^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{6}$$

is known as the squared sample Mahalanobis distance between the vectors $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$. This statistic was developed by P.C. Mahalanobis (1930) at about the same time that Hotelling developed the T^2 statistic. Its square root is used as an estimate of the distance between the two means of the populations relative to the common covariance matrix.

The T^2 test statistic in (4) is also equivalent to the likelihood ratio test statistic for a two-population mean test. A form of the statistic was used by Fisher (1936) as a means for classifying a new observation $\mathbf{x}$ into one of two groups and is known as Fisher's linear discriminant function (e.g., see Johnson and Wichern 2008).

In addition to its use in multivariate analysis, the T^2 statistic also has been applied extensively in the area of multivariate statistical process control (MVSPC). This first

occurred in 1947 when Harold Hotelling used the statistic to solve problems concerning bombsights (Hotelling 1947). Since the advent of inexpensive computing power became available in the 1980's, the T^2 statistic has become one of the most popular charting statistics for use in monitoring the many variables of a multivariate process (see Mason and Young 2002).

In this short entry, we have discussed only a few of the numerous applications of Hotelling's T^2 statistic in multivariate analysis. Many more are available and information on them can be obtained in most applied multivariate statistics textbooks.

About the Author

For biographies of both authors *see* the entry ▶Multivariate Statistical Process Control.

Cross References

▶General Linear Models
▶Multivariate Statistical Process Control
▶Multivariate Technique: Robustness
▶Student's *t*-Tests

References and Further Reading

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 8:376–378
Hotelling H (1931) The generalization of Student's ratio. Ann Math Stat 2:360–378
Hotelling H (1947) Multivariate quality control-illustrated by the air testing of sample bombsights. Tech Stat Anal (Eisenhart C, Hastay MW, Wallis WA (eds)) McGraw-Hill, New York, pp 111–184
Johnson RA, DW Wichern J (2008) Applied multivariate statistical analysis, 6th edn. Prentice Hall, New Jersey
Mahalanobis PC (1930) On tests and measures of group divergence. I.F. Proc. Asiat. Soc. Bengal 26:541
Mason RL, Young JC (2002) Multivariate statistical process control with industrial applications. ASA-SIAM, Philadelphia, PA

Hyperbolic Secant Distributions and Generalizations

Matthias Fischer
University of Erlangen-Nürnberg, Erlangen, Germany

The hyperbolic secant distribution (HSD) has its origin in Fisher (1921), Dodd (1925), Roa (1924) and Perks (1932). Some of its properties are developed by Talacko (1956, 1958, 1958). Given two independent standard normal variables Y_1 and Y_2, the variable $X \equiv \ln|Y_1/Y_2|$ is said to follow a *hyperbolic secant* distribution. Analogue to the ▶logistic distribution, probability density, cumulative distribution function and quantile function of X admit a closed form, namely

$$f(x) = \frac{2}{\pi(e^x + e^{-x})}, \quad F(x) = \frac{2\arctan(e^x)}{\pi} \text{ and}$$
$$F^{-1}(p) = \ln\left(\tan\left(\frac{\pi p}{2}\right)\right), \quad x \in \mathbb{R}.$$

Obviously, the density is symmetric around zero and has mode at zero with $f(0) = 1/\pi$. Since the moment-generating function of a HSD is given by $\mathcal{M}(t) = (\cos(\pi t/2))^{-1}$ for $|t| < \pi/2$, all moments exist. In particular, $\mathbb{E}(X^i) = 0$ for odd i, $Var(X) = \pi^2/4$ and $\mathbb{E}(X^4) = 5\pi^4/16$. Consequently, the kurtosis coefficient (i.e., the fourth standardized moment) of a HSD calculates as $m_4 = 5$, indicating that the HSD has heavier tails and higher peakedness than the normal distribution ($m_4 = 3$) and than the logistic distribution ($m_4 = 4.2$).

Basically, two major generalized hyperbolic distributions emerged.

The first one has its roots in the work of Baten (1934) who derived the probability density function of $X_1 + \ldots + X_n$ where each X_i follows a HSD for finite $n \in \mathbb{N}$ as well as for $n = \infty$. More generally, Harkness & Harkness (1968) discuss distribution families with characteristic function $\text{sech}(t)^\rho$, $\rho > 0$ which can be identified as ρ-th convolution of a hyperbolic secant variable. This symmetric distribution family is commonly termed as *generalized hyperbolic secant* (GHS) distribution with kurtosis parameter ρ. A corresponding skew GHS distribution is known in the statistical literature as NEF-GHS or Laha–Lukacs distribution (e.g., Morris 1982) or as Meixner distribution in the mathematical and financial literature (e.g., Meixner 1934 or Schoutens 2003).

The second generalization dates back to Perks (1932) who discussed probability densities of the form

$$f(x) = \frac{a_0 + a_1 e^{-x} + a_2 e^{-2x} + \ldots + a_m e^{-mx}}{b_0 + b_1 e^{-x} + b_2 e^{-2x} + \ldots + b_n e^{-nx}}$$

with parameters a_i, b_j such that f is a proper probability density. Setting $m = 1$, $a_0 = 0$, $a_1 = 1$ and $n = 2$, $b_0 = 1$, $b_1 = 0$, this equation reduces to the density of a HSD. More generally, Talacko (1956, 1958) focused on $m = 1$, $a_0 = 0$ and $n = 2$, $b_0 = b_2$. It took about 50 years until Talacko's generalized secant hyperbolic (GSH) distribution was re-examined by Vaughan (2002) and Klein and Fischer (2008) under a slightly different parameterization. Skew versions of the GSH distribution where introduced and successfully applied to financial return data in Fischer (2004, 2006) and Palmitesta and Provasi (2004).

Cross References

►Financial Return Distributions
►Generalized Hyperbolic Distributions
►Location-Scale Distributions

References and Further Reading

Baten WD (1934) The probability law for the sum of n independent variables, each subject to the law $(1/(2h))\,\mathrm{sech}\,(\pi x/(2h))$ Bull Am Math Soc 40:284–290
Dodd EL (1925) The frequency law of a function of variables with given frequency laws. Ann Math 27(2):13
Fisher RA (1921) On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1(4):3–32
Fischer M (2004) Skew generalized secant hyperbolic distributions: unconditional and conditional fit to asset returns. Aust J Stat 33(3):293–304
Fischer M (2006) A skew generalized secant hyperbolic family. Aust J Stat 35(4):437–444
Harkness WL, Harkness ML (1968) Generalized hyperbolic secant distributions. J Am Stat Assoc 63(321):329–337
Klein I, Fischer M (2008) A note on the kurtosis ordering of the generalized secant hyperbolic family. Commun Stat-Theor M 37(1):1–7
Meixner J (1934) Orthogonale Polygonsysteme mit einer besonderen Gestalt der erzeugenden Funktion. J London Math Soc 9:6–13
Morris CN (1982) Natural exponential families with quadratic variance functions. Ann Stat 10(1):65–80
Palmitesta P, Provasi C (2004) GARCH-type models with generalized secant hyperbolic innovations. Stud Nonlinear Dyn E 8(2):1–17
Perks W (1932) On some experiments in the graduation of mortality statistics. J Inst Actuaries 63:12–57
Roa E (1924) A number of new generating functions with applications to statistics. Thesis, University of Michigan
Schoutens W (2003) Lévy processes in finance – pricing financial derivatives. Wiley Series in Probability and Statistics. Wiley, New York
Talacko J (1958) A note about a family of Perks' distribution. Sankhya 20(3,4):323–328
Talacko J (1956) Perks' distribution and their role in the theory of Wiener's stochastic variables. Trabajos de Estatistica 17:159–174
Vaughan DC (2002) The generalized hyperbolic secant family and its properties. Commun Stat-Theor M 31(2):219–238

Hypergeometric Distribution and Its Application in Statistics

Anwar H. Joarder
Professor
King Fahd University of Petroleum and Minerals,
Dhahran, Saudi Arabia

An important discrete distribution encountered in sampling situations is the hypergeometric distribution. Suppose that a finite population of N items contains two types of items in which K items are of one kind (say defective) and $N-K$ items are of a different kind (say non-defective). If n items are drawn at random in succession, without replacement, then X denoting the number of defective items selected follows a hypergeometric distribution. The probability of the event $D_1D_2\cdots D_xD'_{x+1}\cdots D'_n$ denoting x successive defectives items and $n-x$ successive non-defective items is given by

$$P\left(D_1D_2\cdots D_xD'_{x+1}\cdots D'_n\right)=\frac{C_{K-x}^{N-n}}{C_K^N},\quad \max\{0,n-(N-K)\}\le x\le \min\{n,K\},\qquad (1)$$

where C_x^n is the number of combinations of x items that can be chosen from a group of n items and is equal to $n!/[x!(n-x)!]$. The probability of any other particular sequence in the sample space is also the same as (1). Interested readers may refer to Joarder and Al-Sabah (2007).

Since there are C_x^n outcomes having x defective items and $(n-x)$ non-defective items out of at most 2^n elements in the sample space, the probability of x successes in n trials is given by

$$P(X=x)=\frac{C_x^nC_{K-x}^{N-n}}{C_K^N},\quad \max\{0,n-(N-K)\}\le x\le \min\{n,K\},\qquad (2)$$

(cf. Kendall and Stuart 1969, p. 133). The probability of x successes in n trials is commonly written as

$$P(X=x)=\frac{C_x^KC_{n-x}^{N-K}}{C_n^N},\quad \max\{0,n-(N-K)\}\le x\le \min\{n,K\}.\qquad (3)$$

Vandermonde's identity justifies the equivalence of the two forms in (2) and (3). The proof of (3) is available in most textbooks on statistics (e.g., Johnson 2007) and discrete mathematics (e.g., Barnett 1998). There are C_x^K ways of choosing x of the K items (say defective items) and C_{n-x}^{N-K} ways of choosing $(n-x)$ of the $(N-K)$ non-defective items, and hence there are $C_x^KC_{n-x}^{N-K}$ ways of choosing x defectives and $(n-x)$ non-defective items. Since there are C_n^N ways of choosing n of the N elements, assuming C_n^N sample points are equally likely, the probability of having x defective items in the sample is given by (3).

The name hypergeometric is derived from a series introduced by the Swiss mathematician and physicist, Leonard Euler, in 1769. The probabilities in (3) are the successive terms of

$$\frac{(N-n)!(N-K)!}{N!(N-K-n)!}\;{}_2F_1(-n,-K;N-K-n+1;1),\qquad (4)$$

where ${}_2F_1(a_1, a_2; b; x)$ is the generalized hypergeometric function (Johnson et al. 1993, p. 237).

Suppose that a random sample of $n = 3$ items is selected from a lot of $N = 5$ items in which there are $K = 3$ defective items (distinguishable or indistinguishable) and 2 non-defective items (distinguishable or indistinguishable). Let $D_i\,(i = 1, 2, 3)$ be the event that we have a defective item in the ith selection, and $N_i\,(i = 1, 2, 3)$ be the event that we have a non-defective item in the ith selection. Also let X be the number of defective items selected in the sample. The elements of the sample space are given by $D_1D_2D_3, D_1D_2N_3, D_1N_2D_3, D_1N_2N_3, N_1D_2D_3$, $N_1D_2N_3$, and $N_1N_2D_3$. By (1), the probabilities are given by 0.10, 0.2, 0.2, 0.10, 0.2, 0.10 and 0.10 respectively. Note that elements in the sample space are not equiprobable. The probability of having two defective items in the sample is given by $P(X = 2) = 0.2 + 0.2 + 0.2 = 0.6$. If n is large, it is not feasible to write out the sample space but one can use (2) directly.

Note that if the items in each of the two categories are distinguishable, or labeled to make them distinguishable, the sample space can be written out with all distinguishable items. Then the sample outcomes are equally likely or equiprobable resulting in a Simple Random Sampling. In the above example, let the defective items be labeled as D^1, D^2 and D^3 while the non-defective items be labeled as N^1 and N^2 to make the items in the population distinguishable. Then $C_n^N = 10$ elements in the sample space are given by $D^1D^2D^3$, $D^1D^2N^1$, $D^1D^2N^2$, $D^1D^3N^1$, $D^1D^3N^2$, $D^1N^1N^2$, $D^2D^3N^1$, $D^2D^3N^2$, $D^2N^1N^2$, and $D^3N^1N^2$ each with the probability 0.10, and hence $P(X = 2) = 0.6$. In case C_n^N is large, it is not feasible to write out the sample space but one can use (3) directly.

Suppose that n items are drawn at random, with replacement, and X denotes the number of defective items selected. The probability that any item is defective at any draw is $p = K/N$ (say). Then with arguments similar to above, the probability of having x defectives and $(n - x)$ non-defectives in any of the C_x^n sequence is given by $p^x q^{n-x}$ so that $P(X = x) = C_x^n p^x q^{n-x}$. Now if $N \to \infty$, and $p = K/N$, it is easy to prove that (2) has a limiting value of $C_x^n p^x q^{n-x}$. This shows the equivalence of binomial and hypergeometric distribution in the limit.

The mean and variance of hypergeometric distribution are given by np and $(1 - f)npq$ respectively, where $p = K/N$, $q = 1-p$, and f is the finite population correction factor defined by $(N-1)f = N - n$. The mode of the distribution is the greatest integer not exceeding $\frac{(n+1)(K+1)}{N+2}$.

The coefficient of skewness and that of kurtosis are given by

$$\frac{(N-2K)(N-2n)(N-1)^{1/2}}{[nK(N-K)(N-n)]^{1/2}(N-2)}, \tag{5}$$

and

$$\frac{N^2(N-1)}{n(N-2)(N-3)(N-n)}\left[\frac{N(N+1)-6n(N-n)}{K(N-K)} + \frac{3n(N-n)(N+6)}{N^2} - 6\right], \tag{6}$$

respectively (Evans et al. 2000, p. 111).

The maximum likelihood estimator of the number of defectives K in a lot is the greatest integer not exceeding $x(N + 1)/n$; if $x(N + 1)/n$ is an integer, then $[x(N+1)/n]-1$ also maximizes the likelihood function (Johnson et al. 1993, p. 263).

The distribution has got a number of important applications in the real world. In the industrial quality control, lots of size N containing a proportion of p defectives are sampled by using samples of fixed size n. The number of defectives X per sample follows a hypergeometric distribution. If $X \le c$ (the acceptance number), the lot is accepted; otherwise it is rejected. The design of suitable sampling plans requires the calculation of confidence intervals of Np, given N, n and c. Tables of these have been published by Chung and DeLury (1950) and Owen (1962). It is worth mentioning that in many cases binomial or Poisson approximations to the hypergeometric distribution suffice.

Another useful application is the estimation of the size of the animal populations from capture–recapture data. This kind of application dates back to Peterson (1896), quoted by Chapman (1952). Consider, for example, the estimation of the number N of animals in a population. A sample of size K is captured, tagged and then released into the population. After a while a new catch of n animals is made, the number of tagged animals (X) in the sample is noted. Assume that the two catches are random samples from the population of all animals. Indeed, if we assume that there were no births or deaths, then the proportion of tagged animals in the sample (X/n) is approximately the same as that in the population (K/N). That is, an estimate of N is $\hat{N} = nK/X$. It may be noted that this estimate maximizes the probability of observed value of X. Evidently, X has a hypergeometric distribution with probability mass function given by (2) or (3).

The hypergeometric distribution can be approximated by Poisson distribution with parameter λ if K, N and n all tend to infinity for K/N small and nK/N tending to λ. It

can also be approximated by normal distribution if n is large, but x/N is not too small. A concise description of many other types of hypergeometric distribution and their properties are available in Johnson et al. (1993).

About the Author

Dr. Anwar H Joarder has been working at the Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia since 1997. He also worked at the University of Western Ontario, Jahangir Nagar University, University of Dhaka, North South University, Monash University and the University of Sydney. He is an Elected member of the Royal Statistical Society and the International Statistical Institute. He authored and co-authored around 60 research papers. He serves on the editorial board of five journals including *Model Assisted Sampling and Applications* and *International Journal of Mathematical Education in Science and Technology*.

Cross References

►Exact Inference for Categorical Data
►Fisher Exact Test
►Minimum Variance Unbiased
►Most Powerful Test
►Multivariate Statistical Distributions
►Poisson Distribution and Its Application in Statistics
►Proportions, Inferences, and Comparisons
►Statistical Distributions: An Overview
►Univariate Discrete Distributions: An Overview

References and Further Reading

Barnett S (1998) Discrete mathematics: numbers and beyond. Pearson Education Limited, Essex, England

Chapman DG (1952) Inverse, multiple and sequential sample censuses. Biometrics 8:286–306

Chung JH, DeLury DB (1950) Confidence limits for the hypergeometric distribution. University of Toronto Press, Toronto, Canada

Evans M, Hastings N, Pecock B (2000) Statistical distributions. Wiley, New York

Joarder AH, Al-Sabah WS (2007) Probability issues in without replacement sampling. Int J Math Educ Sci Technol 38(6):823–831

Johnson R (2007) Miller and Freund's Probability and statistics for engineers. Pearson Educational International, New Jersey, USA

Johnson NL, Kotz S, Kemp AW (1993) Univariate discrete distributions. Wiley, New York, USA

Kendall MG, Stuart A (1969) The advanced theory of statistics, vol 1: distribution theory. Charles Griffin, London

Owen DB (1962) Handbook of statistical tables. Addison-Wesley, Reading, MA

Peterson CGJ (1896) The yearly immigration of young plaice into the Limfjord from the German sea. Danish Biol Station Rep 6:5–48

Identifiability

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

Consider a vector $\boldsymbol{Y}$ of random variables having a distribution $F(\boldsymbol{y};\boldsymbol{\theta})$ that depends on an unknown parameter vector $\boldsymbol{\theta}$. $\boldsymbol{\theta}$ is *identifiable* by observation of $\boldsymbol{Y}$ if distinct values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ for $\boldsymbol{\theta}$ yield distinct distributions for $\boldsymbol{Y}$, that is, if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ implies $F(\boldsymbol{y};\boldsymbol{\theta}_1) \neq F(\boldsymbol{y};\boldsymbol{\theta}_2)$ for some $\boldsymbol{y}$ (Bickel and Doksum 1977). A function $g(\boldsymbol{\theta})$ is *identifiable* by observation of $\boldsymbol{Y}$ if $g(\boldsymbol{\theta}_1) \neq g(\boldsymbol{\theta}_2)$ implies $F(\boldsymbol{y};\boldsymbol{\theta}_1) \neq F(\boldsymbol{y};\boldsymbol{\theta}_2)$ for some $\boldsymbol{y}$. Note that $\boldsymbol{\theta}$ is identifiable if and only if all functions of $\boldsymbol{\theta}$ are identifiable.

There is some variation in the definition of identifiability, the preceding being the most general and central to topics such as ►*Bias analysis*. Variants typically employ the density $f(\boldsymbol{y};\boldsymbol{\theta})$ or the expectation $E(\boldsymbol{Y};\boldsymbol{\theta})$ in place of the distribution. The latter variants may explicitly involve a design matrix X of regressors; for example, $E(\boldsymbol{Y};X,\boldsymbol{\theta})$. The basic concept, however, is that $\boldsymbol{\theta}$ [or $g(\boldsymbol{\theta})$] is a function of the $\boldsymbol{Y}$ distribution, and hence observations from the distribution of $\boldsymbol{Y}$ can be used to discriminate among distinct values of $\boldsymbol{\theta}$ [or $g(\boldsymbol{\theta})$].

The term *estimable* is sometimes used as a synonym for identifiable, but is also used in more specific ways, especially in the context of linear models. For example, Scheffé (1959) defines a linear function $\boldsymbol{c}'\boldsymbol{\theta}$ of $\boldsymbol{\theta}$ to be estimable if there exists an unbiased estimator of $\boldsymbol{c}'\boldsymbol{\theta}$ that is a linear function of the observed realizations of $\boldsymbol{Y}$. This property has also been referred to as linear estimability. In epidemiology, estimability of $g(\boldsymbol{\theta})$ is sometimes used to mean that $g(\boldsymbol{\theta})$ can be consistently estimated from observable realizations of $\boldsymbol{Y}$. Several other definitions have been given, e.g., see Lehman (1983), McCullagh and Nelder (1989), and Seber and Wild (1989).

About the Author

For biography *see* the entry ►Confounding and Confounder Control.

Cross References

►Best Linear Unbiased Estimation in Linear Models
►Bias Analysis
►Mixture Models

References and Further Reading

Bickel PJ, Doksum KA (1977) Mathematical statistics. Holden-Day, Oakland
Lehmann EL (1983) Theory of point estimation. Wiley, New York
McCullagh P, Nelder JA (1989) Generalized linear models. Chapman and Hall, New York
Scheffé H (1959) The analysis of variance. Wiley, New York
Seber GAF, Wild CJ (1989) Nonlinear regression. Wiley, New York

Imprecise Probability

Frank P. A. Coolen[1], Matthias C. M. Troffaes[1], Thomas Augustin[2]
[1]Durham University, Durham, UK
[2]Ludwig Maximilians University, Munich, Germany

Overview

Quantification of uncertainty is mostly done by the use of precise probabilities: for each event A, a single (classical, precise) probability $P(A)$ is used, typically satisfying Kolmogorov's axioms (Augustin and Cattaneo 2010). Whilst this has been very successful in many applications, it has long been recognized to have severe limitations. Classical probability requires a very high level of precision and consistency of information, and thus it is often too restrictive to cope carefully with the multi-dimensional nature of uncertainty. Perhaps the most straightforward restriction is that the quality of underlying knowledge cannot be adequately represented using a single probability measure. An increasingly popular and successful generalization is available through the use of *lower and upper probabilities*, denoted by $\underline{P}(A)$ and $\overline{P}(A)$ respectively, with $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$, or, more generally, by lower and upper expectations (previsions) (Smith 1961; Walley 1991; Williams 2007). The special case with $\underline{P}(A) = \overline{P}(A)$ for

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

all events A provides precise probability, whilst $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$ represents complete lack of knowledge about A, with a flexible continuum in between. Some approaches, summarized under the name *nonadditive probabilities* [18], directly use one of these set-functions, assuming the other one to be naturally defined such that $\underline{P}(A^c) = 1 - \overline{P}(A)$, with A^c the complement of A. Other related concepts understand the corresponding intervals $[\underline{P}(A), \overline{P}(A)]$ for all events as the basic entity (Weichselberger 2000, 2001). Informally, $\underline{P}(A)$ can be interpreted as reflecting the evidence certainly in favour of event A, and $1 - \overline{P}(A)$ as reflecting the evidence against A hence in favour of A^c.

The idea to use imprecise probability, and related concepts, is quite natural and has a long history (see Hampel 2009 for an extensive historical overview of nonadditive probabilities), and the first formal treatment dates back at least to the middle of the nineteenth century (Boole 1854). In the last twenty years the theory has gathered strong momentum, initiated by comprehensive foundations put forward by Walley (1991) (see Miranda (2008) for a recent survey), who coined the term *imprecise probability*, by Kuznetsov (1991), and by Weichselberger (2000, 2001), who uses the term *interval probability*. Walley's theory extends the traditional subjective probability theory via buying and selling prices for gambles, whereas Weichselberger's approach generalizes Kolmogorov's axioms without imposing an interpretation. Usually assumed consistency conditions relate imprecise probability assignments to non-empty closed convex sets of classical probability distributions. Therefore, as a welcome by-product, the theory also provides a formal framework for models used in frequentist robust statistics (Augustin and Hable 2010) and robust Bayesian approaches (Rios Insua and Ruggeri 2000). Included are also concepts based on so-called two-monotone (Huber and Strassen 1976) and totally monotone capacities, which have become very popular in artificial intelligence under the name (Dempster–Shafer) belief functions (Dempster 1967; Shafer 1976). Moreover, there is a strong connection (de Cooman and Hermans 2008) to Shafer and Vovk's notion of game-theoretic probability (Shafer and Vovk 2001).

The term "imprecise probability" – although an unfortunate misnomer as lower and upper probability enable more accurate quantification of uncertainty than precise probability – appears to have been established over the last two decades, and actually brings together a variety of different theories. In applications, clear advantages over the established theory of precise probability have been demonstrated (see sections "►Applications in Statistics and Decision Theory" and "►Further Applications"). This justifies the further development of imprecise probability, particularly toward building a complete methodological framework for applications in statistics, decision support, and related fields. Imprecise probability provides important new methods that promise greater flexibility for uncertainty quantification. Its advantages include the possibility to deal with conflicting evidence, to base inferences on weaker assumptions than needed for precise probabilistic methods, and to allow for simpler and more realistic elicitation of subjective information, as imprecise probability does not require experts to represent their judgements through a full probability distribution, which often does not reflect their beliefs appropriately.

The Society for Imprecise Probability: Theories and Applications (www.sipta.org) organizes conferences, workshops and summer schools, and provides useful introductory information sources and contacts through its web-page.

The increased attention to imprecise probability during the last two decades has led to many new methods for statistical inference and decision support, with applications in a wide variety of areas.

Applications in Statistics and Decision Theory

Following Walley (1991), many of the imprecise probability-based contributions to statistics follow a generalized Bayesian approach. Typically, a standard precise parametric sampling model with a set of prior distributions is used. In particular, the use of models from the exponential family is popular in conjunction with classes of conjugate priors. Walley's Imprecise Dirichlet Model (IDM) for inference in case of multinomial data (Walley 1996) has attracted particular attention (Bernard 2009). One successful application area for the IDM is classification (Zaffalon 2002), where the use of lower and upper probabilities makes the learning process more stable and enables in a quite natural way an item to be explicitly not classified into a unique category, indicating that no clear decision for a single category can be made on the basis of the information available. In these models, updating to take new information into account is effectively done by updating all elements of the set of prior distributions as in Bayesian statistics with precise prior distributions, leading to a set of posterior distributions which forms the basis for inferences. From the technical perspective this procedure is therefore closely related to robust Bayesian inference, but, by reporting the indeterminacy resulting from limited information, use and interpretation of the resulting imprecise posterior goes far beyond a simple sensitivity and robustness analysis.

Other approaches to statistical inference with imprecise probabilities have been developed, which tend to move further away from the precise probabilistic approaches. Examples of such approaches are Nonparametric Predictive Inference (Coolen 2010), generalizations of the frequentist approach, see e.g., Augustin (2002) for hypotheses testing and Hable (2010) for estimation, as well as several approaches based on logical probability (Kyburg 1974; Levi 1980; Weichselberger 2005).

Imprecise probabilities have also proven their use in decision support (Troffaes 2007), where, in tradition of Ellsberg's experiments (Ellsberg 1961), ambiguity (or non-stochastic uncertainty) plays a crucial role (Hsu et al. 2005). If only little information is available about a variable, then it is often more natural to refuse to determine a unique optimal decision, when gains and losses depend on that variable. Imprecise probability theory grasps this in a rigorous manner, resulting in a set of possibly optimal decisions, rather than providing only a single, arbitrarily chosen, optimal decision from this set. Imprecise probability theory is especially useful in critical decision problems where gains and losses heavily depend on variables which are not completely known, such as for instance in pollution control (Chevé and Congar 2000) and medical diagnosis (Zaffalon et al. 2003).

Further Applications

Recent collections of papers (Augustin et al. 2009; Coolen-Schrijner et al. 2009; de Cooman et al. 2007) give an impression of the huge variety of fields of potential application. In artificial intelligence, for example in pattern recognition (Loquin and Strauss 2010, see also ►Pattern Recognition, Aspects of and ►Statistical Pattern Recognition Principles) and information fusion (Benavoli and Antonucci 2010), uncertain expert knowledge can be represented more accurately by means of imprecise probability. Because imprecise probability methods can process information without having to add unjustified assumptions, they are of great importance in risk and safety evaluations, design engineering (Aughenbaugh and Paredis 2006) and reliability (Coolen and Utkin 2010). The ongoing intensive debate on bounded rationality makes reliable decision theory based on imprecise probability particularly attractive in microeconomics and in social choice theory. In finance, imprecise probability is gaining strong influence given its very close connection to risk measures (Artzner et al. 1999; Vicig 2008). Imprecise probability also yields deeper insight into asset pricing (Richmond et al. 2008). The study of ►Markov chains with imprecise transition probabilities (de Cooman et al. 2009) is also important for many areas of application.

Challenges

Imprecise probability and its applications in statistical inference and decision support offer a wide range of research challenges. On foundations, key aspects such as updating have not yet been fully explored, and different approaches have different conditioning rules. The relation between imprecision and information requires further study, and many of the most frequently used statistical methods (such as complex regression models) have not yet been fully generalized to deal with imprecise probability. In cases where generalizations are easily found, it may be unclear which of many possible approaches is most suitable. Of course, early developments of new theoretic frameworks tend to include illustrative applications to mostly text-book style problems. The next stage required toward widely applicable methods involves upscaling, where in particular computational aspects provide many challenges. Even methods such as simulation, mostly straightforward with precise probabilities, become non-trivial with imprecise probabilities.

For applications which require the use of subjective information, elicitation of expert judgements is less demanding when lower and upper probabilities are used, but while practical aspects of elicitation have been widely studied this has, thus far, only included very few studies involving imprecise probabilities.

In decision making, algorithms to find optimal solutions need to be improved and implemented for large-scale applications. As many problems have a sequential nature, ways of representing sequential solutions efficiently also need to be developed, the more so as classical techniques such as backward induction and dynamic programming often cannot be extended directly.

About the Authors

For the bibliography of Frank Coolen, *see* the entry ►Nonparametric Predictive Inference.

For the bibliography of Thomas Augustin, *see* the entry ►Foundations of Probability.

Since 2006, Dr Troffaes is lecturer in statistics at the Department of Mathematical Sciences, Durham University, UK. He has authored and co-authored various papers in the field of imprecise probability theory and decision making, and co-edited special issues for the *Journal of Statistical Practice and Applications* (2009), *International Journal of Approximate Reasoning* (2010), and the *Journal of Risk and Reliability* (2010). Since 2009, he is member of the executive committee of the Society for Imprecise Probability: Theories and Applications (SIPTA).

Cross References

►Foundations of Probability
►Imprecise Reliability
►Nonparametric Predictive Inference

References and Further Reading

Artzner P, Delbaen F, Eber J-M, Heath D (1999) Coherent measures of risk. Math Finance 9(3):203–228

Aughenbaugh JM, Paredis CJJ (2006) The value of using imprecise probabilities in engineering design. J Mech Design 128(4): 969–979

Augustin T (2002) Neyman-Pearson testing under interval probability by globally least favorable pairs – reviewing Huber-Strassen theory and extending it to general interval probability. J Stat Plann Infer 105(1):149–173

Augustin T, Cattaneo M (2010) International encyclopedia of statistical sciences, chapter Foundations of probability. Springer

Augustin T, Coolen FPA, Moral S, Troffaes MCM (eds) (2009) ISIPTA'09: Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications, Durham University, Durham, UK, July 2009. SIPTA

Augustin T, Hable R (2010) On the impact of robust statistics on imprecise probability models: a review. Structural Safety. To appear

Benavoli A, Antonucci A (2010) Aggregating imprecise probabilistic knowledge: application to zadeh's paradox and sensor networks. Int J Approx Reason. To appear

Bernard J-M (2009) Special issue on the Imprecise Dirichlet Model. Int J Approx Reason 50:201–268

Boole G (1854) An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities. Walton and Maberly, London

Chevé M, Congar R (2000) Optimal pollution control under imprecise environmental risk and irreversibility. Risk Decision Policy 5:151–164

Coolen FPA (2010) International encyclopedia of statistical sciences, chapter Nonparametric predictive inference. Springer

Coolen FPA, Utkin LV (2010) International encyclopedia of statistical sciences, chapter Imprecise reliability. Springer

Coolen-Schrijner P, Coolen F, Troffaes M, Augustin T (2009) Special issue on statistical theory and practice with imprecision. J Stat Theor Practice 3(1):1–303. Appeared also with Sat Gupta as additional editor under the title Imprecision in statistical theory and practice. Grace, Greensboro

de Cooman G, Hermans F (2008) Imprecise probability trees: Bridging two theories of imprecise probability. Artif Intell 172(11):1400–1427

de Cooman G, Hermans F, Quaeghebeur E (2009) Imprecise Markov chains and their limit behavior. Probab Eng Inform Sci 23(4):597–635

de Cooman G, Vejnarová J, Zaffalon M (eds) (2007) ISIPTA'07: Proceedings of the Fifth International Symposium on Imprecise Probabilities and Their Applications, Charles University, Prague, Czech Republic, July 2007. SIPTA

Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38:325–339

Denneberg D (1994) Non-additive measure and integral. Kluwer, Dordrecht

Ellsberg D (1961) Risk, ambiguity, and the savage axioms. Q J Econ 75:643–669

Hable R (2010) Minimum distance estimation in imprecise probability models. J Stat Plann Infer 140:461–479

Hampel F (2009) Nonadditive probabilities in statistics. J Stat Theor Practice 3(1):11–23

Hsu M, Bhatt M, Adolphs R, Tranel D, Camerer CF (2005) Neural systems responding to degrees of uncertainty in human decision-making. Science 310:1680–1683

Huber PJ, Strassen V (1973) Minimax tests and the Neyman–Pearson lemma for capacities. Ann Stat 1:251–263

Kuznetsov VP (1991) Interval statistical models. Radio i Svyaz Publ., Moscow, 1991. In Russian

Kyburg HE Jr (1974) The logical foundations of statistical inference. Springer, Synthese Library

Levi I (1980) Enterprise of knowledge: essay on knowledge, credal probability and chance. MIT, Cambridge, MA

Loquin K, Strauss O (2010) Noise quantization via possibilistic filtering. Int J Approx Reason. To appear

Miranda E (2008) A survey of the theory of coherent lower previsions. Int J Approx Reason 48(2):628–658

Richmond V, Jose R, Nau RF, Winkler RL (2008) Scoring rules, generalized entropy, and utility maximization. Oper Res 56(5):1146–1157

Rios Insua D, Ruggeri F (eds) (2000) Robust Bayesian analysis. Springer, New York

Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton

Shafer G, Vovk V (2001) Probability and finance: it's only a game. Wiley, New York

Smith CAB (1961) Consistency in statistical inference and decision. J R Stat Soc B 23:1–37

Troffaes MCM (2007) Decision making under uncertainty using imprecise probabilities. Int J Approx Reason 45(1):17–29

Vicig P (2008) Imprecise probabilities in finance and economics. International Journal of Approximate Reasoning 49(1): 99–100

Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall, London

Walley P (1996) Inferences from multinomial data: learning about a bag of marbles. J R Stat Soc B 58(1):3–34

Weichselberger K (2000) The theory of interval probability as a unifying concept for uncertainty. Int J Approx Reason 24: 149–170

Weichselberger K (2001) Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I – Intervallwahrscheinlichkeit als umfassendes Konzept. Physica, Heidelberg, 2001. In cooperation with T. Augustin and A. Wallner

Weichselberger K (2005) The logical concept of probability and statistical inference. In: Cozman FG, Nau R, Seidenfeld T (eds) ISIPTA '05: Proceedings of Fourth International Symposium on Imprecise Probabilities and Their Applications. pp 396–405

Williams PM (2007) Notes on conditional previsions. Technical report, School of Math. and Phys. Sci., Univ. of Sussex, 1975. Republished In Int J Approx Reason 44(3):366–383

Zaffalon M (2002) The naive credal classifier. J Stat Plann Infer 105(1):5–21

Zaffalon M, Wesnes K, Petrini O (2003) Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. Artif Intell Med 29(1–2):61–79

Imprecise Reliability

Frank P. A. Coolen[1], Lev V. Utkin[2]
[1]Professor
Durham University, Durham, UK
[2]Professor, Vice-rector for Research work
St. Petersburg State Forest Technical Academy,
St. Petersburg, Russia

Overview

Reliability analysis is an important application area of statistics and probability theory in engineering, with several specific features which often complicate application of standard methods. Such features include data censoring, for example due to maintenance activities, lack of knowledge and information on dependence between random quantities, for example if failures occur due to competing risks, and required use of expert judgements, for example when new or upgraded versions of units are used. While mathematical approaches for dealing with such issues have been presented within the framework of statistics using precise probabilities, the use of imprecise probability (Coolen et al. 2010) provides exciting new ways for dealing with such challenges in reliability. One of the first approaches that generalized probability in reliability was fuzzy reliability theory (Cai 1996), but this suffers from vagueness about axioms and rules for combination of information, and lack of clear interpretation of the results. We restrict attention to generalized uncertainty quantification in reliability via lower and upper probabilities, also known as "imprecise probability" (Walley 1991) or "interval probability" (Weichselberger 2000, 2001). During the last two decades, imprecise probability (Coolen et al. 2010) has received increasing attention, and interesting applications have been reported. It is widely accepted that, by generalizing precise probability theory in a mathematically sound manner, with clear axioms and interpretations, this theory provides a better approach to generalized uncertainty quantification then its current alternatives.

In classical probability theory, a single probability $P(A) \in [0,1]$ is used to quantify uncertainty about event A. Imprecise probability theory (Walley 1991; Weichselberger 2000, 2001) generalizes this by using lower probability $\underline{P}(A)$ and upper probability $\overline{P}(A)$ such that $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$, where the difference between $\overline{P}(A)$ and $\underline{P}(A)$ represents lack of perfect information about the uncertainty involving event A, see Coolen et al. (2010) for a further introduction. For reliability, attractive features of this generalization include that one does not need to make strong assumptions in order to derive at precise probabilities for all situations. For example, one may have partial information about dependence of failure times for different components, or one may have to rely on expert judgement with an only limited elicitation possible due to time constraints. All kinds of partial information that might be available in practice can be formulated as constraints on underlying probabilities, which can be satisfied by sets of probabilities.

Applications

A recent extensive introduction to imprecise reliability, together with a discussion of many applications, has been presented by Utkin and Coolen (2007). As an example of an imprecise reliability application, Fetz and Tonon (2008) consider bounds for the probability of failure of a series system when no information is available about dependence between failure probabilities of different modes. They consider several models, including random sets and p-boxes, and they provide a detailed list of references to the literature on such topics. They also discuss some computational methods, which is an important aspect of application of imprecise reliability to medium or larger size practical problems.

One of the possible ways in which output from imprecise probability methods can be useful is in the study of sensitivity of model outputs with respect to variations in input parameters. An interesting recent study by Oberguggenberger et al. (2009) presents such an approach to a large-scale modeling problem to assess reliability in an aerospace engineering application, comparing the use of classical probabilities and a variety of imprecise probability methods.

Imprecise probabilistic approaches to statistics are of great value to reliability problems. Recent examples include the use of Walley's imprecise Dirichlet model (Walley 1996) for system reliability without detailed assumptions on dependence of components (Troffaes and Coolen 2009), and system reliability from the perspective of nonparametric predictive inference (Coolen 2010). Another recent development is combination of imprecise Bayesian methods for some parameters with a generalized maximum likelihood approach for other parameters in an inferential problem, where the former can be used to explicitly deal with incomplete expert judgements while the latter can be appropriate on aspects of the problem for which data are available but no strong expert opinions. This has been explored for software reliability growth models, using the maximum likelihood approach for the temporal growth aspect together with imprecise Bayesian methods

for the parameters modeling the stationary aspects of the model (Utkin et al. 2009).

Challenges

Imprecise reliability is a relatively new area of research, with methods presented that are inspired by practical problems but that are not yet suitable for applications of substantial size. The main challenges are in upscaling the methods to become useful for practical problems, together with aspects of implementation which include consideration of elicitation and model choice. The combination of substantial optimization problems and statistical modeling and updating may also lead to a level of complexity that requires attention to methods for computation, for example it is not clear how modern simulation-based methods, that are for example popular in ►Bayesian statistics, can best be used or adapted for imprecise approaches.

The models for imprecise reliability that have been presented so far are still pretty basic, and for example inclusion of covariates requires further research. Generally, imprecise approaches can be found that generalize the established methods in varying ways, so in addition to developing new methods one must find ways to decide on how useful they are, which requires careful consideration of fundamental aspects of uncertainty and information. Hybrid methods, which combine imprecise models where useful to model indeterminacy with precise models where possible due to sufficient data or information, provide exciting opportunities for research, with issues that must be addressed including interpretation of results, choice of models and methods, and computation.

About the Authors

For biography of Frank P.A. Coolen *see* the entry ►Nonparametric Predictive Inference.

Lev V. Utkin received his M.Sc. (1986) from the Institute of Electrical Engineering in Leningrad. He obtained his Ph.D. in information processing and control systems (1989) from the same institute. Currently, he is Vice-rector for Research work and professor in the Department of Computer Science, St. Petersburg Forest Technical Academy. His research papers are on fuzzy reliability theory, the theory of imprecise probabilities, imprecise probability models in the reliability theory, optimization and uncertainty representation.

Cross References

►Imprecise Probability
►Industrial Statistics
►Parametric and Nonparametric Reliability Analysis

References and Further Reading

Cai KY (1996) Introduction to fuzzy reliability. Kluwer Academic, Boston
Coolen FPA (2010) Nonparametric predictive inference. International encyclopedia of statistical sciences. Springer (this volume)
Coolen FPA, Troffaes MCM, Augustin T (2010) Imprecise probability. International encyclopedia of statistical sciences. Springer (this volume)
Fetz T, Tonon F (2008) Probability bounds for series systems with variables constrained by sets of probability measures. Int J Reliability and Safety 2:309–339
Oberguggenberger M, King J, Schmelzer B (2009) Classical and imprecise probability methods for sensitivity analysis in engineering: a case study. Int J Approx Reason 50:680–693
Troffaes M, Coolen FPA (2009) Applying the imprecise dirichlet model in cases with partial observations and dependencies in failure data. Int J Approx Reason 50:257–268
Utkin LV, Coolen FPA (2007) Imprecise reliability: an introductory overview. In: Levitin G (ed) Computational intelligence in reliability engineering, vol 2: New metaheuristics, neural and fuzzy techniques in reliability. Springer, New York, pp 261–306
Utkin LV, Zatenko SI, Coolen FPA (2009) Combining imprecise Bayesian and maximum likelihood estimation for reliability growth models. In: Proceedings ISIPTA'09 www.sipta.org/isipta09, pp 421–430
Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman and Hall, London
Walley P (1996) Inferences from multinomial data: learning about a bag of marbles. J R Stat Soc B 58:3–34
Weichselberger K (2000) The theory of interval-probability as a unifying concept for uncertainty. Int J Approx Reason 24:149–170
Weichselberger K (2001) Elementare Grundbegriffe einer Allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als Umfassendes Konzept. Physika, Heidelberg

Imputation

Roderick J. Little
Richard D. Remington Collegiate Professor of Biostatistics
University of Michigan, Ann Arbor, MI, USA

Missing data is a difficult problem that degrades the quality of empirical studies. There are no foolproof methods, so the best approach is to seek to avoid missing values by careful design and prevention strategies, such as avoiding excessive respondent burden and strong efforts to elicit responses. Despite our best attempts to limit levels of nonresponse, missing values inevitably occur. For example, individuals in a sample survey refuse to answer items that are sensitive or difficult to answer; in a longitudinal study, individuals drop out prior to the end to a study because of relocation, or study fatigue. Analysis methods for data

subject to missing values are thus needed. A limited objective is to provide valid point estimates of population quantities from the incomplete data. Since good statistical analyses usually also require an assessment of statistical uncertainty of estimates through confidence intervals or test of hypotheses, it is also important that such inferences reflect the loss of information arising from missing data, so that, for example nominal 95% confidence intervals cover the true population quantity at least approximately 95% of the time, in repeated sampling.

Imputation is a method for the analysis of data with missing values, where missing values are replaced by estimates and the filled-in data are analyzed by complete-data methods. Often a single value is imputed (single imputation). Multiple imputation imputes more than one set of imputations of the missing values, allowing the assessment of imputation uncertainty.

A common alternative to imputation in public use files is to include incomplete cases in a data set, with missing value codes to indicate which values are missing. A drawback of this procedure is that different users may get different answers for the same research question, because they use different methods for dealing with the missing values. Most statistical software discards cases that have missing values on any of the variables included in the analysis, leading to complete-case analysis or listwise deletion (Little and Rubin 2002, Chap. 3). The information contained in observed variables in the incomplete cases is thus lost, and the complete cases may not be representative of the original sample.

Imputation is a method that allows incomplete cases to be included in the analysis. In fact, the main reason for imputation is not to recover the information in the missing values, which is lost and usually not recoverable, but rather to allow the information in observed values in the incomplete cases to be retained. If there is little information to be recovered in the cases with missing values, as for example cases in regression for which the outcome variable is missing, then imputation is not very useful.

A common naive imputation approach imputes missing values by their simple unconditional sample means (i.e., marginal means). This can yield satisfactory point estimates of unconditional means and totals, but it yields inconsistent estimates of other parameters, for example variances or regression coefficients (Kalton and Kasprzyk 1982; Little and Rubin 2002, Sect. 4.2). Inferences (tests and confidence intervals) based on the filled-in data are seriously distorted by bias and overstated precision. Thus the method cannot be generally recommended.

An improvement over unconditional mean imputation is *conditional mean* imputation, in which each missing value is replaced by an estimate of its conditional mean given the values of observed values. For example, in the case of univariate nonresponse with $Y_1, \ldots, Y_{p-1}$ fully observed and Y_p sometimes missing, one approach is to classify cases into cells based on similar values of observed variables, and then to impute missing values of Y_p by the within-cell mean from the complete cases in that cell. A more general approach is regression imputation, in which the regression of Y_p on $Y_1, \ldots, Y_{p-1}$ is estimated from the complete cases, including interactions as needed, and the resulting prediction equation is used to impute the estimated conditional mean for each missing value of Y_p. For a general pattern of missing data, the missing values for each case can be imputed from the regression of the missing variables on the observed variables, computed using the set of complete cases. Iterative versions of this method lead (with some important adjustments) to maximum likelihood estimates under multivariate normality (Little and Rubin 2002, Sect. 8.2).

Although conditional mean imputation incorporates information from the observed variables and yields best predictions of the missing values in the sense of mean squared error, it leads to distorted estimates of quantities that are not linear in the data, such as percentiles, correlations and other measures of association, variances and other measures of variability. A solution to this problem is to use random draws rather than best predictions to preserve the distribution of variables in the filled-in data set. An example is *stochastic regression* imputation, in which each missing value is replaced by its regression prediction plus a random error with variance equal to the estimated residual variance. Other imputation methods impute values observed in the dataset. One such method is the *hot-deck,* as used by the Census Bureau for imputing income in the Current Population Survey (CPS) (Hanson 1978). Each nonrespondent is matched to a respondent based on variables that are observed for both; the missing items for the nonrespondent are then replaced by the respondent's values. For the CPS, matching is achieved by classifying respondents and nonrespondents into adjustment cells based on the observed variables. When no match can be found for a nonrespondent based on all of the variables, the CPS hot-deck searches for a match at a lower level of detail, by omitting some variables and collapsing the categories of others. A more general approach to hot-deck imputation is to define a distance function based on the variables that are observed for both nonrespondents and respondents. The missing values for each nonrespondent are then imputed from a respondent that is close to the nonrespondent in terms of the distance function. For a review of hot-deck methods see Andridge and Little (2010).

A serious defect with imputation is that it invents data. More specifically, a single imputed value cannot represent

all of the uncertainty about which value to impute, so analyses that treat imputed values just like observed values generally underestimate uncertainty, even if nonresponse is modeled correctly and random imputations are created.

Two approaches to this deficiency are (a) to apply a replication method of variance estimation, and recompute the imputations on each replicate sample (Fay 1996, Rao 1996, Efron 1994), and (b) multiple imputation (MI) (Rubin 1987, 1996). Instead of imputing a single set of draws for the missing values, a set of M (say $M = 10$) datasets are created, each containing different sets of draws of the missing values from their predictive distribution. We then apply the analysis to each of the M datasets and combine the results in a simple way. In particular, for scalar estimants, the *MI* estimate is the average of the estimates from the M datasets, and the variance of the estimate is the average of the variances from the five datasets plus $1 + 1/M$ times the sample variance of the estimates over the M datasets (The factor $1 + 1/M$ is a small-M correction). The last quantity here estimates the contribution to the variance from imputation uncertainty, missed by single imputation methods. Another benefit of multiple imputation is that the averaging over datasets results in more efficient point estimates than does single random imputation. Often *MI* is not much more difficult than doing a single imputation – the additional computing from repeating an analysis M times is not a major burden and methods for combining inferences are straightforward. Multiple imputation is available for the multivariate normal model in a variety of software packages (e.g., Proc *MI* in SAS), and more flexible sequential multiple imputation methods are available in IVEware (http://www.isr.umich.edu/src/smp/ive/) and MICE (http://www.multiple-imputation.com/).

Most approaches imputation to date have assumed that the missing data are MAR (Rubin 1976, Little and Rubin 2002, Sect. 6.2), which means that differences in the distribution of the missing variables for respondents and nonrespondents can be captured using observed variables. Non-MAR models are needed when missingness depends on the missing values. For example, suppose a subject in an income survey refused to report an income amount because the amount itself is high (or low). If missingness of the income amount is associated with the amount, after controlling for observed covariates (such as age, education or occupation) then the mechanism is not MAR, and methods for imputing income based on MAR models are subject to bias. A correct analysis must be based on the full likelihood from a model for the joint distribution of the data and indicators for the missing values. ►Sensitivity analysis is the preferred approach to assess the impact of non-MAR missing values (e.g., Little and Rubin 2002, Chap. 15).

About the Author

Roderick Little Chaired the Biostatistics Department at the University of Michigan from 2007–1993–2001 and 2007–2010. Dr. Little is also a professor in the University's Statistics Department and a research professor in the University's Institute for Social Research. He is an Elected ASA Fellow (1985) an Elected Member of International Statistical Institute (1989), and a Fellow of the American Academy of Arts and Sciences (2010). He has over 160 publications (including *Statistical Analysis with Missing Data*, with D.B. Rubin, John Wiley, 1987), notably on methods for the analysis of data with missing values and model-based survey inference, and the application of statistics to diverse scientific areas Professor Little was awarded the Samuel S. Wilks Medal, American Statistical Association (2005) for "significant and pioneering contributions to the development of statistical methodology." Currently, he is Vice-President of the American Statistical Association (2010–2012).

Cross References

►Incomplete Data in Clinical and Epidemiological Studies
►Multi-Party Inference and Uncongeniality
►Multiple Imputation
►Nonresponse in Surveys
►Nonsampling Errors in Surveys
►Sampling From Finite Populations

References and Further Reading

Andridge RH, Little RJ (2010) A review of hot deck imputation for survey nonresponse. Int Stat Rev 78(1): 40–64
Efron B (1994) Missing data, imputation and the bootstrap. J Am Stat Assoc 89:463–479
Fay RE (1996) Alternative paradigms for the analysis of imputed survey data. J Am Stat Assoc 91:490–498. With discussion
Hanson RH (1978) The current population survey: design and methodology. Technical Paper, No. 40, U.S. Bureau of the Census
Kalton G, Kasprzyk D (1982) Imputing for missing survey nonresponses. In: Proceedings of the Survey Research Methods Section, American Statistical Association, pp 22–31
Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
Rao JNK (1996) On variance estimation with imputed survey data. J Am Stat Assoc 91:499–506. With discussion
Rubin DB (1976) Inference and missing data. Biometrika 63:581–592
Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
Rubin DB (1996) Multiple imputation after 18+ years. J Am Stat Assoc 91:473–489. With discussion

Incomplete Block Designs

Nam-Ky Nguyen[1], Kalina Trenevska Blagoeva[2]
[1]Vietnam National University, Hanoi, Vietnam
[2]Faculty of Economics
University "SS. Cyril and Methodius," Skopje, Macedonia

Introduction

Blocking is the division of experimental material into blocks or sets of homogeneous experimental units. Proper blocking can control the source of variability which is not of primary interest and thus can reduce the experimental error. If the number of treatments is the same as the block size, a randomized block design can be used. However, if the number of treatments exceeds the block size, an incomplete block design (IBD) should be considered.

An IBD of size (v, k, r) is an arrangement of v treatments set out in b blocks, each of size $k(< v)$ such that each treatment occurs in r blocks where $vr = bk$ and no treatment occurs more than once in any block. The following is an IBD of size $(v, k, r) = (4, 2, 3)$ (Note that all IBDs in this article are displayed with blocks as columns.):

0	2	0	1	1	0
1	3	3	2	3	2

An IBD is said to be r/s-resolvable if the blocks can be divided into s *replicate sets* (of blocks) and each set is an IBD of size $(v, k, r/s)$. A 1-resolvable IBD is a resolvable IBD (see the above example). The following is a 2-resolvable IBD of size $(v, k, r) = (6, 4, 4)$:

3	0	3	1	0	2
4	2	4	3	4	3
5	5	2	5	2	4
0	1	1	0	1	5

The books of John and Williams (1995) and Raghavarao and Padgett (2005) give a more complete treatment on the subject.

A Criterion for Comparing IBDs

Associated with each IBD is its (treatment) *concurrence* matrix $NN' = \{\lambda_{ij}\}$ where λ_{ij} $(i, j = 1, \ldots, v)$ is the number of blocks in which treatment i and j both appear. Obviously, $\lambda_{ii} = r$. When $\lambda_{ij} = \lambda$ for all $i \neq j$, the IBD is called a balanced IBD (BIBD). Below is an IBD of size $(v, k, r) = (6, 3, 2)$:

0	0	1	1
2	3	3	2
5	4	5	4

The concurrence matrix of this IBD is

$$\begin{pmatrix} 2 & 0 & 1 & 1 & 1 & 1 \\ 0 & 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 2 & 0 \\ 1 & 1 & 1 & 1 & 0 & 2 \end{pmatrix}.$$

As $\sum \lambda_{ij}$ is constant $(=vkr)$, $\sum \lambda_{ij}^2$ is minimized if λ_{ij}'s $(i \neq j)$ differ by at most 1. IBDs with this property were called regular graph designs (RGDs) by John and Mitchell (1977) who conjectured that D-, A- and E- optimal IBDs are also RGDs. Thus, RGDs include BIBDs, i.e., IBDs whose λ_{ij}'s $(i \neq j)$ do not differ and all near-BIBDs whose λ_{ij}'s $(i \neq j)$ differ by 1. RGD is an important class of IBDs not only because it has been conjectured that optimal IBDs are RGDs but also because most IBDs used by researchers in practice are actually RGDs. This fact has prompted Nguyen (1994) to search for RGDs as the first step in constructing optimal IBDs.

A common criterion for comparing IBDs of the same size is the *efficiency factor* defined as $E = (v-1)/\sum e_i^{-1}$ where e_i's are the $v-1$ nonzero eigenvalues of $r^{-1}C$ and $C = rI - k^{-1}NN'$ is the information matrix for the adjusted treatment effects. Here, we assume that the IBD is *connected*, i.e., $\text{rank}(C) = v-1$. An IBD which has the maximal value of E is said to be A-optimal (John and Williams 1995, Sect. 2.4). The upper bounds for the efficiency factor of an IBD have been discussed extensively in Sect. 2.6 of John and Williams (1995). These upper bounds are used to establish the stopping rule for any IBD algorithm.

IBDs with High Efficiency Factors

This section introduces BIBDs and some classes of computer-generated IBDs. These designs have become more popular among designers of experiments as with the advent of the computer, the flexibility and goodness of the design have succeeded ease of analysis as their criteria in design selection.

The parameters (v, b, k, r, λ) of a BIBD satisfy two relationships: (i) $bk = vr$ and (ii) $r(k-1) = \lambda(v-1)$. These two relationships, however, are necessary but not sufficient for

a BIBD to exist. For any combination of v and k ($k < v$), an *unreduced* BIBD can be constructed by taking all $b = \binom{v}{k}$ combinations of v treatments k at a time. The blocks of a BIBD for $v = b = 5$, $r = k = 4$ and $\lambda = 3$ are (0123), (0124), (0134), (0234) and (1234). Another well-known class of BIBDs which is resolvable and requires a smaller number of blocks and replications is *balanced lattice*. An $s \times s$ balanced lattice is a resolvable BIBD of size $v = s^2$, $b = s(s+1)$, $k = s$, $r = s+1$, and $\lambda = 1$ constructed from a complete set of $s \times s$ *mutually orthogonal Latin squares*. Chapters 4 and 9 of Raghavarao and Padgett (2005) describe the combinatorics of BIBDs and lattice designs in detail.

Cyclic IBDs are IBDs generated by the cyclic development of one or more suitably chosen initial blocks. Cyclic IBDs account for a large number of BIBDs in literature (see Table 4.4 of Raghavarao and Padgett 2005) . They also provide efficient alternatives to many partially BIBDs catalogued in Clatworthy (1973). Chap. 3 of John and Williams (1995) gives a summary of cyclic IBDs. When the number of replications r is equal to or is a multiple of the block size k, cyclic IBDs render automatic elimination of heterogeneity in two directions (see Sect. 5.7 of John and Williams 1995). The following is a cyclic BIBD for $v = 7$, $b = 14$, $k = 3$, $r = 6$ and $\lambda = 2$ generated by two initial blocks (1, 2, 6) and (1, 2, 4).

1	2	3	4	5	6	0	1	2	3	4	5	6	0
2	3	4	5	6	0	1	2	3	4	5	6	0	1
6	0	1	2	3	4	5	4	5	6	0	1	2	3

Patterson and Williams (1976a) introduced a new class of resolvable IBDs called α-design. α-designs are available for many (r, k, s) combinations where r is the number of replicates, k is the block size and s is the number of blocks per replicate (the number of treatments $v = ks$). An α-design was generated by an $r \times k$ array α with elements in set of residues mod s. Thus, the construction of an α-design (or a cyclic IBD) resorts to the construction of an array α (or one or more initial blocks). Chapter 4 of John and Williams (1995) gives a summary of resolvable IBDs including α-designs. α-designs and cyclic IBDs can be generated by the CycDesignN software (http://www.cycdesign.co.nz/) and the Gendex DOE toolkit (http://designcomputing.net/gendex/).

Cyclic solutions are not always optimal. Following is a non-cyclic solution for an IBD of size $(v, k, r) = (15, 6, 4)$. This optimal IBD with $E = 0.8861$ was constructed by the algorithm of Nguyen (1994).

7	5	4	6	5	8	8	10	13	14
9	13	6	10	13	2	5	0	12	7
5	7	8	12	14	6	3	9	3	1
12	4	0	4	9	13	2	2	8	12
3	0	9	14	11	7	14	11	1	0
6	10	1	11	1	11	4	3	10	2

Two Tools for IBD Construction

To construct certain IBDs with a large number of treatments effortlessly, we have to note a relationship between (1) an IBD and its *dual* and (2) a 2-replicate resolvable IBD and its *contraction*. An IBD is optimal if its dual (or contraction) is optimal (see Sects. 2.7 and 4.7 of John and Williams 1995). A *dual* of an IBD D of size (v, k, r) is an IBD D' of size $(v', k', r') = (b, r, k)$ obtained by swapping the treatments and blocks symbols in the original design. For example, the dual of the IBD of size $(v, k, r) = (4, 2, 3)$ displayed in section "▶Introduction" is the IBD of size $(v', k', r') = (6, 3, 2)$ displayed in section "▶A Criterion for Comparing IBDs".

Patterson and Williams (1976b) showed that a 2-replicate resolvable IBD D of size $(v, k, r) = (ks, k, 2)$ is uniquely determined by its contraction, a *symmetrical* IBD D^* of size $(v^*, k^*, r^*) = (s, k, k)$. The following 2-replicate resolvable IBD of size $(v, k, r) = (24, 4, 2)$ was obtained from a symmetrical IBD of size $(v^*, k^*, r^*) = (6, 4, 4)$ in section "▶Introduction":

0	4	8	12	16	20
1	5	9	13	17	21
2	6	10	14	18	22
3	7	11	15	19	23

3	7	5	0	1	2
4	11	10	8	9	6
15	12	18	13	17	14
16	19	20	21	22	23

Since the original designs in these two examples are optimal, their derived designs are also optimal. Additional examples on the use of these tools are given in Nguyen (1994) and Sects. 2.7 and 4.7 of John and Williams (1995).

Some Applications of IBDs

IBDs are related to several more complex combinatorial structures. As such they can be used to build these structures. The apparent application of IBDs is to use the blocks of an IBD as the column component of a row-column design (RCD). These designs are used for elimination of heterogeneity in two directions. Nguyen and Williams (1993) and Nguyen (1997) suggested a method for constructing optimal RCDs by permuting the treatments

within the blocks of an IBD used as the column component of the RCD. The following is an optimal RCD for 15 treatments, each replicated four times, set out in a 6 × 10 array obtained by permuting the treatments within the blocks of the IBD of size $(v, k, r) = (15, 6, 4)$ in section ►"IBDs with High Efficiency Factors":

12	7	4	11	9	8	5	10	1	2
3	5	0	12	11	6	4	2	13	1
7	4	9	6	1	13	2	3	10	14
6	10	1	14	5	7	8	11	3	0
9	13	8	4	14	11	3	0	12	7
5	0	6	10	13	2	14	9	8	12

This optimal RCD with $E = 0.856$ has been recommended for a taste test experiment involving 15 food products. The columns represent the tasters and the rows represent the order in which the products are introduced to the tasters.

Supersaturated designs are designs in which the number of factors $m > n - 1$ where n is the number of runs. Nguyen (1996) and Liu and Zhang (2000) described a method of constructing optimal 2-level supersaturated designs from cyclic BIBDs. The following optimal supersaturated design for 14 factors in eight runs was obtained from the cyclic BIBD in section "►IBDs with High Efficiency Factors". Each factor (column) of this design corresponds to a block of this BIBD. The treatments in each block of this BIBD are used to allocate the high level of a factor to a run:

1	1	1	1	1	1	1	1	1	1	1	1	1	1
−1	1	−1	−1	−1	1	1	−1	−1	−1	1	−1	1	1
1	−1	1	−1	−1	−1	1	1	−1	−1	−1	1	−1	1
1	1	−1	1	−1	−1	−1	1	1	−1	−1	−1	1	−1
−1	1	1	−1	1	−1	−1	−1	1	1	−1	−1	−1	1
−1	−1	1	1	−1	1	−1	1	−1	1	1	−1	−1	−1
−1	−1	−1	1	1	−1	1	−1	1	−1	1	1	−1	−1
1	−1	−1	−1	1	1	−1	−1	−1	1	−1	1	1	−1

IBDs have also been used to construct 3-level response surface designs (Box and Behnken 1960 and Nguyen and Borkowski 2008) and orthogonal and near-orthogonal arrays (Nguyen and Liu 2008). Other novel applications of IBDs can be found in Chapts. 5 and 8 of Raghavarao and Padgett (2005).

Cross References

►Experimental Design: An Introduction
►Optimum Experimental Design
►Research Designs

References and Further Reading

Box GEP, Behnken DW (1960) Some new three level designs for the study of quantitative variables. Technometrics 2:455–477

Clatworthy WH (1973) Tables of two-associates-class partially balanced designs. Appl Math Ser 63. National Bureau of Standards, Washington

John JA, Mitchell TJ (1977) Optimal incomplete block designs. J R Stat Soc B 39:39–43

John JA, Williams ER (1995) Cyclic designs and computer-generated designs. Chapman and Hall, New York, NY

Liu MQ, Zhang RC (2000) Construction of $E(s^2)$ optimal supersaturated designs using cyclic BIBDs. J Stat Plann Infer 91:139–150

Nguyen N-K (1994) Construction of optimal block designs by computer. Technometrics 36:300–307

Nguyen N-K (1996) An algorithmic approach to constructing supersaturated designs. Technometrics 38:69–73

Nguyen N-K (1997) Construction of optimal row-column designs by computer. Comput Sci Stat 28:471–475

Nguyen N-K, Borkowski JJ (2008) New 3-level response surface designs constructed from incomplete block designs. J Stat Plann Infer 138:294–305

Nguyen N-K, Liu MQ (2008) Orthogonal and near-orthogonal arrays constructed from incomplete block designs. Comp Stat Data Anal 52:5269–5276

Nguyen N-K, Williams ER (1993) An algorithm for constructing optimal resolvable row-column designs. Aust J Stat 35:363–370

Patterson HD, Williams ER (1976a) A new class of resolvable incomplete block designs. Biometrika 63:83–92

Patterson HD, Williams ER (1976b) Some theoretical results on general block designs. In Proceedings of the 5th British Combinatorial Conference. Congressus Numeratium XV, Utilitas Mathematica, Winnipeg, pp 489–496

Raghavarao D, Padgett LV (2005) Block designs: analysis, combinatorics and applications. World Scientific, Singapore

Incomplete Data in Clinical and Epidemiological Studies

Geert Molenberghs
Professor
Universiteit Hasselt and Katholieke Universiteit Leuven, Leuven, Belgium

In many longitudinal and multivariate settings, not all measurements planned are taken in actual practice. It is important to reflect on the nature and implications of such

incompleteness, or missingness, and properly accommodate it in the modeling process.

When referring to the missing-value process we will use terminology of Little and Rubin (2002, Chap. 6). A non-response process is said to be *missing completely at random* (MCAR) if the missingness is independent of both unobserved and observed data and *missing at random* (MAR) if, conditional on the observed data, the missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR).

Given MAR, a valid analysis that ignores the missing value mechanism can be obtained, within a likelihood or Bayesian framework, provided the parameters describing the measurement process are functionally independent of the missingness model parameters, the so-called parameter distinctness condition. This situation is termed ignorable by Rubin (1976) and Little and Rubin (2002) and leads to considerable simplification in the analysis (Verbeke and Molenberghs 2000). There is a strong trend, nowadays, to prefer this kind of analyses, in the likelihood context also termed *direct-likelihood* analysis, over *ad hoc* methods such as *last observation carried forward* (LOCF), *complete case analysis* (CC), or simple forms of ►imputation (Molenberghs and Kenward 2007). Practically, it means conventional tools for longitudinal and multivariate data, such as the linear and generalized linear mixed-effects models (Verbeke and Molenberghs 2000; Molenberghs and Verbeke 2005) can be used in exactly the same way as with complete data. Software tools like the SAS procedures MIXED, NLMIXED, and GLIMMIX facilitate this paradigm shift.

In spite of direct likelihood's elegance, fundamental model assessment and model selection issues remain. Such issues, occurring under MAR and even more under MNAR, are the central theme of this paper.

Indeed, one can never fully rule out MNAR, in which case the missingness mechanism needs to be modeled alongside the mechanism generating the responses. In the light of this, one approach could be to estimate from the available data the parameters of a model representing a MNAR mechanism. It is typically difficult to justify the particular choice of missingness model, and the data do not necessarily contain information on the parameters of the particular model chosen (Molenberghs and Kenward 2007). For example, different MNAR models may fit the observed data equally well, but have quite different implications for the unobserved measurements, and hence for the conclusions to be drawn from the respective analyses. Without additional information one can only distinguish between such models using their fit to the observed data, and so goodness-of-fit tools typically do not provide a relevant means of choosing between such models. It follows that there is an important role for sensitivity analysis in assessing inferences from incomplete data (Verbeke and Molenberghs 2000; Molenberghs and Verbeke 2005; and Molenberghs and Kenward 2007).

About the Author

For biography *see* the entry ►Linear Mixed Models.

Cross References

►Clinical Trials: An Overview
►Imputation
►Statistical Methods in Epidemiology

References and Further Reading

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
Molenberghs G, Kenward MG (2007) Missing data in clinical studies. Wiley, Chichester
Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. Springer, New York
Rubin DB (1976) Inference and missing data. Biometrika 63:581–592
Tan MT, Tian G-L, Ng KW (2010) Bayesian missing data problems. Chapman and Hall/CRC, Boca Raton
Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer, New York

Index Numbers

Dimitri Sanga
Acting Director
African Centre for Statistics, Addis Ababa, Ethiopia

Definition

An index number can be defined as a single indicator representing the change in the value of a variable relative to its value at some base date or state referred to as the base period. The index is often conventionally scaled so that its base value is 100. The variables considered represent a number of concepts including prices, quantity, volumes, value of a commodity, or other general economic variable such as national income, or gross output, cost of living, value of stock exchange etc. It constitutes a convenient way to standardize the measurement of numbers so that they are directly comparable.

Index numbers are used in several instances. The most commonly used include price indexes, quantity indexes, value indexes, or special-purpose indexes etc. Some of

the most widely known and used indexes include the Consumer Price Index (CPI), the Producer Price Index (PPI), the Human Development Index (HDI), etc. The CPI describes the change in prices of a basket of goods purchased by a representative consumer relative to a base period while the PPI is its equivalent but on the producer side. Stock market indexes, on the other hand report on the change in the prices of stocks on different markets. These include the Dow Jones Industrial Average, which is published daily, and that describes the overall change in common stock prices of 30 large companies during the day, while the Standard and Poor's 500 Stock Average is based on a 500 most important firms which stocks trade on the New York Stock Exchange divided by a factor that is adjusted for stock splits. There are equivalent of these indexes in other countries and cities such as the CAC 40 in Paris, the DAX in Frankfurt, the FTSE 100 in London, and the Nikkei 225 in Tokyo.

Since the development of indices has been dominated by price indices and that many of the developments do also apply to other types of indexes, the remaining of the text will concentrate and use examples of these kinds of indexes.

Types of Indexes

When the measurements over time and/or space are on a single variable, for example, the price of a certain commodity, the index is called a simple index number. Thus, the index of a variable for any year t is defined as:

$$I_t = \frac{X_t}{X_0} \times 100,$$

where I is the index number at period t, X_t and X_0 being the values of the variable at time t and base period respectively. For example, if a commodity costs twice as much in 2010 as it did in 2000, its index number would be 200 relative to 2000.

When the measurements over time and/or space are on the multiple aspects of a concept such as the level of economic development, general disparities, or for two or more items, the index is called composite index. The index related to the concept for any year t, in this case, is defined as:

$$I_t = f\left(\omega_i, \frac{X_t^i}{X_0^i}\right),$$

where I is the index number at period t, X_t and X_0 the values of the concept at time t and base period respectively, f a functional form (might be a product or sum), and ω_i the weight of component (aspect) i. The well-known Laspeyres price index is a special case of the above with the weights being values (prices time quantities) of different products (goods or services) in the basket during the base period as follows: $L_P = \omega_i \sum_{i=1}^{n} \frac{P_t^i}{P_0^i}$, with $\omega_i = \frac{P_0^i Q_0^i}{\sum_{i=1}^{n} P_0^i Q_0^i}$ and P the prices and Q quantities at respective periods. The Laspeyres price index is therefore a weighted average of the relative prices of different goods being part of a basket between periods 0 and t.

Elementary Indexes

Elementary indexes also called unweighted indexes are those that compare the prices in different periods without using weights. They are called elementary because they are computed for a single good (product, item, concept etc.). In fact, the computation of price indexes for a single good does not require the use of weights since only one type of good is being aggregated. Below are some well-known elementary price indices.

The Dutot Index

Developed by the French economist Charles de Ferrare Dutot in 1738, this index is a ratio of average prices as follows:

$$D_P = \frac{\frac{1}{n}\sum_{i=1}^{n} P_t^i}{\frac{1}{n}\sum_{i=1}^{n} P_0^i} = \frac{\sum_{i=1}^{n} P_t^i}{\sum_{i=1}^{n} P_0^i}.$$

The Carli Index

Developed by the Italian economist Rinaldo Carli in 1764, this index is an arithmetic average of price ratios as follows:

$$C_P = \frac{1}{n}\sum_{i=1}^{n} \frac{P_t^i}{P_0^i}.$$

The Jevons Index

Developed by the English economist Stanley Jevons in 1738, this index is a geometric average of prices as follows:

$$J_P = \prod_{i=1}^{n} \left(\frac{P_t^i}{P_0^i}\right)^{1/n}.$$

Because in some instances, prices or price ratios need to be aggregated to derive price indices, arithmetic or geometric means (see ▶Geometric Mean) are used. The debate over the correct method for computing price indexes namely on the use of arithmetic or geometric averages and whether to weight the index and by which quantities goes back as far as the 1800s. Three pioneers on price indexes development exchanged a lot on this. Laspeyres

argued for an arithmetic average weighted by quantities in the first period even if he did not use that himself in practice. Paasche was of the same opinion as Laspeyres on the use of arithmetic means but differed on the weights as he privileged the use of weights from the current period as opposed to Laspeyres who used those of the base period. Unlike the other two economists, Jevons was defending geometric averaging. The debates on these issues still exist nowadays and have been studied for centuries.

To illustrate the debate on arithmetic versus geometric means, let's consider the prices of two goods A and B from period 0 to t. If the price of good A doubles between 0 to t, the index will rise from 100 to 200. If in the meantime the price of good B decreases by half during the same period, the related index will decrease from 100 to 50. The average prices level of the two goods in t will be 125 resulting in an average price change of 25%. Using geometric means, the average price change (square root of the product of the indexes in t) will be 100, meaning no change in average prices. As can be observed from this example, the choice of the aggregation method will influence the index. The geometric means assure that expenditures shares are constant. The argument against arithmetic averages and in favor of geometric ones can be summarized in the fact that buyers substitute towards those goods whose relative price has fallen. The geometric means takes into account substitutions holding expenditures shares constants while the arithmetic means assumes that quantities remain constants.

Fixed-Weights Indexes

Fixed-weights indexes are those indexes that use weights derived for a given period in their calculation. In practice, the computation of price indexes entails only collecting prices in the current period (t) as the indices are explained as a function of the ratios of prices of the items between the current and the base periods. These ratios are then aggregated using either arithmetic or geometric means. Once this decided upon, there is still the issue of the weights to be used. These weights can be computed either for the base period or the current one. The Laspeyres index commonly used to compute the CPI in official statistics of countries across the world uses weights from the base period while the Paasche index uses weights from the current period. While both are fixed-weights indices, the Laspeyres is very attractive in practice because at t the weights in period 0 (the base) can be derived from the expenditures from household budget surveys as they have already been observed in the past. Below are some widely used fixed-weights indexes.

The Laspeyres Index

Developed by the German economist Ernst Louis Étienne Laspeyres in 1871, this index is a weighted arithmetic average of price ratios with weights derived from the base period as follows:

$$L_P = \frac{\sum_{i=1}^{n} P_t^i \, Q_0^i}{\sum_{i=1}^{n} P_0^i \, Q_0^i} = \omega_i \sum_{i=1}^{n} \frac{P_t^i}{P_0^i}, \text{ with } \omega_i = \frac{P_0^i \, Q_0^i}{\sum_{i=1}^{n} P_0^i \, Q_0^i}.$$

The Paasche Index

Developed by the German Statistician/economist Hermann Paasche in 1874, this index is a weighted arithmetic average of price ratios with weights derived from the current (t) period as follows:

$$P_P = \frac{\sum_{i=1}^{n} P_t^i \, Q_t^i}{\sum_{i=1}^{n} P_0^i \, Q_t^i}.$$

By virtue of the use of the base period weights, the Laspeyres index is known to overstate price changes. In fact, according to economic theory, consumers substitute goods that are becoming more expensive with less expensive ones while the index assumes that the basket of goods and services chosen in the base period remains fixed. This index ends up using an outdated fixed structure that does not take into accounts the substitution effect. The Paasche index understates price changes for the same reason. To deal with this problem, Fisher came up with a proposal that lies between the two previous ones.

Chained Indexes

To overcome the problem related to the use of outdated fixed weights structures, indexes are often chained using updated weights. In the case of the Laspeyres price index, its chained version will take the following form:

$$LC_P = \frac{\sum_{i=1}^{n} P_1^i \, Q_0^i}{\sum_{i=1}^{n} P_0^i \, Q_0^i} \cdot \frac{\sum_{i=1}^{n} P_2^i \, Q_1^i}{\sum_{i=1}^{n} P_1^i \, Q_1^i} \cdot \ldots \cdot \frac{\sum_{i=1}^{n} P_t^i \, Q_{t-1}^i}{\sum_{i=1}^{n} P_{t-1}^i \, Q_{t-1}^i} \cdot \ldots \cdot \frac{\sum_{i=1}^{n} P_s^i \, Q_{s-1}^i}{\sum_{i=1}^{n} P_{s-1}^i \, Q_{s-1}^i}$$

where s is the number of periods over which the chain index extends. This has been and remains the practice in the computation of official CPIs of many of the countries across the world.

The Fisher Index

Developed by the American economist Irving Fisher in 1920/1922, this index is a geometric mean of Laspeyres and

Paasche indexes as follows:

$$F_P = (L_P \cdot P_P)^{1/2}.$$

Non-Fixed-Weights Indexes

The non-fixed weights indexes include the following:

The Marshall-Edgeworth Index

Developed by Alfred Marshall (1887) and Edgeworth (1925), this index is a weighted relative of current prices with weights being arithmetic averages of current and base period quantities as follows:

$$ME_P = \frac{\sum_{i=1}^{n}\left[P_t^i \cdot \frac{1}{2} \cdot \left(Q_0^i + Q_t^i\right)\right]}{\sum_{i=1}^{n}\left[P_0^i \cdot \frac{1}{2} \cdot \left(Q_0^i + Q_t^i\right)\right]}.$$

This formulation has however a major drawback as it can be problematic when comparing the price levels of a small entity versus a large one as the quantities of the large entity might dominate those of the small.

Superlative Indexes

Superlative indexes provide close approximations of the true cost of living index. They produce similar results and constitute an exact approximation for a flexible functional form that can provide a second order approximation to other twice differentiable functions around the same point (Diewert 1976).

The Fisher index is a superlative index and is also called "Fisher Ideal Price Index." Another superlative index is the Tornqvist Index. Beside the fact that the Fisher index is superior theoretically to the Laspeyres and the Paasche, it has a number of desirable properties from the National Accounts perspective. In fact, it is reversible over time that is the index showing the change between period 0 and t is the reciprocal of the index showing the change between period t and 0. Moreover, it also has the property of reversibility of factors by which the product of the price and quantity indexes is equal to the change in current values. The other indexes do not have these properties.

The Tornqvist Index

Developed by Tornqvist in 1936, this index is a weighted geometric mean of price ratios with weights being average expenditures shares on each good as follows:

$$T_P = \prod_{i=1}^{n}\left(\frac{P_t^i}{P_0^i}\right)^{\frac{1}{2}\left[\frac{P_0^i \cdot Q_0^i}{\sum_{i=1}^{n} P_0^i \cdot Q_0^i} + \frac{P_t^i \cdot Q_t^i}{\sum_{i=1}^{n} P_t^i \cdot Q_t^i}\right]}.$$

Other Composite Indexes

A composite index is a comprehensive single number representing a vast array of measurements on the multiple aspects of a conceptual entity such as general price level, cost of living, level of economic development, general disparities, statistical development etc. Representatives of these kinds of indices are the HDI, the General Index of Development (GID), Physical Quality of Life Index (PQLI), Index of Social Progress (ISP) etc.

There are several advantages related to the development of composite indexes. The advantages of composite indices include: excellent communication tools for use with practically any constituency including the media, general public, and decision makers; the provision of single targets that facilitate the focus of attention; facilitation of the necessary negotiations about practical value and usefulness due to simplicity; provision of a means to simplify complex, multidimensional phenomena and measures; easy measure and visual representation of overall trends in several distinct indicators over time or across space; and provision of a means to compare diverse phenomena and assessing their relative importance, status or standing on the basis of some common scale of measurement across time and space. On the other hand, there are also some disadvantages. These include sending misleading policy messages if composite indices are poorly constructed or misinterpreted; possibility of inviting simplistic policy conclusions; misused to support desired policy, if the construction process is not transparent and lacks sound statistical or conceptual principles; selection of indicators and weights could be the target of political challenge; and leading to inappropriate policies of dimensions of performance that are difficult to measure.

The bottom line of the relevance of composite indicators is that they are needed as a starting point for initiating discussions and attracting public interest to the phenomenon at stake as they provide a very simple and useful way of presenting complex multidimensional phenomena into an easily understandable measure. Nevertheless, one as to be cautious in its development as the index is meant to be used for important decision-making and expression of views on the considered phenomenon. Therefore, constituencies affected by its use should ascertain their relevance.

Areas of Research

Several issues on index numbers theory and practice are still worth exploring for future research. These include the use of geometric versus arithmetic means, the use or not of weights and from which time, approximation of the cost

of living index, aggregation formulas, change in quality, introduction of new goods etc.

There is a wealth of literature on index construction and applications, desirable properties of index numbers and the relationship between index numbers and economic theory. For further details concerning the discussed matter, the reader may refer, *inter alia*, to the below references.

Acknowledgments

Disclaimer: The views expressed in this paper are personal to the author and do not necessarily represent those of the United Nations Economic Commission for Africa or its subsidiary organs.

About the Author

For biography *see* the entry ►Role of Statistics: Developing Country Perspective.

Cross References

►Business Statistics
►Economic Statistics
►National Account Statistics

References and Further Reading

Booysen F (2002) An overview and evaluation of composite indices of development. Soc Indic Res 59:115–151. Netherlands

Deaton A (1998) Getting prices right: what should be done? J Econ Perspect 12(1):37–46. Winter

Diewert E (1976) Exact and superlative index numbers. J Econ 46:115–145

Diewert E (1978) Superlative index numbers and consistency in aggregation. Econometrica 46(4):883–900

Diewert E (1987) Index numbers. In: Eatwell J, Milgate M, Newman P (eds) The new palgrave: a dictionary of economics, vol 1. MacMillan, London, pp 767–780

Diewert E (1998) Index number issues in the consumer price index. J Econ Perspect 12(1):47–58. Winter

Drewnowski J (1972) Social indicators and welfare measurement: remarks on methodology. J Dev Stud 8:77–90

Edgeworth FY (1925) The plurality of index-numbers. The Economic Journal, vol 35(139) pp 379–388

Morris MD (1979) Measuring the conditions of world poor: the physical quality of life index. Perganon Policy Studies, p 42. Perganon Press, New York, pp 20–56

Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffman A, Giovannini E (2005) Handbook on constructing composite indicators: methodology and user guide. OECD Statistics Working Papers, OECD, Paris

Salzman J (2003) Methodological choices encountered in the construction of composite indices of economic and social well-being. Centre for the study of Living Standards, Ottawa, Ontario

Turvey R (2004) Consumer price index manual: theory and practice. International Labor Organization, Geneva, p 11

Industrial Statistics

Ursula Gather[1,2], Sonja Kuhnt[2], Thomas Mühlenstädt[2]
[1]Rector of TU Dortmund University, Dortmund, Germany
[2]Faculty of Statistics
TU Dortmund University, Dortmund, Germany

Industrial statistics deals with the assurance and improvement of quality in industrial (production-) processes and products.

Quality Assurance

One of the most important considerations regarding a production process is the assurance of a stable and steady quality of the process output, i.e., the process is *under control*. Otherwise, if the process shows erratic and undesired behavior, it is *out of control*. An underlying random variable Y measuring the quality of the process, is often assumed to have a distribution P_θ, with θ being a parameter(-vector), possibly mean and variance $\theta = (\mu, \sigma^2)$.

Acceptance Sampling

►Acceptance sampling aims at accepting or rejecting a lot of products by inspecting only a small proportion of the items (Kenett and Zacks 1998). Items are chosen according to an acceptance sampling scheme and rated as either conforming or nonconforming to given quality specifications. An important characterization of an acceptance sampling plan is given by the operating characteristic (OC) function, which yields the probability of accepting a lot with proportion p of defective items.

Control Charts

By observing series of samples $y_{m,1}, \ldots, y_{m,g}$, of size $g \in \mathbb{N}$, $m = 1, 2, 3, \ldots$, one wants to check if the process is under control. A control chart is a graphical tool which plots a summary statistic of the samples against the sample number (Montgomery 2009). To determine whether the process is out of control, control limits $((ucl, lcl))$ are calculated. If the control limits are exceeded the process is rated out of control and thus the reason for the deviation has to be investigated. The control limits are estimated from samples, for which it is known that the process was under control. The summary statistic used most frequently is the arithmetic mean: $\bar{y}_m := \frac{1}{g} \sum_{i=1}^{g} y_{m,i}$.

The performance of the control chart can also be described by the OC-function. Possible extensions of ►control charts are control charts with memory, multivariate control charts or control charts for the variance of the process.

Process Capability Indices

Process capability indices summarize the behavior of a process with a single number in order to have a simple decision rule to determine if the process performance can be accepted or not. Well known process capability indices are the C_p and the C_{pk} index:

$$C_p := \frac{ucl - lcl}{6\sigma}, \qquad C_{pk} := \min\left(\frac{\mu - lcl}{3\sigma}, \frac{ucl - \mu}{3\sigma}\right), \tag{1}$$

using the same notation as before. Here, the *ucl* and *lcl* are most frequently defined by the desired properties for the product. In practice, μ and σ are estimated by the arithmetic mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and the empirical standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$, and then substituted into the formulae. For the C_p index, the mean μ of the process is not included in the formula. This reflects the assumption that the position of a process is (often) easy to adjust, while the variation of the process is difficult to control and hence should be the only value to be used for judging a process. The C_{pk} index also considers the location of the process. For both indices, higher values represent a better process. For the C_p index, this means that the variation of the process compared to the range of the control limits is small, while for the C_{pk} the location of the process has to be inside the control limits as well. There are many possible extensions of the concept of process capability indices for situations like one sided process specifications or skew distributions of Y, e.g., see Kotz and Lovelace (1998).

Quality Improvement

If a new process is under investigation, good configurations of the process are wanted. Here the random variable Y representing the process depends on some covariates $x \in \mathbb{R}^k$. A configuration of x which optimizes Y in some suitable way is searched for. This can be maximizing/minimizing Y, or searching for an x which results in the smallest deviation of Y from a nominal value T. In industrial statistics, methods for this kind of problem are often summarized under the topic of ►response surface methodology (RSM, Myers et al. 2009). Formally, this usually results in a regression model:

$$Y(x) = f(x) + \varepsilon, \tag{2}$$

where $f(x)$ is an unknown function $f : \mathbb{R}^k \to \mathbb{R}$ and ε a random variable representing process variation with expectation $E(\varepsilon) = 0$ and constant variance $var(\varepsilon) = \sigma^2$. Given some data $y_1 = y(x_1), \ldots, y_n = y(x_n)$ with $x_1, \ldots, x_n \in \mathbb{R}^k$ an estimate $\hat{f}$ is calculated and used for optimizing Y.

Design of Experiments (DoE)

A powerful tool for improving process quality is called design of experiments, which refers to statistically planning experiments. DoE leads to a design matrix $X = [x_1, \ldots, x_n]'$ for n runs of the experiments to be conducted, where each row of X specifies the settings of the covariates for one run. The choice of X depends on a number of issues, e.g. the purpose and scope of the experiment, the assumed statistical model and the desired degree of precision of the results and conclusions. Standard designs such as fractional factorial designs and response surface designs exist, primarily for use in situations where the estimation of a regression model is desired.

Robust Parameter Design/Taguchi Methods

Taguchi (e.g., see Myers et al. 2009) was one of the first to consider not only the mean f of the process under investigation but also the variance of Y. It is assumed that there are some control variables $x \in \mathbb{R}^k$ which are easy to adjust in mass production settings but also noise variables $z \in \mathbb{R}^d$, which are adjustable in a laboratory but not during mass production, e.g., fluctuations in raw materials. In order to reduce the influence of noise factors on the process, a configuration for the control variables x is investigated, which not only optimizes the mean of Y but also results in a small variation of Y. As these two aims are not always achievable at the same time, strategies for finding compromises have to be applied. Crucial for achieving this is the use of suitable experimental designs like crossed designs, which combine a design for the control variables and a design for the noise variables.

Further Topics

Six Sigma

►Six Sigma is a methodology for implementing process improvements in companies. Applying Six Sigma implies using a specific organizational form. One of the most important aspects of Six Sigma is its focus on projects. A project should consist of a "potential breakthrough" for product or service improvement, Montgomery and Woodall (2008). Each project is strictly organized by the DMAIC cycle: Define, Measure, Analyze, Improve and Control, which makes research projects more traceable.

Another key aspect of Six Sigma is its belt system, dividing the educational status of an employee into three classes: green belt (lowest level), black belt or master black belt (highest level).

The name Six Sigma is taken from reliability theory. Consider a process with nominal value T and symmetrical control limits lcl, ucl. A Gaussian random variable Y with mean $\mu = T$ and variance $\sigma = (ucl - lcl)/6$ has a probability of 99.73 percent of being within the control limits. This reflects the claim of Six Sigma to achieve processes which have only a very small probability of being outside the control limits.

Reliability Analysis

Reliability analysis deals with the analysis of how reliably a product is performing its task (Kenett and Zacks 1998). A product which functions for a long time without any defects is said to be reliable. A central concept is the reliability function $R(t)$: $R(t) :=$ *Probability that the product is working according to its specifications after t time units.* In order to estimate the reliability function, a sample of products is investigated and the failure times are noted. As this can be very time consuming, accelerated life testing is often used, where the product is set under higher stress than under normal conditions. Reliability analysis is not only applied to products but also, for example, to software.

Computer Experiments

In many industrial research situations, a real world process can be replaced by simulation models, which reproduce the real process in a software environment. As a result, huge cost reductions can be achieved. However, care has to be taken, as computer experiments are very different from conventional experiments. Most of the time, the computer experiment is deterministic and a single run is very time consuming. Thus the computer experiment should be planned carefully (Fang et al. 2006). Designs for computer experiments are summarized as space filling designs while models for a computer experiment should interpolate the observations as no random error is observed. Furthermore, validating the computer experiment as replacement for the real world experiment is crucial in order to assure that valid results for real world applications are derived from the computer experiment.

About the Authors

Professor Gather was appointed rector of TU Dortmund University on 1 September 2008. She is vice-chair of the Senate of the German Aerospace Center, council member of the International Statistical Institute, and spokesperson of the DFG-Review Board Mathematics. Professor Gather was Editor of Metrika (1996–2006).

Dr. Sonja Kuhnt has a temporary professorship of Mathematical Statistics and Industrial Applications in the Statistics Department at TU Dortmund University.

Cross References

►Acceptance Sampling
►Control Charts
►Economic Statistics
►Six Sigma

References and Further Reading

Fang K-T, Li R, Sudjianto A (2006) Design and modeling for computer experiments. Computer Science and Data Analysis Series. Chapman and Hall/CRC, Boca Raton, FL

Kenett R, Zacks S (1998) Modern industrial statistics. Duxbury, Pacific Grove, CA

Kotz A, Lovelace C (1998) Process capability indices in theory and practice. Arnold, London

Montgomery D (2009) Statistical quality control: a modern introduction. Wiley series in probability and statistics. Wiley, Hoboken, NJ

Montgomery D, Woodall W (2008) An overview of six sigma. Int Stat Rev 76(3):329–346

Myers R, Montgomery D, Andersion-Cook C (2009) Response surface methodology – process and product optimization using designed experiments. Wiley series in probability and statistics. Wiley, Hoboken, NJ

Inference Under Informative Probability Sampling

Michail Sverchkov
Bureau of Labor Statistics, Washington, DC, USA

One of the outstanding features of sample surveys is that the samples are often drawn with unequal probabilities, at least at one stage of the sampling process. The selection probabilities are generally known for the sampled units in the form of the sampling weights (inverse of the sampling probability and possibly adjusted for nonresponse or calibration). The sampling weights are in common use for randomization-based inference on finite population quantities of interest, by weighting the sample observations by the corresponding sampling weights. This is discussed and illustrated in every text book on sample surveys. In this paper we focus on model-based inference.

In what follows we distinguish between the model holding for the population outcomes, hereafter the *population model*, and the (conditional) *sample model* holding for the sample outcomes. When the selection probabilities are correlated with the outcome values even after conditioning on the model covariates, the sampling is *informative*

and the two models can be very different, in which case the sampling process cannot be ignored in the modeling process. To see this, denote the population model by $f_U(y|x)$, where y is the outcome variable and x is a set of covariates. Following Pfeffermann et al. (1998), the sample model is defined as

$$f_s(y_i|x_i) \stackrel{def}{=} f(y_i|x_i, i \in s) = \frac{\Pr(i \in s|y_i, x_i) f_U(y_i|x_i)}{\Pr(i \in s|x_i)} = \frac{E_U(\pi_i|y_i, x_i) f_U(y_i|x_i)}{E_U(\pi_i|x_i)}, \quad (1)$$

where $\pi_i = \Pr(i \in s)$ is the sample inclusion probability and $E_U(\cdot)$ defines the expectation under the population model. Note that $\Pr(i \in s|y_i, x_i)$ is generally not the same as π_i, which may depend on all the population values $\{y_i, x_i\}, i \in U$ and possibly also on the values z_i of design variables z, which are not included in the model but are used for the sample selection. By (1), if $\Pr(i \in s|y_i, x_i) \neq \Pr(i \in s|x_i)$, the population and the sample models are different and fitting the population model to the sample data ignoring the sampling process may bias the inference very severely. In the rest of this paper we review approaches that account for the sampling effects under informative probability sampling. See Pfeffermann and Sverchkov (2009) for a more comprehensive discussion with examples.

The first approach utilizes the fact that if for every x, $\Pr(i \in s|y_i, x_i) = \Pr(i \in s|x_i) \forall y_i$, the population and the sample models are the same. Therefore the sampling effects can be accounted for by including among the model covariates all the design variables and interactions that are related to the outcome values and affect the sample selection probabilities. See, e.g., Gelman (2007). However, this paradigm is not always practical because there may be too many variables to include in the model and some or all of them may not be known or accessible to the modeler. Notice also that by including these variables among the model covariates, the resulting model may no longer be of scientific interest, requiring integrating them out of the model at a later stage, which can be complicated and not always feasible.

Alternatively, one could include in the model the sampling weights as surrogates for the design variables as proposed by Rubin (1985). The use of his strategy may again distort the interpretation of the model requiring therefore integrating out the sampling weights at a second stage. This is a feasible procedure for estimating the sample model because the integration is then with respect to the conditional sample distribution of the sampled weights given the covariates, which can be assessed from the observed weights and the covariates. However, for estimating the population model the integration of the sampling weights must be with respect to the population model of the weights given the covariates. Notice also that the vector of sampling weights may not be an adequate summary of all the design variables used for the sample selection.

A third approach, and the one in common use, is to estimate the population model by weighting the sample observations in the model-based estimating equations by the sampling weights. When the estimating equations are defined by the score function, the use of this approach is known in the sampling literature as "pseudo likelihood" estimation. The use of this approach is limited, however, mostly to point estimation, and probabilistic inference such as the construction of confidence intervals or the application of hypothesis testing generally require large samples normality assumptions. The inference is based on the randomization distribution and as such, it does not permit conditioning on the selected sample, for example, conditioning on the observed covariates, or on the selected clusters in a multi-level model. In addition, the use of this approach does not lend itself to prediction problems other than the prediction of the finite population quantities from which the sample is taken, and the estimators often have large variances, depending on the dispersion of the weights. See Pfeffermann and Sverchkov (2009) for further discussion on the use of this approach with many references.

A fourth approach is based on the relationship between the population model and the sample model in (1). Following Pfeffermann and Sverchkov (1999),

$$f_U(y_i|x_i) = \frac{E_s(w_i|y_i, x_i) f_s(y_i|x_i)}{E_s(w_i|x_i)}, \quad (2)$$

where $w_i = 1/\pi_i$ and $E_s(\cdot)$ is the expectation under the sample model. Thus, one can identify and estimate the population model by fitting the sample model $f_s(y_i|x_i)$ to the sample data and estimating the expectation $E_s(w_i|y_i, x_i)$, again using the sample data. For example, suppose that the population model is governed by the vector parameter $\theta = (\theta_0, \theta_1, \ldots, \theta_k)'$ and let $\mathbf{d}_{Ui} = (d_{Ui,0}, d_{Ui,1}, \ldots, d_{Ui,k})' = \partial \log f_U(y_i|x_i; \theta)/\partial\theta$ be the i^{th} score function. Assuming that the conditional expectations $E_s(w_i|y_i, x_i)$ are known or have been estimated and that the expectations $E_s(w_i|x_i) = \int_y E_s(w_i|y, x_i) f_s(y|x_i; \theta) dy$ are differentiable with respect to θ, it follows from (2) that if the sample outcomes are independent (see Remark 1 below), the sample likelihood equations are,

$$W_s(\theta) = \sum_{i \in s} E_s\{[\partial \log f_s(y_i|x_i; \theta)/\partial\theta]|x_i\} = \sum_{i \in s} E_s\{[\mathbf{d}_{Ui} + \partial \log E_s(w_i|x_i)/\partial\theta]|x_i\} = 0. \quad (3)$$

Therefore, θ can be estimated by solving the equations in (3).

Remark 1. Pfeffermann et al. (1998) showed that under some general regularity conditions if the population measurements are independent, the sample measurements are "asymptotically independent" with respect to the sample model. The asymptotic framework requires that the population size increases but the sample size stays fixed. The result is shown to hold for many sampling schemes with unequal probabilities in common use.

Instead of basing the likelihood on the sample distribution, one could use instead the "full likelihood" based on the joint distribution of the sample outcomes and the sample membership indicators $I_i = 1(0)$ for $i \in s (i \notin s)$. Let $\mathrm{I} = (I_1, \ldots, I_N)'$ and denote $\mathrm{x} = \{x_i, i \in U\}$. Then,

$$f(\mathrm{I}, y_s|\mathrm{x}) = \prod_{i \in s} \Pr(i \in s|y_i, x_i) f_U(y_i|x_i) \prod_{j \notin s} [1 - \Pr(j \in s|x_j)], \tag{4}$$

where $\Pr(i \in s|x_i) = \int \Pr(i \in s|y_i, x_i) f_U(y_i|x_i) \mathrm{d}y_i$; see, e.g., Gelman et al. (2003), Pfeffermann and Sverchkov (2003) and Little (2004). The use of (4) has the theoretical advantage of employing the information on the sample selection probabilities for units outside the sample, but it requires knowledge of the covariates for every unit in the population, unlike the use of (3). Other estimation approaches are considered in Breckling et al. (1994) and Pfeffermann and Sverchkov (2003).

So far we considered model estimation but the sample distribution enables also to predict missing population values. For this we define the *sample-complement* model,

$$f_c(y_i|x_i) \stackrel{def}{=} f(y_i|x_i, i \notin s) = \frac{\Pr(i \notin s|y_i, x_i) f_U(y_i|x_i)}{\Pr(i \notin s|x_i)} = \ldots = \frac{E_s[(w_i - 1)|y_i, x_i] f_s(y_i|x_i)}{E_s[(w_i - 1)|x_i]}, \tag{5}$$

with the last equation obtained in Sverchkov and Pfeffermann (2004). The sample-complement model is again a function of the sample model $f_s(y_i|x_i)$ and the expectation $E_s(w_i|y_i, x_i)$, and thus can be estimated from the sample.

Remark 2. When predicting the outcome value for a specific nonsampled unit (say, a unit with a given set of covariates), or the mean of a given nonsampled area in a small area estimation problem, and the sampling process is informative, there seems to be no alternative but to base the prediction on the sample-complement model. Classical randomization based inference is suited for estimating parameters of the finite population from which the sample is drawn, but not for prediction problems. Sverchkov and Pfeffermann (2004) illustrate how many of the classical randomization-based estimators of finite population totals, such as the ►Horvitz-Thompson estimator, Hajek/Brewer estimator and the GREG, are obtained as special cases of the application of the sample-complement model. The authors develop also a method for estimating the MSE of the prediction errors. Small area estimation under informative sampling of areas and within the areas is considered in Pfeffermann and Sverchkov (2007).

About the Author

Disclaimer: The opinions expressed in this paper are those of the author and do not necessarily represent the policies of the Bureau of Labor Statistics.

Cross References

►Sample Survey Methods
►Small Area Estimation

References and Further Reading

Breckling JU, Chambers RL, Dorfman AH, Tam SM, Welsh AH (1994) Maximum likelihood inference from sample survey data. Int Stat Rev 62:349–363

Gelman A (2007) Struggles with survey weighting and regression modeling (with discussion). Stat Sci 22:153–164

Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian data analysis, 2nd edn. CRC, London

Little RJ (2004) To model or not to model? Competing modes of inference for finite population sampling. J Am Stat Assoc 99:546–556

Pfeffermann D, Sverchkov M (1999) Parametric and semi-parametric estimation of regression models fitted to survey data. Sankhya B, 61:166–186

Pfeffermann D, Sverchkov M (2003) Fitting generalized linear models under informative probability sampling. In: Skinner C, Chambers R (eds) Analysis of survey Data. Wiley, New York, pp 175–195

Pfeffermann D, Sverchkov M (2007) Small area estimation under informative probability sampling of areas and within the selected areas. J Am Stat Assoc 102:1427–1439

Pfeffermann D, Sverchkov M (2009) Inference under Informative Sampling. In: Pfeffermann D, Rao CR (eds) Handbook of statistics 29B; sample surveys: inference and analysis. North Holland, Amsterdam, pp 455–487

Pfeffermann D, Krieger AM, Rinott Y (1998) Parametric distributions of complex survey data under informative probability sampling. Stat Sinica 8:1087–1114

Pfeffermann D, Moura FAS, Nascimento-Silva PL (2006) Multilevel modeling under informative sampling. Biometrika 93:943–959

Rubin DB (1985) The use of propensity scores in applied Bayesian inference. In: Bernardo JM, Degroot MH, Lindley DV, Smith AFM (eds) Bayesian statistics 2, Elsevier Science BV, Amsterdam, pp 463–472

Sverchkov M, Pfeffermann D (2004) Prediction of finite population totals based on the sample distribution. Surv Methodol 30: 79–92

Influential Observations

Deng-Yuan Huang
Professor
Fu-Jen Catholic University, Taipei, Taiwan

Fitting liner regression models usually uses the ▶least squares method. The fitted model may be largely influenced by a few observations. These observations are called influential observations. It is necessary to define a criterion to find out these observations. They may include important information.

The analysis of residuals may reveal various functional forms to be suitable for the regression model. Some appropriate criteria to measure the influence of the model were studied for detecting influential data.

We consider the following linear model:

$$\underline{Y} = X\underline{\beta} + \underline{e}, \tag{1}$$

where $E(\underline{e}) = \underline{0}$, $Var(\underline{e}) = \sigma^2 I_n$, and I_n denotes the identity matrix of order n, $\underline{Y}$ is an $n \times 1$ vector of responses, X is an $n \times p$ $(n > p)$ matrix of known constants of rank p, and $\underline{\beta}$ is a $p \times 1$ parameter vector.

Several authors have studied the influence on the fitted regression line when the data are deleted. Let $\hat{\underline{\beta}}$ be the usual least squares estimator of $\underline{\beta}$ based on the full data and let $\hat{\underline{\beta}}_A$ be an alternative least squares estimator based on a subset of the data. The empirical influence function for $\hat{\underline{\beta}}$, namely, IF_A, is defined to be

$$IF_A = \hat{\underline{\beta}} - \underline{\beta}. \tag{2}$$

For a given positive definite matrix M and a nonzero scale factor c, Cook and Weisberg (1980) defined the distance $D_A(M, c)$ between $\hat{\underline{\beta}}$ and $\hat{\underline{\beta}}_A$ as follows:

$$D_A(M, c) = (IF_A)'M(IF_A)/c. \tag{3}$$

They suggested that the matrix M can be chosen to reflect specific interests. They also pointed out that in some applications, measuring the influence of cases on the fitted values, $\hat{\underline{Y}} = X\hat{\underline{\beta}}$, may be more appropriate than measuring the influence on $\hat{\underline{\beta}}$. They mentioned an example to describe the fact that if prediction is the primary goal, it may be convenient to work with a reparameterized model where the regression coefficients are not of interest. They tried to treat their measurement of the influence on the fitted values $X\hat{\underline{\beta}}$ and used the empirical influence function for $\hat{\underline{Y}}$, denoted by $X(IF_A)$.

Welsch (1982) pointed out that in an earlier paper, Cook (1977) chose to measure influence by

$$D_i = \frac{(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})'X'X(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})}{s^2 p}, \tag{4}$$

where s^2 is the residual mean square for full data and $\hat{\underline{\beta}}_{(i)}$ is the least squares estimator of $\underline{\beta}$ based on the data set with the ith component in $\underline{Y}$ deleted.

Welsch (1982) gave an example to explain that when all of the observations but one lie on a line, (4) can give potentially confusing information since it may indicate that some observations on the line are more influential than the one observation not on the line. This is counterintuitive since the deletion of this one observation leads to a perfect fit. Therefore, finding a more reasonable measurement is very important. We shall consider the case of one-at-a-time data deletion, since, for the case of deletion of a subset, computations can be similarly carried out (Cook and Weisberg 1980; Gray and Ling 1984).

Gupta and Huang (1996) derived a suitable choice of M and c in (3) to measure the influence and bias are derived as follows:

$$D_{(i)} = \frac{(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})'X'_{(i)}X_{(i)}(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})}{s^2_{(i)}p}, \tag{5}$$

where $X_{(i)}$ denotes the data set with the ith component in X deleted, and $s^2_{(i)}$ is the residual mean square for the ith component in $\underline{Y}$ deleted.

Gupta and Huang's statistic $D_{(i)}$ in (5) measures the influence on residuals and on $X\hat{\underline{\beta}}$. It should be pointed out that the large influence on $X\hat{\underline{\beta}}$ should have much influence on $\hat{\underline{\beta}}$ though the converse may not hold.

About the Author

For biography *see* the entry ▶Multiple Statistical Decision Theory.

Cross References

▶Cook's Distance
▶Regression Diagnostics
▶Residuals
▶Robust Regression Estimation in Generalized Linear Models
▶Simple Linear Regression

References and Further Reading

Cook RD (1977) Detection of influential observations in linear regression. Technometrics 19:15–18
Cook RD, Weisberg S (1980) Characterizations of an empirical influence function for detecting influential cases in regression. Technometrics 22:495–506

I

Daniel C, Wood FS (1980) Fitting equations to data, 2nd edn. Wiley, New York
Gray JB, Ling RF (1984) *K*-clustering as a detection tool for influential subsets in regression. Technometrics 26:305–318
Gupta SS, Huang DY (1996) On detecting influential data and selecting regression variables. J Stat Plann Infer 53:421–435
Welsch RE (1982) Influence functions and regression diagnostics. In: Launer RL, Siegel AF (eds) Modern data analysis Academic, New York

Information Theory and Statistics

Evgueni Haroutunian
Professor, Head of the Laboratory of Information Theory and Applied Statistics
Institute for Informatics and Automation Problems of the Armenian National Academy of Sciences, Yerevan, Armenia

Information theory is a branch of mathematics based on probability theory and statistical theory. The founder of information theory, Claude Shannon, created the "mathematical theory of communication" and the successors of Shannon thoroughly developed firm mathematical constructions for descriptions of communication processes ensuring data reliable compression, transmission and protection (Blahut 1974, 1987; Cover and Thomas 2006; Rissanen 2007).

Modern information theory is characterized as a "unifying theory with profound intersection with probability, statistics, computer science, and other fields" Verdú (1998), it has main applications in communication engineering, neurobiology, psychology, linguistics, electrical engineering, data analysis, and statistical inference.

What might statisticians learn from information theory? Basic concepts like ►entropy, mutual information, and ►Kullback-Leibler divergence (also called informational divergence, or relative entropy, or discrimination information), along with many various generalizations of them, certainly play an important role in statistics. The definitions of these notions are given in publications cited below Blahut (1987), Cover (2006), Csiszár (2004), Liese (2006), Pardo (2006). The elements of large deviations theory (see ►Large Deviations and Applications), limit theorems, hypothesis testing, estimation of parameters, some modern principles of statistical inference such as the maximum entropy principle, model selection methodologies like AIC and the principle of minimum description length, are explained with usage of information theory methodology in Ahlswede (1987), Kullback (1959), Pardo (2006), and Rissanen (2007).

Statistical theory shares with information theory the common optimal methods of usage of considered random data. Interaction of information theory and statistics creates the opportunity for formulating and solving of many interesting specific theoretical and practical problems Ahlswede (1986), Barron (1997), Rényi (1968). Information theoretic proofs have been given to various limit theorems of probability theory Kendall (1963), Linnik (1959).

A specific scientific field established by a series of works of information theory experts is the hypotheses testing with finite statistics Hellman (1970).

One of new directions of statistical studies initiated by information theorists is development of statistical inference, in particular of hypothesis testing, for models of two or many similar stochastic objects Ahlswede (2005), Haroutunian et al. (2008).

Specific applications of the information theory models to design of statistical experiments are summarized by M. Malytov in the supplement to the Russian translation of Ahlswede (1987).

The early stages of information theory development involved the participation of such mathematicians as C. Shannon, N. Wiener, A. Kolmogorov, J. Wolfowitz, A. Rényi and their disciples. Leaders of the next generation were R. Gallager, A. Wyner, R. Dobrushin, M. Pinsker, J. Ziv, R. Ahlswede, T. Berger, R. Blahut, T. Cover, I. Csiszár, T. S. Han, and others. But in recent years much development of information theory in connection with statistics has taken place in general in work of specialists in electrical engineering. The most active last years have been S. Verdú , V. Poor, N. Merhav, S. Shamai (Shitz), R. Yeung, and many others.

Major information and theoretic journals and scientific meetings regularly incorporate publication of results of statistical investigations.

About the Author

From 1958 to 1970 and since 1987 Professor Evgueni A. Haroutunian has been with the Computing Center (in 1990 it was reorganized to the Institute for Informatics and Automation Problems) of the Armenian National Academy of Sciences. Since 1959, he has taught at the Yerevan State University and at the Armenian State Engineering University. During 1975–1978 and 1981–1984 he was a Visiting Professor at Algerian Universities. During 2002–2004 he was a participant of the research project

"General Theory of Information Transfer and Combinatorics" at the Center of Interdisciplinary Research (ZIF), Bielefeld University, Germany. His research interests lie in Shannon theory, hypotheses testing, and identification, as well as in application of information-theoretical and statistical methods. Professor Haroutunian is a member of the International Statistical Institute (ISI), Bernoulli Society for Mathematical Statistics and Probability and International Association for Statistical Computing (all since 1996), an Associate member of IEEE Information Theory Society (since 1994). In 2002, he was elected to the Russian Academy of Natural Sciences as a foreign member. He was a founder and the President of the Armenian Statistical Computing Society for a number of years. In October 2005 he co-organized a NATO Advanced Study Institute session on "Network Security and Intrusion Detection", Yerevan, Armenia. He is a (co-)author of more than 160 publications including Haroutunian et al. (2008).

Cross References

►Akaike's Information Criterion
►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Entropy
►Entropy and Cross Entropy as Diversity and Distance Measures
►Integrated Statistical Databases
►Kullback-Leibler Divergence
►Maximum Entropy Method for Estimation of Missing Data
►Statistical View of Information Theory

References and Further Reading

Ahlswede R, Csiszár I (1986) Hypothesis testing with communication constraints. IEEE Trans Inform Theory 32:533–542

Ahlswede R, Haroutunian EA (2005) On logarithmically asymptotically optimal testing of hypotheses and identification. Lecture notes in computer science, vol 4123. Springer, New York, pp 462–478

Ahlswede R, Wegener I (1987) Search problems. Wiley-Interscience, New York (German original, Teubner, Stuttgart (1979), Russian translation with supplement on information-theoretical methods in search problems, Mir, Moscow, 1982)

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Proceedings of 2nd international symposium on information theory, Tsahkadzor, Armenia, 1971. Akademiai Kiado, Budapest, pp 267–281

Barron AR (1997) Information theory in probability, statistics, learning, and neural nets. Department of Statistics. Yale University. Working paper distributed at plenary presentation of the 10th annual ACM workshop on computational learning theory

Blahut RE (1974) Hypotheses testing and information theory. IEEE Trans Inform Theory 20(4):405–417

Blahut RE (1987) Principles and practice of information theory. Addison-Wesley, Reading

Cover T, Thomas J (2006) Elements of information theory, 2nd edn. Hoboken

Csiszár I, Shields PS (2004) Information theory and statistics: a tutorial. Foundations and Trends in Communications and Information Theory. Now Publishers, Hanover 1:4

Haroutunian EA, Haroutunian ME, Harutyunyan AN (2008) Reliability criteria in information theory and hypothesis testing. Foundations and Trends in Communications and Information Theory. Now Publishers, Hannover 4:2–3

Hellman M, Cover T (1970) Learning with finite memory. Ann Math Statist 41:765–782

Kendall DG (1963) Information theory and the limit theorem for Markov chains and processes with a countable infinity of states. Ann Inst Stat Math 15:137–143

Kullback S (1959) Information theory and statistics. Wiley, New York

Linnik YuV (1959) An information theoretic proof of the central limit theorem on Lindeberg conditions (in Russian). Teor Veroyat Primen 4:311–321

Liese F, Vajda I (2006) On divergences and informations in statistics and information theory. IEEE Trans Inform Theory 52(10):4394–4413

Pardo L (2006) Statistical inference based on divergence methods. Chapman & Hall/CRC Press, New York

Rényi A (1968) On some problems of statistics from the point of view of information Theory. In: Rényi A (ed) Proceedings of the colloquium on information theory. János Bolyai Mathematical Society, Budapest, pp 343–357

Rissanen J (2007) Information and stability in statisical modeling. Information Science and Statistics. Springer, New York

Verdú S (1998) Fifty years of Shannon theory. IEEE Trans Inform Theory 44(6):2057–2078

Instrumental Variables

Michael P. Murray
Charles Franklin Phillips Professor of Economics
Bates College, Lewiston, ME, USA

Instrumental variables (IV) estimation can provide consistent estimates of a linear equation's parameters when ordinary ►least squares (OLS) is biased because an explanatory variable in the equation is correlated with the equation's disturbance. The necessary ingredient for consistent IV estimation is a "valid" instrument, which is a variable correlated with the offending explanatory variable but uncorrelated with the equation's disturbance term. IV estimation was first used to overcome biases in OLS by Phillip Wright (Wright 1928).

If the attractive large sample properties of the instrumental variable estimator are to be well approximated in finite samples, the correlation between the instrument and

the troublesome explanatory variable must be sufficiently high (Nelson and Startz 1990). Instruments lacking such correlation are called "weak." IV can consistently estimate an equation's parameters if there is for each troublesome explanatory variable at least one valid instrument that is not itself a variable in the model.

Suppose the equation of interest is

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon.$$

IV estimation is desirable when $E(\mathbf{X}'\varepsilon) \neq 0$. If $\mathbf{Z}$ is a matrix of instruments with the same dimensions, *kxn*, as $\mathbf{X}$ (explanatory variables uncorrelated with the disturbances, i.e., non-troublesome explanatory variables, can serve as their own instruments), then the IV estimator is

$$\tilde{\beta} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}. \quad (1)$$

When there are multiple candidate instruments for use as the instrument for a troublesome explanatory variable, any linear combination of the candidate instruments can, in principle, serve as the instrument in $\tilde{\beta}$.

Conventional practice is to form the variables in $\mathbf{Z}$ in a specific way: (i) regress each variable in X on all of the candidate instruments plus any non-troublesome explanatory variables using OLS; and (ii) set each explanatory variable's instrument equal to the explanatory variable's fitted values from (i). When the instruments are constructed in this fashion, so that $\mathbf{Z} = \hat{\mathbf{X}}$, the IV estimator can be expressed as

$$\tilde{\beta} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{Y}, \quad (2)$$

which is the OLS estimator of the equation with the actual values of the explanatory variables replaced by their fitted values. The estimator in (2) is called the *two-stage least squares* (2SLS) estimator; the first stage is the regression of each element of $\mathbf{X}$ on the m candidate instruments (Plus any non-troublesome explanatory variables, and the second stage is the OLS regression indicated by (2). When the number of candidate instruments, m, equals k (the number of explanatory variables in the equation of interest) the equation of interest is said to be *exactly identified*. The 2SLS estimator has only $m - k$ finite moments.

There are other IV estimators of the form (1) besides 2SLS. The limited information maximum likelihood (LIML) estimator simultaneously estimates the first stage equations for the troublesome variables and the equation of interest by maximum likelihood, assuming the disturbances are normally distributed. LIML is an IV estimator. Indeed, when the equation of interest is exactly identified, LIML is equivalent to 2SLS. The LIML estimator is also generally asymptotically equivalent to 2SLS, but LIML has no finite moments. Despite its lack of moments, when the number of observations is small and m is appreciably larger than k, LIML has been found to perform better than 2SLS (Davidson and MacKinnon 1993). Fuller (1977) proposed an IV estimator that performs better than 2SLS or LIML when instruments are weak (Hahn et al. 2004).

IV estimation is also applied to non-linear regression models as part of a generalized method of moments estimation. The lack of correlation between an instrument and a model's disturbances provides a moment restriction to be exploited in estimation: $E(\mathbf{Z}'\varepsilon) = 0$.

An equation's explanatory variables can be correlated with the disturbances because an explanatory variable is omitted, mis-measured, or endogenous, or because an explanatory variable is a lagged dependent variable. IV estimation can, in principle, overcome any of these problems. But IV estimation is not a panacea. At least one valid instrument must be at hand for each troublesome explanatory variable, and the cloud of uncertainty about instruments' validity that hovers over IV estimates is hardly ever entirely dispelled.

There are steps one can take to partially assess the claim that one's instruments are valid. When an equation is over-identified ($m > k$), one can formally test whether all of the instruments agree about what the parameter's actual value (this is called an over-identification test). But failing to reject this null hypothesis is consistent with all of the instruments being valid *and* with none of them being valid. Only if an instrument has been randomly assigned or if other data besides the sample in hand establish the validity of an instrument can one be secure about validity. Nonetheless, even in the case of exact identification, one can buttress confidence in the validity of an instrument by formally testing specific proposals about how an invalidating correlation between the disturbance and the instrument occurs or by appealing to either economic theory or intuition. One can also check to see whether the sign on the instrumental variable in the first stage regression of 2SLS accords with the intuitive or theoretical rationale for the instrument's validity. A significant negative coefficient in the first stage when one expects a positive sign undercuts one's faith in the instrument.

Weak instruments pose two problems: (1) the IV estimator can be almost as seriously biased as OLS in even very large samples; and (2) t- and F-tests based on the IV estimator can suffer serious size distortions. Weakness can be tested for formally using the first stage regression from 2SLS. In the case of a single troublesome explanatory variable, the classic F-statistic for the null hypothesis that variables appearing in $\mathbf{Z}$, but not in $\mathbf{X}$, have coefficients of zero in the first stage is a statistic for testing the null hypothesis that the instruments are weak, but its distribution is non-standard. Stock and Yogo (2005) offer suitable critical values. For examples, they offer critical values both for the

null hypothesis that the bias of 2SLS is greater than 10% of that of OLS and for the null hypothesis that the size distortion in a nominally 5% significance level is greater than 10%. When there are several troublesome variables, the appropriate test statistic is Cragg and Donald's multivariate extension of the F-test (Cragg and Donald 1993). Stock and Yogo provide critical values for the case of two troublesome variables.

When instruments are weak, estimation with Fuller's IV estimator is robust, even in moderate sized samples. When instruments are weak and the disturbances homoskedastic, a conditional likelihood ratio test proposed by Moreira (2003) provides optimal two-sided tests for hypotheses about linear combinations of an equation's coefficients. Moreira's critical regions can also be used to construct confidence intervals that are robust to weak instruments.

When the coefficient on an explanatory variable in a linear equation is itself a random variable, the interpretation of the IV estimator becomes more arcane. In such "heterogeneous response" cases, the IV estimator converges in probability to a weighted average of the realizations of the random coefficient *for a specific subset of the population*. In the most common of such applications, a binary instrumental variable is used to estimate the heterogeneous effect of a binary treatment variable. Imbens and Angrist (1994) named the estimand of IV estimation in this case the "local average treatment effect" (LATE). They show that the local average treatment effect is the mean effect on Y of the binary explanatory variable X ($X = 1$ indicates treatment) *for those people who would have* $X = 0$ *if* Z *were zero and would have* $X = 1$ *if* Z *were one*. IV estimation consistently estimates the LATE if the instrument: (1) is not itself an explanatory variable for Y, given X; (2) is, in essence, randomly assigned, and (3) does not both increase X for some people and decrease X for others. (The last is a constraint not required when the estimated coefficient is not random.) Sometimes the LATE is what one wants to know sometimes it is not – one might actually be interested in the mean effect of X on Y for a different subgroup of the population or for the whole population; special care must be taken when estimating heterogeneous responses by IV methods.

About the Author

Michael P. Murray received his Ph.D. in Economics in 1974, Iowa State University. He is the Charles Franklin Phillips Professor of Economics, Bates College (since 1986). His recent publications include a text *Econometrics, A Modern Introduction* (Addison-Wesley, 2005).

Cross References

►Causal Diagrams
►Method Comparison Studies
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Simulation Based Bayes Procedures for Model Structures with Non-Elliptical Posteriors
►Two-Stage Least Squares

References and Further Reading

Andrews DWK, Stock JH (eds) (2005) Identification and inference for econometric models – essays in honor of Thomas Rothenberg. Cambridge University Press, Cambridge

Andrews DWK, Moreira M, Stock JH (2006) Optimal two-sided invariant similar tests for instrumental variables regression. Econometrica 74(3):715–752

Cragg JG, Donald SG (1993) Testing identifiability and specification in instrumental variable models. Econmet theor 9:222–240

Davidson R, MacKinnon JG (1993) Estimation and inference in econometrics. Oxford University Press, New York

Fuller WA (1977) Some properties of a modification of the limited information maximum likelihood estimator, Econometrica 45(4):939–954

Hahn J, Hausman J, Kuersteiner G (2004) Estimation with weak instruments: accuracy of higher order bias and MSE approximations. Econmet J 7:272–306

Heckman JJ, Vytlacil E (2005) Structural equations, treatment effects and econometric policy evaluation. Econometrica 73(3):669–738

Imbens G, Angrist JD (1994) Identification and estimation of local average treatment effects. Econometrica 62:467–476

Moreira M (2003) A conditional likelihood ratio test for structural models. Econometrica 71(4):1027–1048

Nelson C, Startz R (1990) The distribution of the instrumental variables estimator and its F-ratio when the instrument is a poor one. J Bus 63(1):125–140

Stock JH, Yogo M (2005) Testing for weak instruments in IV regression. In: Donald WK, Andrews, James H, Stock (eds) Identification and inference for econometric models: A Festschrift in honor of Thomas Rothenberg, Cambridge University Press, pp 80–108

Wright PG (1928) The tariff on animal and vegetable oils. Macmillan, New York

Insurance, Statistics in

Steve Drekic
Associate Professor
University of Waterloo, Waterloo, ON, Canada

In attempting to analyze insurance losses arising in connection with health coverages as well as property and casualty insurance situations involving homeowner and automobile coverages, it is imperative to understand that

a portfolio of insurance business is very complicated in terms of the nature of its past and future risk-based behavior. There are many deterministic and stochastic influences at play, and the precise prediction of the future claims experience necessitates that all such influences and their effects be identified. The role of probability and statistics is vitally important in this regard, not only in terms of providing the required statistical methodology to properly analyze any data collected by the business, but also in assessing whether a quantitative (i.e., theoretical) model is able to accurately predict the claims experience of a portfolio of insurance business.

First of all, in situations when the underlying data are very extensive and have been collected in the most appropriate form for its intended purpose, it is indeed possible to answer many of the questions which arise in general insurance using observed claim size distributions and/or observed claim counts. However, it is quite often the case that data are far from extensive and may not actually be in the most convenient form for analysis. In such circumstances, calculations are only possible if certain (mathematical) assumptions are made. In other words, a quantitative model is formulated involving the use of theoretical probability distributions. Moreover, even in situations where the data are extensive, the use of theoretical distributions may still be essential. Several reasons emphasizing the importance of their use include:

1. Knowledge of their convenient and established properties, which facilitate the analysis of many problems
2. The fact that the distribution is completely summarized by a relatively small number of parameters (which characterize its location, spread, and shape) and it is not necessary to work with a lengthy set of observed data
3. The fact that they enable one to make statistical inferences concerning the behavior of insurance portfolios
4. Their tractability in terms of mathematical manipulation, permitting the development of useful theoretical results.

The Central Limit Theorem justifies why normal distributions play such an important role in statistics. In particular, the well-known law of large numbers is employed in the literature on risk management and insurance to explain pooling of losses as an insurance mechanism. For most classes of general insurance, the claim size distribution is markedly skew with a long tail to the right. If an insurer were to experience a large number of claims with respect to a particular block of business, its total payout (i.e., aggregate claims) might, however, be expected to be approximately normal distributed, being the sum of a large number of individual claims. This assumption is certainly reasonable for many purposes. There may, however, be problems associated with the extreme tails of the distribution, and these tails are particularly important for reinsurance purposes. Serious consequences could result from an insurance business basing financial risk management decisions on a model which understates the probability and scope of large losses. As a result, other parametric models, such as the gamma, log-normal, and Pareto distributions, are often much better suited to capture the positively skewed nature of the claim size distribution, and would therefore be much safer to use for estimating reinsurance premiums with regard to very large claims.

The most common and certainly best known of the claim frequency models used in practice is the Poisson distribution (see ►Poisson Distribution and Its Application in Statistics). In particular, the compound Poisson model for aggregate claims is far and away the most tractable analytically of all the compound models, as it is useful in a wide variety of insurance applications. It is also consistent with various theoretical considerations including the notion of infinite divisibility, which has practical implications in relation to the subdivision of insurance portfolios and business growth. On the other hand, the Poisson model inherently assumes that the individual risks within a portfolio of business are homogeneous from the point of view of risk characteristics, and this unfortunately leads to an inadequate fit to insurance data in some coverages. Consequently, perhaps the most important application of the negative binomial distribution, as far as general insurance applications are concerned, is in connection with the distribution of claim frequencies when the risks are heterogeneous, providing a significantly improved fit to that of the Poisson distribution.

In reference to the probability models above, it is also critical to realize that the parameters of a distribution are seldom known a priori. As a result, they need to be estimated from claims data before the distribution can be applied to a particular problem. Oftentimes, several different functions of the observed data will suggest themselves as possible estimators, and one needs to decide which one to use. The following criteria provide a good basis for determination:

1. The estimator should be *unbiased*, so that its expectation is equal to the true value of the parameter,
2. The estimator should be *consistent*, so that for an estimate based on a large number of observations, there is

a remote probability that its value will differ seriously from the true value of the parameter,
3. The estimator should be *efficient*, so that its variance is minimal.

Statisticians have developed a variety of different procedures for obtaining point estimates of parameters, including the method of moments, ►least squares, and maximum likelihood. In simple situations, the various methods often produce identical results. When sample sizes are large, they all tend to provide more or less the same answers, even in more complicated cases. In other instances, however, markedly different results can emerge, and the three criteria above are frequently used by risk practitioners in deciding which estimator to use for a given insurance application.

In conclusion, thorough treatments of these topics can be found in several reference texts including Boland (2007), Bowers et al. (1997), Daykin et al. (1994), Dickson (2005), Hossack et al. (1999), Kaas et al. (2001), and Klugman et al. (2008).

About the Author

Steve Drekic is an Associate Professor in the Department of Statistics and Actuarial Science at the University of Waterloo. Dr. Drekic's research primarily involves the use of stochastic techniques with advanced computational methods to analyze mathematical problems arising in various applications. His work has garnered particular attention in the fields of applied probability, insurance risk/ruin theory, and queueing theory. He has been actively involved in the Canadian Operational Research Society (CORS), serving as CORS Bulletin Editor (1998–2002) as well as CORS President (2005–2006). In addition, Steve Drekic became an elected member of the International Statistical Institute in 2005 and an Associate Editor for the *INFORMS Journal on Computing* in 2009.

Cross References

►Actuarial Methods
►Copulas in Finance
►Decision Theory: An Introduction
►Geometric and Negative Binomial Distributions
►Laws of Large Numbers
►Quantitative Risk Management
►Risk Analysis
►Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts
►Statistics of Extremes
►Stochastic Processes: Applications in Finance and Insurance

References and Further Reading

Boland PJ (2007) Statistical methods in insurance and actuarial science. Chapman and Hall/CRC, Boca Raton
Bowers NL, Gerber HU, Hickman JC, Jones DA, Nesbitt CJ (1997) Actuarial mathematics, 2nd edn. Society of Actuaries, Schaumburg
Daykin CD, Pentikäinen T, Pesonen M (1994) Practical risk theory for actuaries. Chapman and Hall, London
Dickson DCM (2005) Insurance risk and ruin. Cambridge University Press, Cambridge
Hossack IB, Pollard JH, Zehnwirth B (1999) Introductory statistics with applications in general insurance, 2nd edn. Cambridge University Press, Cambridge
Kaas R, Goovaerts MJ, Dhaene J, Denuit M (2001) Modern actuarial risk theory. Kluwer Academic, Dordrecht
Klugman SA, Panjer HH, Willmot GE (2008) Loss models: from data to decisions, 3rd edn. Wiley, New York

I

Integrated Statistical Databases

Samir Pradhan
Senior Researcher
Economics Program at the Gulf Research Center, Dubai, United Arab Emirates

In the current age of "information revolution," statistical information is critical for the development of society and the competitiveness of an economic system. Over the last few decades, statistical information has become the cornerstone of public decision making processes across the globe. To address the contemporary multidimensional issues of sustaining economic and social development amidst overriding changes at the national, regional and global level, statistical information systems, both official and commercial, have adopted a unified approach to integrate varied statistical information sources and tool-sets to achieve a coherent framework for production, reference and dissemination activities. In other words, statistical authorities have developed a seamless network in which users have transparent access to statistical information from a variety of sources, which are popularly known as integrated statistical databases.

An integrated statistical database (ISD) refers to amalgamation of data from varied sources through statistical integration. Statistical integration implies a framework of applications, techniques and technologies (Willis Oluoch-Kosura 2009) for combining data from different sources in order to provide the user with a unified view of the data (Lenzerini 2002) or the process of combining data

from separate sources for making use of the information in estimating accurately the missing values in any of the single datasets or an approach for enhancing the information content of separate statistical collections by ensuring consistency (Colledge 1999). These databases could be augmented with summary measures of regional and neighborhood characteristics. Integrated data offer a potent and flexible platform for analyzing various dimensions of economic behavior and the consequence of public policies. This integrated platform provides reliable, robust tools, processes and procedures for transforming disparate, unrefined data into understandable, credible and easily accessible information for action (Polach and Rodgers 2002), which could be attributed to the stupendous success of the Information and Communication Technology (ICT) Revolution in manipulating data repositories for the purpose of bridging the information deficit.

Integrated statistical databases produce and disseminate information out of data through the process called "information life cycle" (Gregory E. Farmakis et al. 2009) which transforms elementary data collected from a multitude of different sources, into valid and valuable information, suitable for statistical analysis and delivers them to information consumers, with different and often ad hoc requirements. Being conceptually different from other databases and statistical warehouses, an integrated statistical database is "a set of measurable transformations from an abstract sample space into a Euclidean space with known norm. This set of variables is accompanied by its underline structure (variance-covariance matrix, hierarchies) and its metadata specifications. These specifications refer both to the data set as a whole and to the individual variables separately, describe and where possible quantify the individualities of the statistical data set like the sampling scheme, non-response, editing, etc." (Gregory E. Farmakis et al. 2009).

As an example, an integrated economic statistical database for a particular country or region could comprise statistical reconciliation of the systems of national accounts, balance of payments, government financial statistics and other monetary and financial statistics. Conventionally, for an integrated economic statistical database, statistical integration involves three-dimensional processes – horizontal, vertical and temporal (Carol A. Hert et al. 2004). For horizontal integration, the various primary statistics on production, trade, labor and consumption need to be reconciled before they enter macroeconomic accounts (national accounts and balance of payments). Vertical integration is about reconciling primary statistics and macroeconomic accounts as well as national and international economic statistics. Temporal integration refers to reconciliation of short-term and structural economic statistics produced at different points in time but referring to the same phase in the business cycle.

An integrated statistical database involves primarily three important issues, namely, conceptual issues, statistical production issues and institutional issues. Apart from numerical consistency, ISD should also provide a framework for coordination of data in order to facilitate conceptual consistency based upon certain universally accepted concepts, definitions, classifications, and accounting rules. These classifications are applied to appropriately defined statistical units for the coherence of statistics produced. Usually a comprehensive registration process is used as a platform for structuring different units and assigning respective classifications for those units. Finally, a proper institutional arrangement is envisaged to effectively coordinate the mechanisms of the whole system.

With the current surge in regional economic integration witnessed in the proliferation of regional economic blocs across the world economy, integrated statistical databases have become an important platform for dissemination of statistical information. Good quality statistics constitute the cornerstone of success of any unified economic and monetary policy framework at the regional level (Fasano and Iqbal 2002). This, therefore, makes the harmonizing of statistical standards at the national level and developing of new statistics at the regional level imperative. The experience of Eurostat and Afristat in Europe and sub-Saharan Africa, respectively provide examples of two such successful regional ventures.

The six member states of the Gulf Cooperation Council (GCC)–Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and United Arab Emirates (UAE)–have intensified regional economic integration through establishing a common market and are on the way to a complete monetary union by 2010 based on economic convergence. In order to synergize the economic and monetary policies of the member states for the fulfillment of the convergence criteria, members are putting in place various measures to develop comparable economic data for member countries and data for the region as a whole. Efforts to coordinate statistical activity to achieve cross-country comparability are being taken in the Gulf States, particularly in the form of biannual meetings of the heads of statistical agencies. This group is developing a common vision, and each country is making an effort to bring key economic indicators up to international standards, which should contribute to greater comparability (Al-Mansouri and Dziobek 2006). In this context, the purchasing power parity (PPP) data, generated by the International Comparison Program for the six GCC

countries, are being used as a valuable tool for monitoring of economic convergence. In addition, the GCC Secretariat publishes a statistical bulletin–a starting point for a regional program of data dissemination. In this regard, the member countries are striving for an integrated statistical database for the GCC called Gulfstat in the near future. In April 2005, the heads of national statistical offices of the six countries agreed to intensify the program of statistical coordination and to consider in particular the institutional structure of statistics serving the monetary union. In principle, the heads agreed to conduct a regionally coordinated household survey in 2006 and a population census in 2010. The survey and the census will provide source data for regional economic statistics.

Acknowledgments

The views expressed here are personal. Usual disclaimer applies. The author can be contacted at spradhan@grc.ae.

Cross References

►Data Privacy and Confidentiality
►Eurostat

References and Further Reading

Al-Mansouri AKL, Dziobek C (2006) Providing official statistics for the common market and monetary union in the gulf cooperation council (GCC) countries–a case for "Gulfstat" IMF Working Paper, WP/06/38. International Monetary Fund, Washington

Afristat (2001) Contribution to statistical capacity building in member states during the 1996–2000 Period. Seminar on the launching of the study: Afristat after 2005, 7–9 May, 2001, Bamako, available on the web at http://www.Afristat.org

Agrawal R, Gupta A, Sarawagi S (1997) Modeling multidimensional databases, paper presented at the 13th International Conference on Data Engineering (ICDE'97), April 07-April 11, University of Birmingham, Birmingham, U.K, available online at http://www.computer.org/portal/web/csdl/doi?doc=doi/10.1109/ICDE.1997.581777

Colledge MJ (1999) Statistical integration through metadata management, International statistical review/revue internationale de statistique, Vol. 67, No. 1, pp. 79–98

European Commission, Eurostat, and European Central Bank (2003) Memorandum of understanding on economic and financial statistics, Brussels, March 2003, available on the web at http://www.ecb.int/

Eurostat, http://epp.eurostat.ec.europa.eu/portal/page/portal/about_eurostat/corporate/introduction

Farmakis GE, Kapetanakis Y, Petrakos GA, Petrakos MA (2009) Architecture and design of a flexible integrated information system for official statistics surveys, based on structural survey metadata", *Research Paper*, available at http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/22.pdf

Fasano U, Iqbal Z (2002) Common currency, finance & development, Vol. 39, December

Hert CA, Denn S, Haas S (2004) The role of metadata in the statistical knowledge network: an emerging research agenda, social science computing reviews.

Lenzerini M (2002) Data integration: a theoretical perspective. PODS 2002: 243–246

Polach R, Rodgers M (2002) The importance of data integration, IIM National, available at; www.iim.org.au/national/htm/default.cfm

Khawaja S, Morrison TK (2002) Statistical legislation: towards a more general framework, International Monetary Fund Working paper, WP/02/179, available on line at http://www.imf.org/external/pubs/ft/wp/2002/wp02179.pdf

United Nations Economic Commission for Europe (UN/ECE) (1995) Guidelines for the modelling of Statistical Data and Metadata, research paper of the Conference of European Statisticians, available online at http://www.unece.org/stats/publications/metadatamodeling.pdf

United Nations Economic Commission for Europe (UN/ECE) (1999) Information Systems Architecture for national and international statistical offices: Guidelines and Recommendations, research paper of the Conference of European Statisticians, available online at http://www.unece.org/stats/documents/information_systems_architecture/1.e.pdf

Willis Oluoch-Kosura (2009) What does integration imply in choosing a unit of enumeration: enterprise, holding or individual? does it Matter? perspectives from Africa, available at, www.stats.gov.cn/english/icas/. . ./P020071112580841256239.pdf

Interaction

Sander Greenland
Professor
University of California-Los Angeles, Los Angeles, CA, USA

In ordinary English, the term "interaction" usually connotes some type of causal interaction among two or more factors in producing an effect (►Causation and Causal Inference). Formal versions of these ideas are discussed in the article ►Effect Modification and Biological Interaction. The present article concerns instead the common use of the term "interaction" in the statistics literature without explicit reference to causality.

In the context of regression modeling, the phrase "interaction term" or "interaction" is most often used as a synonym for a model term involving the product of two or more variables. Consider a ►logistic regression to predict a man's actual sexual preference A ($A = 1$ for men, 0

for women) from his self-reported preference R in an interview, with G indicating the interviewer's gender ($G = 1$ for male, 0 for female):

$$P(A = 1|R, G) = \text{expit}(\alpha + \beta R + \gamma G + \delta \cdot R \cdot G),$$

where $\text{expit}(x) = e^x/(1 + e^x)$ is the logistic function. Such a model can be useful for correction of misreporting. The term $\delta \cdot R \cdot G$ is often called an "interaction," although sometimes the product $R \cdot G$ is called the "interaction term" and the coefficient δ is called the "interaction" of R and G. Nonetheless, $\delta \cdot R \cdot G$ is more accurately called a "product term," for presumably neither self-report nor interviewer status have any causal effect on actual preference, and thus cannot interact causally or modify each other's effect (because there is no effect to modify).

If $\delta \neq 0$, the product term implies that the regression of A on R depends on G: For male interviewers the regression of A on R is

$$\begin{aligned} P(A = 1|R, G = 1) &= \text{expit}(\alpha + \beta R + \gamma \cdot 1 + \delta \cdot R \cdot 1) \\ &= \text{expit}(\alpha + \gamma + (\beta + \delta)R) \end{aligned}$$

whereas for female interviewers the regression of A on R is

$$\begin{aligned} P(A = 1|R, G = 0) &= \text{expit}(\alpha + \beta R + \gamma \cdot 0 + \delta \cdot R \cdot 0) \\ &= \text{expit}(\alpha + \beta R). \end{aligned}$$

Thus we can say that the gender of the interviewer affects or modifies the logistic regression of actual preference on self-report. Nonetheless, since neither interviewer gender nor self-report affect actual preference (biologically or otherwise), they have no biologic interaction.

When both the factors in the regression causally affect the outcome, it is common to take the presence of a product term in a model as implying biologic interaction, and conversely to take absence of a product term as implying no biologic interaction. Outside of linear models, neither inference is even remotely correct: The size and even direction of the product term can change with choice regression model (e.g., linear versus logistic), whereas biologic interaction is a natural phenomenon oblivious to the model chosen for analysis (Greenland et al. 2008; Rothman 1976). The chief connection is that absence of biologic interaction leads to an absence of a product term in a linear causal model (structural equation) for risks (Greenland and Poole 1988; VanderWeele and Robins 2007). On the other hand, by definition, the presence of a product term in a causal model corresponds to effect modification when the coefficients of the model are taken as measures of effect. See ►Effect Modification and Biological Interaction for further explanation.

About the Author

For biography *see* the entry ►Confounding and Confounder Control.

Cross References

►Causation and Causal Inference
►Effect Modification and Biological Interaction
►Factorial Experiments
►Modeling Randomness Using System Dynamics Concepts
►Moderating and Mediating Variables in Psychological Research
►Multicriteria Decision Analysis
►Multivariable Fractional Polynomial Models
►Multivariate Analysis of Variance (MANOVA)
►Nonparametric Models for Anova and Ancova Designs
►Nonsampling Errors in Surveys
►Research Designs
►Sensitivity Analysis
►Statistical Design of Experiments (DOE)
►Statistical Fallacies: Misconceptions, and Myths

References and Further Reading

Greenland S, Poole C (1988) Invariants and noninvariants in the concept of interdependent effects. Scand J Work Env Hea 14:125–129
Greenland S, Lash TL, Rothman KJ (2008) Concepts of interaction, ch 5. In: Modern epidemiology, 3rd edn. Lippincott, Philadelphia, pp 71–83
Rothman KJ (1976) Causes. Am J Epidemiol 104:587–592
VanderWeele TJ, Robins JM (2007) The identification of synergism in the sufficient-component cause framework. Epidemiology 18:329–339

Interactive and Dynamic Statistical Graphics

Jürgen Symanzik
Associate Professor
Utah State University, Logan, UT, USA

Interactive and dynamic statistical graphics allow data analysts from all statistical disciplines to quickly carry out multiple visual investigations with the goal of obtaining insights into relationships for all kinds of data – from simple to complex. Often, there are no previously established hypotheses for these data, or the data set may be too big and heterogeneous for simple summaries and statistical models.

Interactive graphics and dynamic graphics are two closely related, but different, terms. On one hand, interactive graphics allow a data analyst to interact with graphical displays, typically via a computer. Depending on keystrokes, movements of the mouse, or clicks on the mouse buttons, different follow-up graphics will be produced. On the other hand, dynamic graphics display a sequence of plots without any additional user interaction. This can be some continuous rotation of a point cloud or updating plots in real time, based on additional data obtained from simulations or via data streams. Often, these two methods are used side-by-side for the investigation of the same data set. Interactive and dynamic statistical graphics play an important role in the context of ►Exploratory Data Analysis (EDA) and visual data mining (VDM).

Main Concepts of Interactive and Dynamic Statistical Graphics

Interactive and dynamic statistical graphics commonly make use of multiple, but relatively simple, plots. Frequently used plots for quantitative variables include:

- *Scatterplots*, where different symbols are plotted at horizontal ($x-$) and vertical ($y-$) positions in a two-dimensional plot area to represent the values of two quantitative variables
- *Scatterplot Matrices* (for more than two quantitative variables), where multiple scatterplots are arranged in a systematic way in a matrix
- *Parallel Coordinate Plots,* where a d-dimensional observation is represented by a continuous line drawn through d parallel coordinate axes
- *Histograms*, where area is used to display frequencies or percentages for multiple classes or intervals
- *Spinograms* (that are similar to histograms) where height is kept constant but width differs to represent the frequency or percentage for each class

In case of additional categorical variables (such as gender, race, nationality, etc.) different colors or plotting symbols are used in the previously mentioned plots to distinguish among the different groups represented by these variables.

Categorical variables themselves can be displayed via:

- *Bar Charts*, where bar length is proportional to the observed frequencies or percentages
- *Spine Plots* (that are similar to bar charts), where bar length is kept constant and bar width differs
- *Mosaic Plots,* a complex hierarchical structure that allows to display several categorical variables and visually explore questions such as independence of these variables
- *Pie Charts,* where angular areas in a circle are used to display frequencies or percentages

For data with a geographic (spatial) context, *choropleth maps* are a common component of interactive displays.

The main idea behind interactive graphics is to *link* multiple graphical displays and *brush* (or highlight) subsets of observations in these linked displays. For example, for a given data set (age, income, gender, educational level, ethnicity, nationality, and geographic subregion), consider the following scenarios:

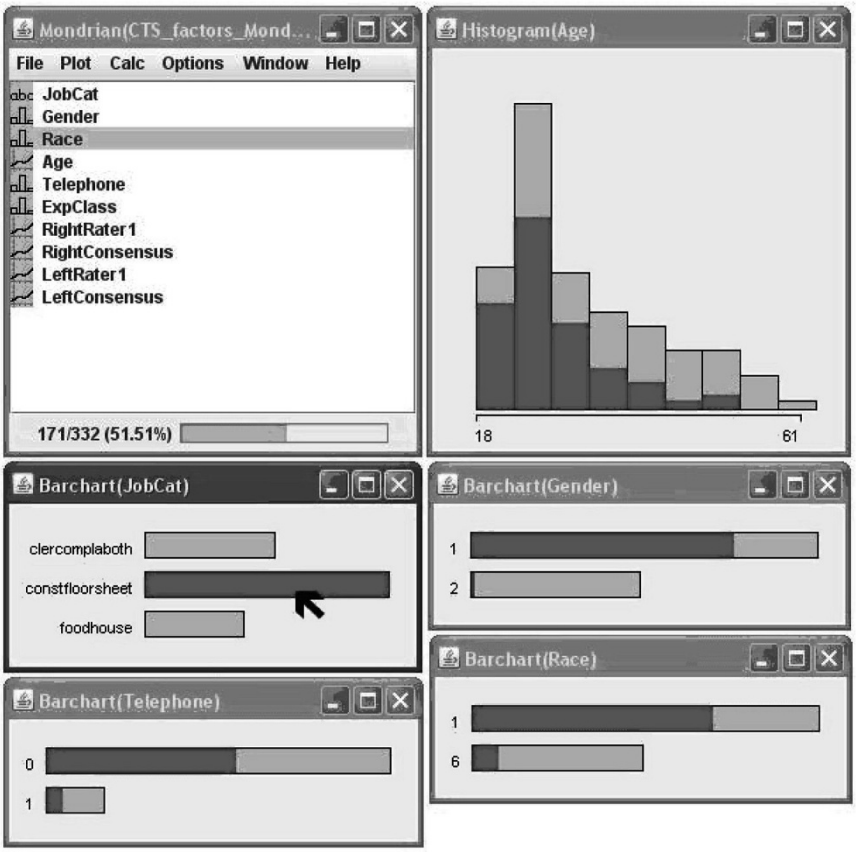
Scenario 1: In order to compare the income of individuals with a doctoral degree to those with no completed degree, then for a scatterplot showing age and income, one might want to brush different educational levels in a bar chart.

Scenario 2: In order to compare men's and women's income who are at a certain age and educational level, one might want to brush gender in another bar chart.

Scenario 3: In order to investigate the effect of ethnicity, nationality, or geographic subregion (in case a map is linked to the other displays) that may further affect the relationship between age and income, additional brushing can be performed in various linked plots.

Interactive and dynamic graphics often reveal the unexpected. For example, when distinguishing between female workers with young children and female workers without young children, a data analyst may observe that on average female workers with young children earn more money than female workers without young children as further discussed in the ►Econometrics entry of this encyclopedia.

In addition to linking and brushing, some other techniques frequently can be found in applications of interactive and dynamic statistical graphics. These techniques include *focusing* on a subset of the data (via *zooming* into a small plot area in case of overplotting of multiple nearby points or lines, or via *slicing* a high-dimensional data set into sections or slices), *rescaling* the data (e.g., by taking the log, standardizing the data, or mapping the data to a 0–1 scale), and *reformatting* the data (e.g., by swapping the axes in a scatterplot or changing the order of the axes in a parallel coordinate plot). *Rotations* are used to give the illusion of a third dimension, and even more, to find interesting views that do not align with the main coordinate axes and therefore cannot be seen in a scatterplot matrix. *Projections* often are used in sophisticated ways (such as the *grand tour*, a continuous sequence of projections) to display high-dimensional data on a 2-dimensional computer screen.



Interactive and Dynamic Statistical Graphics. Fig. 1 Data from a CTS study explored in *Mondrian*. The `Job Category` **constfloorsheet** has been brushed to further investigate the `Gender`, `Race`, and `Age` variables for this category

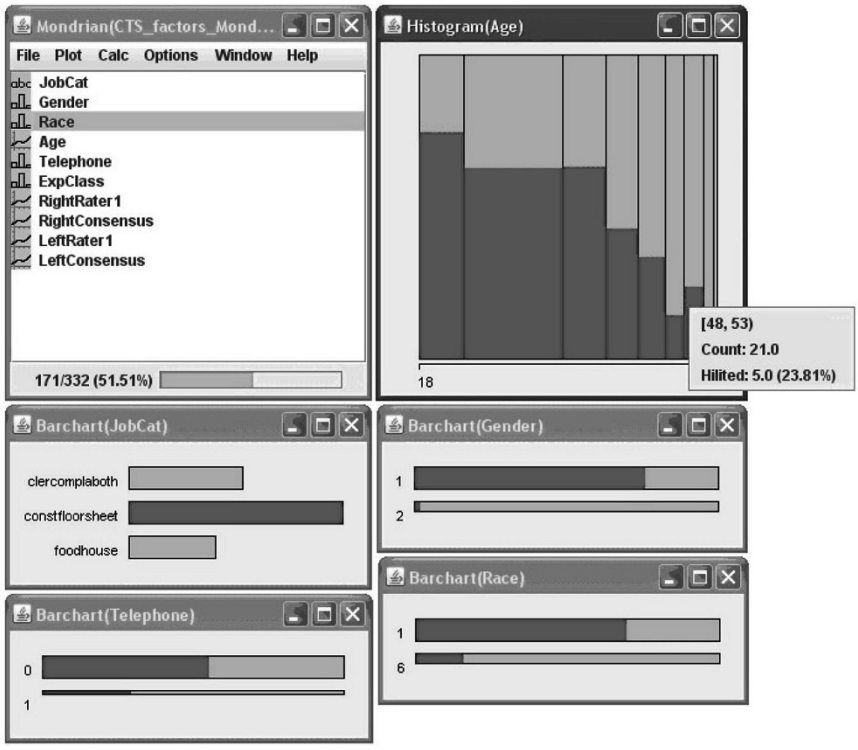
Software for Interactive and Dynamic Statistical Graphics

PRIM-9 (Picturing, Rotation, Isolation and Masking in up to nine dimensions), developed in the early 1970s, is the landmark example of software for interactive and dynamic statistical graphics. Many of the software packages developed over the following decades (and even developed these days) contain features that can be traced back to *PRIM*-9. There are three main families of software packages for interactive and dynamic statistical graphics that are freely available and widely used in the statistical research community:

- The *REGARD/MANET/Mondrian* family was initiated in the late 1980's. *Mondrian* can be freely downloaded from http://rosuda.org/mondrian/. The R package *iplots* is closely related to this family.
- The *HyperVision/ExplorN/CrystalVision* family also was initiated in the late 1980s. *CrystalVision* can be freely downloaded from http://www.galaxy.gmu.edu/pub/so%9Fware
- The *Data Viewer/XGobi/GGobi* family was already initiated in the mid 1980s. *GGobi* can be freely downloaded from http://ggobi.org. The R package *rggobi* is closely related to this family.

All of these software packages are based on the main concepts presented in the previous section. Linking and linked brushing are key components. To point out differences among these packages, *Mondrian* et al. have their strengths for categorical data and maps. *GGobi* et al. have their strengths for higher-dimensional quantitative data, in particular with respect to the *grand tour*. *CrystalVision* et al. have their main focus on parallel coordinate plots.

Web-based applications of interactive and dynamic statistical graphics that are based on *linked micromaps* (which are series of small maps linked to multiple statistical displays) can be found at the *National Cancer Institute* (NCI) State Cancer Profiles Web page at http://statecancerpro%9dles.cancer.gov/micromaps/ and at the *Utah State University* (USU) West Nile Virus Micromaps Web page at http://webcat.gis.usu.edu:.ý.ý/index.

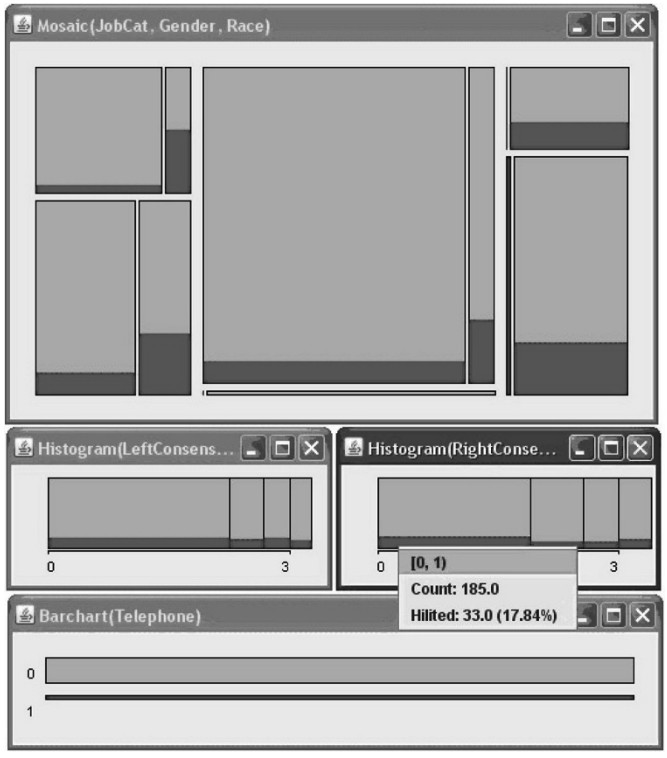


Interactive and Dynamic Statistical Graphics. Fig. 2 Histogram and bar charts interactively converted to spinogram and spine plots to better compare proportions

html. The *Gapminder Foundation* provides access to more than 200 factors of global development via its impressive interactive and dynamic Web-based software at http://www.gapminder.org/.

An Interactive Graphical Example

Some of the concepts from the previous sections will be demonstrated using a medical data set from a carpal tunnel syndrome (CTS) study. CTS is a common diagnosis of the upper extremity that may affect up to 5% of the US population - and up to 10% in some specific industries. Figure 1 shows some graphical displays for some of the variables of this study: Bar charts of `Job Category` (where *clercomplaboth* represents office and technical workers, including computer and laboratory workers; *constfloorsheet* represents construction workers, carpenters, floorlayers, and sheetmetal workers; and *foodhouse* represents service jobs, including housekeepers and food service workers), `Gender` (1 = male, 2 = female), `Race` (1 = Caucasians, 6 = others, including African Americans, Asians, and Native Americans), and `Telephone` (0 = self-administered, 1= telephone interview), and a histogram of `Age`. The main window in the upper left allows the user to interactively select additional variables, create plots, and perform statistical modeling. Currently, the `Job Category` *constfloorsheet* has been brushed in red (via a mouse click) and it turns out that most of the people in this `Job Category` are Caucasian male workers aged 18–53. Via mouse clicks and menus in these plots, the histogram is converted into a spinogram and three of the bar charts are converted into spine plots (Fig. 2). While moving the mouse over the various bars, a user can read for example that about 76% of all male workers but only about 2% of all female workers work in the *constfloorsheet* `Job Category` in this study. Similarly, about 75% of all 18–23 year olds, but only about 24% of all 48–53 year olds work in this `Job Category` - the percentage almost steadily decreasing by `Age`. About 15% of the data for this CTS study were collected via telephone interviews (`Telephone` = 1) while the remaining data were collected via self-administered questionnaires. In Fig. 3, `Telephone` = 1 has been brushed in red (via a mouse click) in the bottom plot. A mosaic plot is shown in the top plot. The first horizontal split separates `Job Category` (*cler-comp-lab-oth*, *constfloorsheet*, and *foodhouse* from left to right), the first vertical split separates `Gender` (male top and female bottom), and the second horizontal split separates `Race` (Caucasians left



Interactive and Dynamic Statistical Graphics. Fig. 3 Mosaic plot for `Job Category/Gender/Race` (*top*) and spinograms of CTS severity for *left* and *right* hands (*middle*), based on the fact whether the data were obtained via a telephone survey (brushed in *red* in the bottom spine plot) or via a self-administered questionnaire

and others right). It is apparent that the need to conduct a telephone interview highly depends on the combination of `Job Category/Gender/Race`. However, the main response variables of this study, the severity of CTS (ranging from 0 to 3) are not particularly affected by the data collection method as can be seen in the two spinograms in the middle plots. For each severity level of CTS, about 10–20% of the study participants submitted their data via telephone interviews.

The next stages of an interactive exploratory analysis could be to determine whether there is a relationship between the severity of CTS for the left hand and the right hand or how closely the diagnosis of a particular medical expert (`RightRater1` and `LeftRater1` in Fig. 1) matches the joint diagnosis of three experts (`RightConsensus` and `LeftConsensus` in Fig. 1). It should be noted that the plots created by most current software packages for interactive and dynamic graphics are crude plots with no titles, labels, or legends. It will take an additional step to transform selected graphics from an interactive exploratory session into graphics that can be used for presentations or publications.

Further Reading and Viewing

This encyclopedia entry is a brief summary of Symanzik (2004). Interested readers should refer to Symanzik (2004) for further details and for a detailed list of references that covers about 40 years of developments in the field of interactive and dynamic statistical graphics. Cook and Swayne (2007) and Theus and Urbanek (2009) are recently published textbooks that focus on interactive and dynamic graphics via *GGobi/rggobi* and *Mondrian/iplots*, respectively. Additional details on *iplots* can be found in Theus and Urbanek (2004). The *Gapminder* software is further discussed in Rosling and Johansson (2009). The Video Library of the *Sections on Statistical Computing and Statistical Graphics of the American Statistical Association* (ASA) contains 38 graphics videos covering the period from 1962 to 1996. Recently, these videos have been converted to digital format and can be watched for free at http://stat-graphics.org/movies/. The data underlying the figures of this article have been analyzed in more detail in Dale et al. (2008) and have been reused by permission. The original study was supported by Centers for Disease Control and Prevention, National Institute

for Occupational Safety and Health, Grant number R01OH008017-01.

About the Author

Dr. Jürgen Symanzik is an Associate Professor, Department of Mathematics and Statistics, Utah State University, Logan, Utah, USA. He is a Co-Editor of *Computational Statistics* (since 2005) and he previously was an Associate Editor of the same journal (1998–2005). He is currently (2010) the Chair-Elect of the Section of Statistical Graphics of the American Statistical Association (ASA) and he will take office as Chair of this section in 2011. He is also a Vice President of the International Association for Statistical Computing (IASC, 2009–2011). Dr. Symanzik has authored and co-authored a combined total of more than 60 journal papers, proceedings papers, and book chapters. He became an Elected Member of the International Statistical Institute (ISI) in 2007.

Cross References

►Data Analysis
►Exploratory Data Analysis

References and Further Reading

Cook D, Swayne DF (2007) Interactive and dynamic graphics for data analysis – with R and GGobi. Springer, New York, NY

Dale AM, Strickland J, Symanzik J, Franzblau A, Evanoff B (2008) Reliability of hand diagrams for epidemiologic case definition of carpal tunnel syndrome. J Occup Rehabil 18(3):233–248

Rosling H, Johansson C (2009) Gapminder: liberating the x-axis from the burden of time. Stat Comp Stat Gr Newslett 20(1):4–7

Symanzik J (2004) Interactive and dynamic graphics. In: Gentle JE, Härdle W, Mori Y (eds) Handbook of computational statistics – concepts and methods. Springer, Berlin, Heidelberg, pp 293–336

Theus M, Urbanek S (2004) iPlots : interactive graphics for R. Stat Comp Stat Gr Newslett 15(1):11–14

Theus M, Urbanek S (2009) Interactive graphics for data analysis: principles and examples. Chapman and Hall/CRC, Boca Raton, FL

Internet Survey Methodology: Recent Trends and Developments

Silvia Biffignandi
Professor of Statistics, DMSIA, Faculty of Economics,
Director of the Center for Statistical Analyses and Surveys
Bergamo University, Bergamo, Italy

Data collected over the Internet (Internet surveys) includes e-mail surveys and Web surveys. In e-mail surveys the ►questionnaire is completed off-line. The respondent returns the questionnaire as an email attachment and responses are given in Word or Excel format, or via other software that may be available to the respondent. In Web surveys the questionnaire is presented to respondents as a set of Web pages, answers being submitted immediately by clicking a submit/next button. Thus, Web surveys are completed on-line. Many methodological problems are common to both e-mail and Web surveys. The key advantages of Internet surveys are said to include easy access to a large group of potential respondents, low cost (costs not specifically related to the number of interviews) and timeliness. However, despite such appealing characteristics, there are serious methodological problems and even the extent of the above-mentioned advantages has yet to be fully verified in terms of actual benefit.

The University of Ljubljana (Vasja Vehovar), Bergamo University (Silvia Biffignandi), Linkoeping University (Gosta Forsman), ZUMA (Wolfing Bandilla) have been working as partners in the WebSm (WebSurvey Methodology) European project. As part of the project they have set up a website which publishes literature, software references and relevant comments within a well-organized framework (website: www.websm.org; Lozar Manfreda and Vehovar 2006). Since 2005, the group created in the context of this project has been meeting in informal workshops, under the heading of Internet Survey Methodology (ISM). The most recent Internet Survey Methodology Workshop (ISM09), organized by S. Biffignandi, was held in Bergamo from September 17–19th, 2009, and brought together highly-qualified specialists involved in analyzing and dealing with Internet survey research results.

One of the major trends in Internet survey methodology is a growing importance of Web surveys across different application fields. Initially, the main field of application was in the area of social surveys, with just a handful of papers related to ►business statistics; nowadays social surveys and business statistics are equally represented, with increasingly greater attention paid to official statistics. New research on methodological issues is emerging, together with complex research strategies. The original focus was on response rates and questionnaire design and these topics have remained of key importance, although they are now studied via more sophisticated approaches (examples would be mixed mode, complex design, questionnaire usability). These days, in fact, a great deal of attention is devoted to problems arising from the use of mixed mode, estimation, panels, and the use of incentives.

The major research areas of prime importance are: (1) survey questionnaire design; (2) methodological aspects,

especially: (a) timeliness; (b) response rate improvement, (c) participant recruitment and statistical inference.

With regard to point (1) (web questionnaire design), since web surveys are self-administered, user-friendliness and limited questionnaire length are highly important. Visual elements, pictures, color and sound are all tools that can make the questionnaire more attractive, although further research is required in order to understand whether, and how, such tools improve survey participation and quality, and what the potential disadvantages and problems may be (for example, technical limitations and information overload). For a discussion on web questionnaire design see Couper (2008) and Dillman (2007).

The first methodological aspect quoted at point (2), timeliness, requires study to establish whether this is of effective benefit to surveys. Timeliness therefore should be analyzed with reference both to the whole survey period length and to participation behavior within the survey period, as well as to the reminders timetable and mode. Some studies have shown that response times to Internet or Web surveys are different to other modes (e.g., Biffignandi and Pratesi 2000, 2002). Other research has highlighted that while there is no gain in shorter survey periods using Internet surveys, reactions to first contact and reminders are more timely.

With regard to the second methodological aspect quoted at point (2), a crucial point in web surveys is that they mostly achieve low response rates. Some studies concentrate on keeping respondents focused on the relevant parts of the computer screen, and keeping distraction to a minimum (eye-tracking analysis is used for such purposes). An interesting strategy for improving response rates is to use mixed modes. However, new problems arise with the mixed approach, since mode effects are to be taken into account in analyzing survey results. Occurrence and treatment of mixed-mode effects need further investigation.

As to the third methodological aspect quoted at point (2) (participants recruitment and statistical inference), the problem is how to select a representative sample, or, if a sample is not representative, how to correct data so as to obtain statistically representative results.

Web surveys are substantially affected by coverage, because not every member of a target population may have access to the Internet. Moreover, it is often difficult to select a proper probability sample because a sampling frame is lacking and the subpopulation with Internet access may not represent the population of interest. In general, good frames should provide a list of sampling units from which a sample can be selected and sufficient information on the basis of which the sample units can be uniquely identified in the field. In general, well-defined sampling frames are not available for most Internet surveys. At present, attempts to widen the scope of Internet-based samples beyond the population of active Internet users are unusual, being a task that is difficult to achieve. Thus, in such surveys, the respondents are a selective sample of the population; moreover, they are obviously the most skilled computer users and may therefore be much quicker than others in understanding and answering Internet interview questions. Because they may respond differently, one needs to find a way to generalize from such a sample to a target population. Major problems arise in household surveys, since many households are not Internet users and therefore cannot be recruited via Internet. In addition, even those with Internet access are potentially not expert in using the Web. Therefore, despite being set up for the Internet (possibly in an "ad hoc" way), they probably show differing survey behavior.

Scientifically meaningful results can only be obtained if proper probability samples are selected and the selection probabilities are known and positive for every member of the population. Unfortunately, many web surveys rely on a process of self-selection of respondents. The survey is simply put on the web. Respondents are those people who happen to have Internet access, visit the website and decide to participate in the survey. If an invitation for participation to the survey is sent, this invitation cannot reach the whole target population. The survey researcher is not in control of the selection process. These surveys are called self-selection surveys. See, for instance, Bethlehem (2008).

Due to imperfect frames in Web surveys, traditional probabilistic samples are in many cases not easy to implement. In order to bypass the problem of frame coverage, a number of Internet-based panels are maintained. Internet-based panels (so-called access panels) are constructed by wide participation requests on well-visited sites and Internet portals. Thus, they are also based on self-selection. Many access panels consist of volunteers and it is impossible to evaluate how well these volunteers represent the general population; in any case, they represent a non-probability sample and no statistical inference applies. However, recent research attempts to tackle the task of how to apply probabilistic recruitment to panels and how to draw inferences from them. Put briefly, trends in Web survey frames are moving in two directions. One is towards the improvement of frames by enlarging and completing the list of web users; such an approach might find greatest success with business frames and closed populations. The second direction is in the treatment of panels by

using special recruitment approaches and/or methodologies for handling results in the light of adequate auxiliary variables.

At the time of registration, basic demographic variables are recorded. A large database of potential respondents is created in this way. For future surveys, samples are selected from this database. Only panel members can participate in these Web panel surveys. The target population is however unclear. One can only say that it consists of people who have an Internet connection, who have a non-zero probability of being confronted with the invitation, and who decide to participate in the panel. Research in the Netherlands has shown that panel members differ from the general population. Access panels have the advantage that values of basic demographic variables are available for all participants. So the distribution of these variables in the survey can be compared with their distribution in the population. Over- or under-representation of specific groups can be corrected via weighting adjustment techniques. However, there is no guarantee that this leads to unbiased estimates.

To allow for the unbiased estimation of the population distribution, a reference survey can be conducted that is based on a true probability sample from the entire target population. Such a reference survey can be small in terms of the number of questions asked. It can be limited to so-called "webographic" or "psychographic" questions. Preferably, the sample size of the reference survey should be large enough to allow for precise estimations. A small sample size results in large standard errors of estimates (Bethlehem 2008).

Since most access panels are based on self-selection, it is impossible to compute unbiased estimates of population characteristics. In order to apply statistical inference properly, probability sampling, in combination with a variety of data collection modes, can be applied for panel recruitment. For example, a sample is selected from the general population and respondents are recruited using, perhaps, CAPI or CATI. Respondents without access to the Internet are provided with Internet facilities. A probabilistic sampling design can be achieved using specific methods, such as random digit dialing (RDD). Some probability-based Internet panels have already been constructed in this way (for instance, Huggins and Krotki 2001).

Another approach to correcting a lack of representativity is to apply propensity scoring methodology. Propensity scores (based on the Rosenbaum and Rubin methodology) have been used to reweight web survey results (Schonlau et al. 2009; Biffignandi and Pratesi 2005, 2003 and ISI, Berlin, August 2003).

About the Author

Dr Silvia Biffignandi is Full professor of Business and Economic Statistics at the Bergamo University and Director of CASI (Center for Statistical Analyses and Surveys). She is past Head of the (DMSIA) Department of Mathematics, Statistics, Informatics and Application (academic period 2002/2003–2008/2009 and period 1990/1991–1995/1996). She has organized the Internet Survey Methodology 2009 workshop (ISM09) held in Bergamo (17–19 September 2009). She was co-editor of the *Review Statistical Methods and Application* (Springer, 2003–2009), and previously Associate Editor of the *International Journal of Forecasting*, (Wiley editor) and of the International Journal of Forecasting (Elsevier editor). She is a Fellow of the American Statistical Association, International Statistical Institute, Italian Statistical Association, International Association Social Surveys, and International Association Official Statistics. She has been a member of the scientific committee and organization of many international conferences. Professor Biffignandi is the author of the text Handbook of Web Surveys (with J.G. Bethlehem, Wiley, Hoboken, NJ. USA, forthcoming 2011).

Cross References

►Non-probability Sampling Survey Methods
►Nonresponse in Web Surveys
►Sample Survey Methods
►Statistics: Controversies in Practice

References and Further Reading

Bethlehem JG (2008) How accurate are self-selection Web surveys? Discussion Paper 08014, Statistics Netherlands, The Hague/Heerlen, The Netherlands

Biffignandi S, Pratesi M (2000) Modelling firms response and contact probabilities in Web surveys. ICESII Proceeding Conference, Buffalo, USA

Biffignandi S, Pratesi M (2002) Modeling the respondents' profile in a Web survey on firms in Italy. In: Ferligoj A, Mrvar A (eds) Developments in social science methodology. Metodoloski zvezki, vol 18, pp 171–185

Biffignandi S, Pratesi M (2003) Potentiality of propensity scores methods in weighting for Web surveys: a simulation study. Quaderni del Dipartimento di Matematica, Statistica, Informatica e Applicazioni, n. 1

Biffignandi S, Pratesi M (2005) Indagini Web: propensity scores matching e inferenza. Un'analisi empirica e uno studio di simulazione. In: Falorsi P, Pallara A, Russo A (eds) Integrazione di dati di fonti diverse. F. Angeli, Milano, pp 131–158

Couper MP (2008) Designing effective Web surveys. Cambridge University Press, New York

Dillman D (2007) Mail and Internet surveys, the tailored design method, 2nd edn. Wiley, New York

Huggins V, Krotki K (2001) Implementation of nationally representative web-based surveys. Proceedings of the Annual Meeting of the American Statistical Association, August 5–9

I

Lozar Manfreda K, Vehovar V (2006) Web survey methodology (WebSM) portal. In: Aaron B, Aiken D, Reynolds RA, Woods R, Baker JD (eds) Handbook on research on electronic surveys and measurement, Hershey PA, pp 248–252

Schonlau M, van Soest A, Kapteyn A, Couper M (2009) Selection bias in Web surveys and the use of propensity scores. Sociol Method Res 37(3):291–318

Intervention Analysis in Time Series

Florance Matarise
Acting Department Chairperson
University of Zimbabwe, Harere, Zimbabwe

Intervention analysis is the application of modeling procedures for incorporating the effects of exogenous forces or interventions in time series analysis. These interventions, like policy changes, strikes, floods, and price changes, cause unusual changes in time series, resulting in unexpected, extraordinary observations known as ►outliers. Specifically, four types of outliers resulting from interventions, additive outliers (AO), innovational outliers (IO), temporary changes (TC), and level shifts (LS), have generated a lot of interest in literature. They pose nonstationarity challenges, which cannot be represented by the usual Box and Jenkins (1976) autoregressive integrated moving average (ARIMA) models alone.

The most popular modeling procedures are those where "intervention" detection and estimation is paramount. Box and Tiao (1975) pioneered this type of analysis in their quest to solve the Los Angeles pollution problem. Important extensions and contributions have been made by Chang et al. (1988), Chen and Liu (1993), and Chareka et al. (2006). Others, like Kirkendall (1992) and Abraham and Chuang (1993), propose the use of robust procedures where model estimation is insensitive to the effects of the interventions.

Intervention Model

An intervention model for a time series $\{Y_t\}$ according to Box and Tiao (1975) is of the dynamic form

$$Y_t = f(\kappa, X_t, t) + Z_t \tag{1}$$

where $Y_t = F(Y_t)$ *is some appropriate transformation of* Y_t *such as* $\log Y_t$ *or* $\sqrt{Y_t}$, $f(\kappa, X_t, t)$ *is a function incorporating the effects of exogenous variables* X_t, *in particular interventions,* κ *is a set of unknown parameters, and* Z_t *is the stochastic background or noise.*

The Noise model $Z_t = \pi(B) = \frac{\theta(B)}{\phi(B)\alpha(B)}$ is the usual Box and Jenkins (1976) ARIMA models, while the outliers or exogenous variables X_t themselves follow dynamic models like the following:

$$f(\delta, \omega, X_t, t) = \sum_{j=1}^{k} \frac{\omega(B)}{\delta(B)} X_t^{(i)}$$

where κ parameters are denoted by δ and ω. This intervention model can then be simplified to the general transfer function model given by

$$Y_t = Z_t + \frac{\omega(B)}{\delta(B)} X_t^{(i)} \tag{2}$$

where $Z_t = \pi(B)$ as above and

$$X_t^{(i)} = \begin{cases} 1, & t = i \\ 0, & otherwise \end{cases} \tag{3}$$

which is an indicator variable taking values 0 or 1 denoting nonoccurrence and occurrence of an intervention, respectively. For simplicity of derivation, a single exogenous variable at a time has been considered below for each of the four outlier models.

1. *The additive outlier (AO)* is usually referred to as a gross error affecting the tth observation as shown in Fig. 1. The AO model where $\delta = 0$ in the general form

$$Y_t = Z_t + \frac{\omega}{1 - \delta B} X_t^{(i)} \tag{4}$$

is represented by

$$Y_t = Z_t + \omega X_t^{(i)} \tag{5}$$

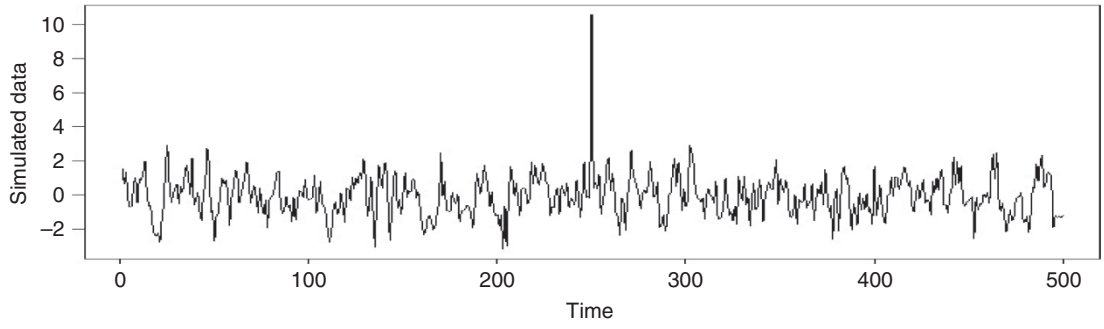
with the residual model given by $e_t = \omega\pi(B)X_t^{(i)} + a_t$.

2. *The level shift (LS)* shown in Fig. 2 is an abrupt but permanent shift by ω in the series caused by an intervention and takes on the maximum value of $\delta = 1$ in Eq. 4 so that the model becomes

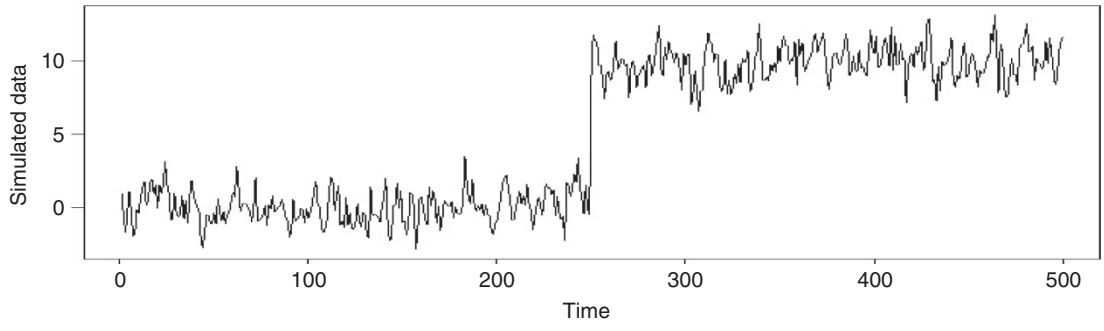
$$Y_t = Z_t + \frac{\omega}{1 - B} X_t^{(i)} \tag{6}$$

The resulting residual model is $e_t = \omega \frac{\pi(B)}{1-B} X_t^{(i)} + a_t$

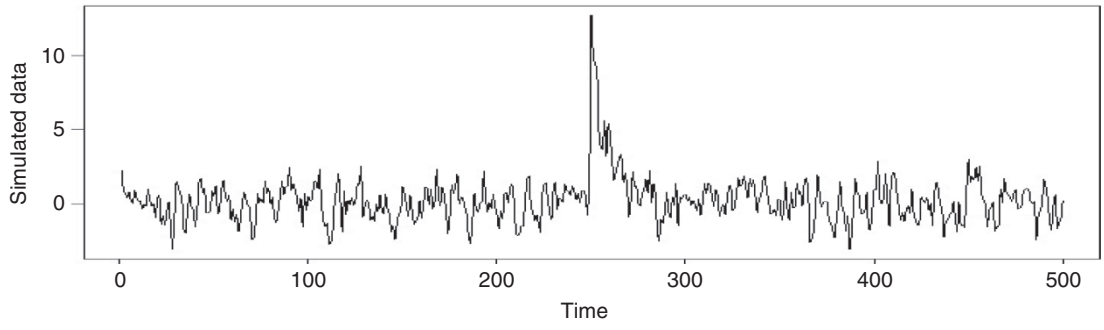
3. *The temporary change (TC)* shown in Fig. 3 is an intervention that occurs when $0 < \delta < 1$ and takes up the general form in Eq. 4. The resulting outlier has an effect ω at time t_1, which dies out gradually and has the residual model $e_t = \omega \frac{\pi(B)}{1-\delta B} X_t^{(i)} + a_t$ (see Chen and Liu 1993).
4. *The innovational outlier (IO)* shown in Fig. 4 is an extraordinary shock at time t influencing $Y_t, Y_{t+1} \ldots$



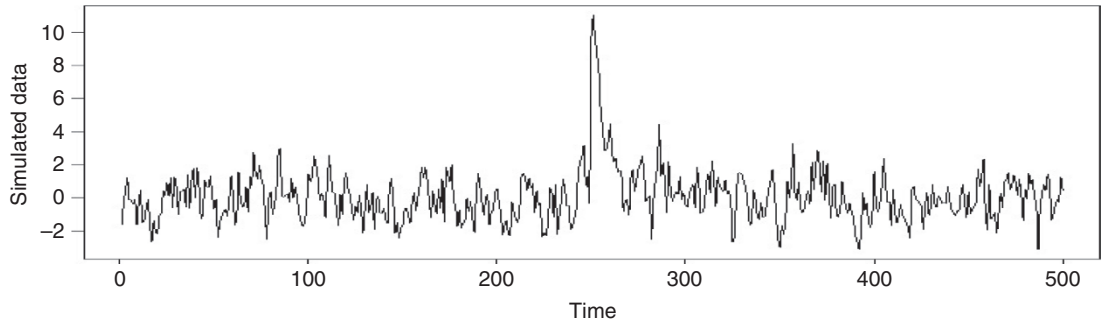
Intervention Analysis in Time Series. Fig. 1 Additive outlier effect



Intervention Analysis in Time Series. Fig. 2 Level shift



Intervention Analysis in Time Series. Fig. 3 Temporary change effect



Intervention Analysis in Time Series. Fig. 4 Innovational outlier effect

through the dynamic system resulting in the model

$$Y_t = \pi(B)\left(\omega X_t^{(i)} + a_t\right) \tag{7}$$

with residuals given by $e_t = \omega_0 X_t^{(i)} + a_t$. The underlying ARIMA model affects the overall innovational effect and can result in a gradual change, a permanent level shift or a seasonal level shift.

Stages in Intervention Analysis

There are four main stages in intervention analysis applicable to both long and short memory time series. These are intervention or outlier detection, model estimation, model diagnostics, and forecasting. However, the major challenge in intervention analysis is determining whether an intervention has actually occurred and what type it is, as described below.

Intervention/Outlier Detection

1. Plot the data to get a picture of the type of series and possible outliers in the data.
2. Assume that the underlying autoregressive moving average (ARMA) series $\{Y_t\}$ contains no outliers and use maximum likelihood estimation or ►least squares procedures to estimate its parameters.
3. State the hypothesis being tested, which is

$$H_0 : \omega_0 = 0 \quad \text{against} \quad H_1 : \omega_0 \neq 0$$

4. Compute the residuals, the impact ω and the test statistic like the popular Chang et al. (1988) likelihood ratio test statistic given by

$$T = \max\{|t_n(1)|, \ldots, |t_n(n)|\} = \max\left\{\frac{|\hat{\omega}_0(1)|}{s_n(1)}, \ldots, \frac{|\hat{\omega}_0(n)|}{s_n(n)}\right\} \tag{8}$$

where $s_n(i)$ is an estimate of the standard error of $\hat{\omega}_0(i)$; $\hat{\omega}_0(i)$ is the estimated intervention or impact at time $t = i$; the Chareka et al. (2006) statistic $C_n = \frac{T_{nn}^2 - d_n}{c_n}$, which by extreme value theory converges to the Gumbel distribution $\Lambda(x) = \exp(-e^{-x})$, $-\infty < x < \infty$; or the Abraham and Chuang (1989)s statistic $Q_{k(t)} \approx \sum_{i=t}^{t+k-1} e_i^2/(1-h_{ii})$, which asymptotically is χ^2.
5. Determine the critical values to use in the test. These can be Chang et al. (1988) critical values simulated in SPLUS, R, or others using distribution-free methods for each particular series and each outlier type at different levels of significance; Chareka et al. (2006) critical values of the Gumbel distribution 2.2504, 2.9702, and 4.6102 for $\alpha = 0.10, 0.05$, and 0.01, which can be used for all outliers as shown in recent research; while χ^2 critical values, which can be used for Abraham and Chuang (1989) test for the AO and the IO.
6. Determine whether observations are outliers and remove each outlier from the series by subtracting the value of the impact ω_i. Then apply the ARIMA modeling procedure to obtain the most adequate model and use it for forecasting future values of the series.

Robust Model Estimation

Model estimation and forecasting are the main goals of robust estimation procedures, which are insensitive to the effect of interventions. The E-M algorithm is one option proposed by Abraham and Chuang (1993). Each observation in $\mathbf{Y_t}$ is assumed to have two states, namely, the observable state and the outlier or unobservable state where $\mathbf{Y_t}$ is viewed as incomplete data and $\mathbf{X} = \left(\mathbf{Y}, \mathbf{X_t^{(i)}}\right)$ is complete data. The algorithm involves maximizing the incomplete data's likelihood function using the conditional expectation of the complete data in each iteration, resulting in parameter converging if the likelihood function meets the set conditions.

State space modeling using the Kalman filter as described in Kirkendall (1992) is another robust approach, which is based on the Markov property of ►Gaussian processes that allows the likelihood function and the ►Akaike Information Criteria to be calculated recursively for a given set of parameters. The Kalman filter model consists of the state equation $X_t = X_{t-1} + w_t$ and the measurement equation $Y_t = X_t + v_t$ with normally distributed w_t and v_t. The standard Kalman recursions are applied, and using criteria such as Bayes' rule, the observations are classified into steady state or outlier models and the various parameters are determined by minimizing the likelihood functions with respect to the parameters.

Cross References

►Kalman Filtering
►Outliers
►Time Series

References and Further Reading

Abraham B, Chuang C (1989) Outlier detection and time series modeling. Technometrics 31(2):241–247

Abraham B, Chuang C (1993) Expectation-maximization algorithms and the estimation of time series models in the presence of outliers. J Time Ser Anal 14(3):221–234

Box GEP, Jenkins GM (1970, 1976) Time series analysis forecasting and control. Holden Day, San Francisco

Box GEP, Tiao GC (1975) Intervention analysis with application to economic and environmental problems. J Am Stat Assoc 70(34):70–79

Chang I, Tiao GC, Chen C (1988) Estimation of time series parameters in the presence of outliers. Technometrics 30(2):193–204
Chareka P, Matarise F, Turner R (2006) A test for additive outliers applicable to long memory time series. J Econ Dyn Control 30(6):595–621
Chen C, Liu L-M (1993) Joint estimation of model parameter and outlier effects in time series. J Am Stat Assoc 88:284–297
Kirkendall N (1992) Monitoring outliers and level shifts in Kalman filter implementations of exponential smoothing. J Forecasting 11:543–550

Intraclass Correlation Coefficient

William D. Johnson[1], Gary G. Koch[2]
[1]Professor of Biostatistics
Louisiana State University, Baton Rouge, LA, USA
[2]Professor of Biostatistics
University of North Carolina, Chapel Hill, NC, USA

A sampling plan that has two (or more) stages of selecting study units is called a two-stage (multistage) cluster sample. A sample of the primary sampling units (the *clusters*) is selected at the first stage and a sample of the component sampling units is selected from each cluster at the second stage and so on throughout any subsequent stages. In a broader context, if observations on a set of study units are arranged in classes, categories, groups or *clusters* and some of the cluster means vary significantly, the within cluster observations will tend to be more homogeneous than those from different clusters and in this sense they will tend to be correlated. The *intraclass correlation coefficient* ρ_I quantifies the propensity for observations on units within the same cluster to be more homogeneous than those in different clusters. Early writers referred to the groups of observations as classes, hence the terminology intraclass correlation coefficient, but the evolving widespread use of cluster samples (see ▶Cluster Sampling) in both survey sampling and designed experiments has made the term "cluster" more popular; nevertheless, "intraclass" is used by most authors in referring to within cluster correlation.

Consider a finite set of M observations y_{ij} on each of N clusters where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, M$; for ease of notation but without loss of generality the present discussion is restricted to the case of all clusters having an equal number of observations. Further note that the observations are in arbitrary order within the clusters. Let μ represent the finite set mean of all NM observations and $v = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)^2/(NM)$ denote the corresponding variance. Further, let $v_a = \sum_{i=1}^{N}(\bar{y}_{i\cdot}-\mu)^2/N$ represent the variance among the cluster means and $v_w = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{i\cdot})^2/(NM)$ represent the within cluster variance where $\bar{y}_{i\cdot} = \sum_{j=1}^{M} y_{ij}/M$. A computational formula for the intraclass correlation coefficient can be expressed as (Koch 1982)

$$\rho_I = \frac{\sum_{1=1}^{N}\sum_{j\neq j'}(y_{ij}-\mu)(y_{ij'}-\mu)}{NM(M-1)v} = \frac{v_a - v_w/(M-1)}{v}$$

where $-1/(M-1) \le \rho_I \le 1$ and the lower bound is attained when the cluster means $\bar{y}_{i\cdot}$ are all equal so that $v_a = 0$ and the upper bound is attained when the cluster means are not all equal but there is no within cluster variation so $v_w = 0$ and $v = v_a$. As $M \to \infty$, ρ_I simplifies to

$$0 \le \rho_I = 1 - v_w/v \le 1.$$

In this form, the intraclass correlation coefficient is seen to be the *additive complement* of the proportion of the total variance that is attributable to variability among observations within clusters, or simply, $\rho_I = v_a/v$, the proportion of the total variance that is attributable to the component of variance for *among cluster means*. As in Koch (1982), an estimator of interest for ρ_I is that for a two stage random sample of size n from a set of N clusters and a sample of size m from each of the selected clusters where both N and M are very large relative to n and m (e.g., $N = 1,000$, $n = 50$, $M = 40$, and $m = 2$); this estimator is

$$r_I = \left(s_a^2 - s_w^2\right)/\left[s_a^2 + (m-1)s_w^2\right]$$

where $s_a^2 = m\sum_{k=1}^{n}(\bar{y}_{k\cdot}-\bar{y}_{\cdot\cdot})^2/(n-1)$ and $s_w^2 = \sum_{k=1}^{n}\sum_{l=1}^{m}(y_{kl}-\bar{y}_{k\cdot})^2/[n(m-1)]$, respectively, are the among and within cluster mean squares; $\bar{y}_{k\cdot}$ is the sample mean for observations in the k^{th} cluster and $\bar{y}_{\cdot\cdot}$ is the overall sample mean. The observations can be represented by the two-stage nested "random effects" model $y_{kl} = \mu + c_k + e_{kl}$ where μ is the overall mean, M and N are both very large and $n/N \doteq m/M \doteq 0$; the cluster effects $\{c_k\}$ and the residual errors $\{e_{kl}\}$ are mutually uncorrelated random variables with $E(c_k) = 0$, $\text{var}(c_k) = v_a$, $E(e_{kl}) = 0$ and $\text{var}(e_{kl}) = v_w$. If the $\{c_k\}$ and $\{e_{kl}\}$ also have mutually independent normal distributions, then $Q_1 = (n-1)s_a^2/(mv_a + v_w)$ has the $\chi^2(n-1)$ distribution; $Q_2 = n(m-1)s_w^2/v_w$ has the $\chi^2[n(m-1)]$ distribution; and Q_1 and Q_2 are independent. A $100(1-\alpha)$ % confidence interval for ρ_I is (Scheffe 1959)

$$\frac{\left(s_a^2/s_w^2\right) - F_2}{\left(s_a^2/s_w^2\right) + (m-1)F_2} \le \rho_I \le \frac{\left(s_a^2/s_w^2\right) - F_1}{\left(s_a^2/s_w^2\right) + (m-1)F_1}$$

where $F_1 = F_{\alpha_1}[(n-1), n(m-1)]$ and $F_2 = F_{1-\alpha_2}[(n-1), n(m-1)]$ with $\alpha_1 + \alpha_2 = \alpha$. For the case with $m = M$ and N very large, there is a random sample of clusters but all observations are used within the clusters in the sample so no subsampling is conducted at the second stage. In this situation, the sample is called a *single-stage cluster sample* and formulation of confidence intervals is based on large sample approximations.

A well-known use of the intraclass correlation is to measure observer reliability (Fleiss 1999). Suppose each of a sample of m observers follows a standard protocol to independently and in random order measure the waist circumference (WC) of each of a sample of n subjects (clusters). The two stage nested random effects model mentioned above is appropriate for the resulting data and the s_a^2 and s_w^2 can be obtained as mean squares in an ▶analysis of variance (ANOVA) and used to calculate r_I as a quantitative assessment of *inter-observer correlation*. Now, suppose a single observer measures the waist circumference (WC) on each subject in a sample of n subjects on p occasions where the observer is unaware of values of previous measurements on a given subject. Calculation of r_I from the two-stage nested random effects ANOVA is again of interest but in this case to assess *intra-observer correlation*. The closer r_I is to zero the stronger the evidence the observer *can not reliably measure* WC and conversely the closer it is to one the stronger the evidence the observer *can reliably measure* WC. Analogously, if multiple observers are used with each observer making replicate measurements, then a three-stage nested random effects model can be used to estimate variance components required for calculating both inter- and intra-observer correlation coefficients.

Intraclass correlation has been used for many years to investigate hereditary properties of human characteristics (Fisher 1918, 1946). In twin studies, for example, the clusters are defined as twin pairs and there are $m = 2$ sibs per cluster; hence, the estimator of ρ_I simplifies to $r_I = \left(s_a^2 - s_w^2\right)/\left(s_a^2 + s_w^2\right)$. The closer r_I is to zero the stronger the evidence a characteristic of interest *is not heritable* and conversely the closer it is to one the stronger the evidence it *is heritable*. Monozygotic (MZ) twins develop from a single fertilized egg and therefore share virtually all their genetic similarities whereas dizygotic (DZ) twins develop from two fertilized eggs and share on average half their similarities just as non-twin siblings. A characteristic y_{ij} can be modeled as $y_{ij} = \mu + g_i + s_i + e_{ij}$ where μ is the overall mean, the $\{g_i\}$ and $\{s_i\}$ are random genetic effects and shared environment effects, respectively, and the $\{e_{ij}\}$ are residual errors including unshared environmental effects. Let G and S denote the estimator of the proportion of total variance for y that is attributable to genetics and shared environment, respectively; further, let $r_{MZ} = G + S$ and $r_{DZ} = 0.5G + S$ represent the calculated intraclass correlation coefficients for the monozygotic and dizygotic twins, respectively. Then, the genetic or heritability coefficient is $H^2 = 2(r_{MZ} - r_{DZ})$.

For a sample of clusters, the required sample size for the total number of subjects needs to be larger than for a corresponding ▶simple random sample to account for greater variability inherent in cluster based estimators. If n_S is the required number of subjects under simple random sampling, $[n_{CS} = (DEFF)\, n_S]$ will be the required total number of subjects under cluster sampling where $DEFF \geq 1$, called the *design effect* by authors such as Kish (1965) and Cochran (1977), is the ratio of an estimator's variance using cluster sampling to its variance using simple random sampling. The design effect also can be expressed as $DEFF = [1 + (\overline{m} - 1)\rho_I]$ where $\overline{m}$ is the average cluster size and ρ_I is the intraclass correlation coefficient.

Intraclass correlation plays a fundamental role in many experimental designs that parallel its importance in survey sampling studies. For example, *cluster* or *group randomized trials* (Green 1998; Murray 1998) randomly allocate clusters rather than individual subjects to intervention conditions. Thus, community level randomization plans may be employed in trials to compare effectiveness of smoking cessation interventions where different combinations of advertisements, group lectures and drugs to attenuate nicotine craving are randomly allocated to communities. A subsample of subjects is selected within each community and, for logistic purposes, all participating subjects within a community receive the same protocol for intervention. Following the notation established above and recognizing that the number of clusters determines the residual degrees of freedom used to test the difference between two interventions, the usual formula for estimating the required total number of subjects per intervention is

$$n_{CS} = N_G \overline{m} = \left\{\left(z_\alpha + z_\beta\right)^2 \left(2\sigma^2\right)[1 + (\overline{m} - 1)\rho_I]\right\}/\Delta^2$$

where N_G is the number of clusters per intervention; z_α and z_β are the $100(1 - \alpha/2)$ and $100(1 - \beta)$ percentiles of the standard normal distribution as is appropriate for a two-directional test with significance level α and for power = $(1 - \beta)$; σ^2 is the applicable variance; and Δ is the anticipated outcome difference between groups. For these trials, it is of interest to compare interventions with respect to change in response between two observation times such as baseline to end of study. The variance for such changes can be expressed as $2\sigma_t^2(1 - \rho_t)$ where σ_t^2 is the outcome variance at a particular time t, ρ_t is the correlation between the outcome observations at two specified times, and ρ_I is

the intraclass correlation coefficient for change from baseline. The total required number of subjects in this context is

$$n_{CS} = N_G\overline{m} = \left\{\left(z_\alpha + z_\beta\right)^2\left(4\sigma_t^2\right)(1-\rho_t) \times \left[1+(\overline{m}-1)\rho_I\right]\right\}/\Delta^2.$$

Acknowledgments

The authors thank Michael Hussey and Margaret Polinkovsky for helpful comments with respect to earlier versions of this entry.

About the Authors

William D. Johnson, Ph.D. is Professor and Director of Biostatistics at Pennington Biomedical Research Center and Director of Biostatistics at Botanical Research Center at Louisiana State University. He currently serves as Biostatistics Associate Editor for *Annals of Allergy, Asthma and Immunology* and Guest Associate Editor for *Statistics in Biopharmaceutical Research*.

Gary G. Koch, Ph.D. is Professor of Biostatistics and Director of the Biometric Consulting Laboratory at the University of North Carolina at Chapel Hill. He served as Editor of *The American Statistician* during 1981–1984, and he currently has continuing service of at least 20 years on the Editorial Boards of *Statistics in Medicine* and *The Journal of Biopharmaceutical Statistics*.

Cross References

►Cluster Sampling
►Correlation Coefficient
►Sampling From Finite Populations

References and Further Reading

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York
Fisher RA (1918) The correlation between relatives on the supposition of mendelian inheritance. T R Soc Edin 52:399–433
Fisher RA (1946) Statistical methods for research workers. Oliver and Boyd, London
Fleiss JL (1999) Design and analysis of clinical experiments. Wiley, New York
Green SB (1998) Group-randomization designs. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 2. pp 1781–1784
Kish L (1965) Survey sampling. Wiley, New York
Koch GG (1982) Intraclass correlation coefficient. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 4, pp 212–217
Koch GG, Paquette DW (1997) Design principles and statistical considerations in periodontal clinical trials. Ann Periodontol 2:42–63
Murray DM (1998) Design and analysis of group-randomized trials. Oxford University Press, New York
Scheffe H (1959) The analysis of variance. Wiley, New York

Inverse Gaussian Distribution

Kalev Pärna
President of the Estonian Statistical Society, Professor of Probability, Head of the Institute of Mathematical Statistics
University of Tartu, Tartu, Estonia

Introduction

The inverse Gaussian distribution (IG) (also known as **Wald distribution**) is a two-parameter continuous distribution given by its density function

$$f(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi}}x^{-3/2}\exp\left\{-\frac{\lambda}{2\mu^2 x}(x-\mu)^2\right\},\quad x>0.$$

The parameter $\mu > 0$ is the mean and $\lambda > 0$ is the shape parameter. For a random variable (r.v.) X with inverse Gaussian distribution we write $X \sim IG(\mu,\lambda)$.

The inverse Gaussian distribution describes the distribution of the time a Brownian motion (see ►Brownian Motion and Diffusions) with positive drift takes to reach a given positive level. To be precise, let $X_t = \nu t + \sigma W_t$ be a Brownian motion with drift $\nu > 0$ (here W_t is the standard Brownian motion). Let T_a be the first passage time for a fixed level $a > 0$ by X_t. Then T_a has inverse Gaussian distribution, $T_a \sim IG\left(\frac{a}{\nu}, \frac{a^2}{\sigma^2}\right)$.

The inverse Gaussian distribution was first derived by E. Schrödinger in 1915. It belongs to the wider family of Tweedie distributions.

Characteristics of IG Distribution

The cumulative distribution function (c.d.f.) of IG is

$$F(x) = \Phi\left(\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}-1\right)\right) + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left(-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}+1\right)\right),$$

where $\Phi()$ is cumulative distribution function of the standard normal distribution. The characteristic function of IG is

$$\phi(t) = \exp\left\{\frac{\lambda}{\mu}\left(1-\sqrt{1-\frac{2\mu^2 it}{\lambda}}\right)\right\}$$

and the ►moment generating function is

$$m(t) = \exp\left\{\frac{\lambda}{\mu}\left(1-\sqrt{1-\frac{2\mu^2 t}{\lambda}}\right)\right\}.$$

Using the latter, the first four raw moments (i.e., $\alpha_n \equiv E(X^n) = \frac{\partial^n m(t)}{\partial t^n}|_{t=0}$) of the IG distribution are calculated as

$$\alpha_1 = \mu,$$
$$\alpha_2 = \mu^2 + \frac{\mu^3}{\lambda},$$
$$\alpha_3 = \mu^3 + \frac{3\mu^4}{\lambda} + \frac{3\mu^5}{\lambda^2},$$
$$\alpha_4 = \mu^4 + \frac{6\mu^5}{\lambda} + \frac{15\mu^6}{\lambda^2} + \frac{15\mu^7}{\lambda^3}.$$

Accordingly, the mean, variance, skewness, and kurtosis are obtained as

$$E(X) = \mu,$$
$$Var(X) = \frac{\mu^3}{\lambda},$$
$$Skewness\ (X) = 3\sqrt{\frac{\mu}{\lambda}},$$
$$Kurtosis\ (X) = 15\frac{\mu}{\lambda}.$$

Summation of IG-Distributed Random Variables

An important property of the IG distribution is that, with certain limitations, the sum of random variables with IG distribution is again IG distributed. More precisely, if X_i are independent and $X_i \sim IG\left(\mu_0 w_i, \lambda_0 w_i^2\right)$ then

$$\sum_{i=1}^{n} X_i \sim IG(\mu_0 \overline{w}, \lambda_0 \overline{w}^2),$$

where $\overline{w} = \sum_{i=1}^{n} w_i$. It follows, by taking $w_i \equiv 1/n$, that the inverse Gaussian distribution is infinitely divisible.

Estimation of Parameters

Let $X_1, \ldots, X_n$ be a random sample from the IG distribution $IG(\mu, \lambda)$. Then the maximum likelihood estimators for the parameters μ and λ are

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \frac{1}{\hat{\lambda}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{X_i} - \frac{1}{\hat{\mu}}\right).$$

The statistics $\hat{\mu}$ and $\hat{\lambda}$ are independent and their distributions are given by

$$\hat{\mu} \sim IG(\mu, n\lambda), \quad \frac{n}{\hat{\lambda}} \sim \chi^2_{n-1}.$$

Simulation from IG Distribution

In order to generate random numbers from the inverse Gaussian distribution, the following algorithm can be used:

1. Generate a random number z from a standard normal distribution $N(0,1)$. Let $y = z^2$.
2. Calculate $x = \mu + \frac{\mu^2 y}{2\lambda} - \frac{\mu}{2\lambda}\sqrt{4\mu\lambda y + \mu^2 y^2}$.
3. Generate a random number u from a uniform distribution $U(0,1)$.
4. If $u \le \frac{\mu}{\mu+x}$ then return x, else return μ^2/x.

Cross References

►Brownian Motion and Diffusions
►Dispersion Models
►Non-Uniform Random Variate Generations
►Normal Distribution, Univariate

References and Further Reading

Chhikara RS, Folks JL (1989) The inverse Gaussian distribution. Marcel Dekker, New York

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol 1, 2nd edn. Wiley, New York

Seshadri V (1993) The inverse Gaussian distribution. Oxford University Press, Oxford

Inverse Sampling

Man Lai Tang[1], Hon Keung Tony Ng[2]
[1]Associate Professor
Hong Kong Baptist University, Kowloon, Hong Kong
[2]Associate Professor
Southern Methodist University, Dallas, TX, USA

Introduction and Applications

The inverse sampling, first proposed by Haldane (1945), suggests one continues to sample subjects until a prespecified number of events of interest is observed. In contrast to the commonly used binomial sampling wherein the sample size is prefixed and the number of events of interest observed is random, the number of events of interest observed is prefixed for inverse sampling and the sample size is a random variable follows a negative binomial distribution. Therefore, inverse sampling is also known as *negative binomial sampling*. It is generally considered to be more appropriate than the usual binomial sampling when the subjects come sequentially, when the response probability is rare, and when maximum likelihood estimators of some epidemiological measures are undefined under binomial sampling. For instance, in epidemiological investigations, the estimation of the prevalence of a given disease on public health in a community or the variation of a disease distribution between geographical regions to locate the potential causes is one of the aims of biostatisticians, epidemiologists or medical researchers. The prevalence

here is referred to the population proportion of subjects having the disease. Usually, the estimation is done under the assumed case-binomial sampling. However, when the underlying disease is rare, the probability of obtaining only a few or zero cases in a sample under binomial sampling can be large or non-negligible. So the estimate of the population prevalence under binomial sampling can be subjected to a large relative error (Cochran 1977). To ensure that a reasonable number of cases are obtained, we may consider the use of inverse sampling.

Inverse sampling has long been appealing to practitioners and statisticians in various medical, biological and engineering research areas. For example, Smith et al. (1994) studied the level of HIV-1 mRNA-expressing (positive) mononuclear cells within the esophageal mucosa of patients with acquired immune deficiency syndrome (AIDS) and esophageal disease. Since the process of identifying positive cells could be quite tedious, they measured the prevalence of positive cells via inverse sampling. Briefly, in each slide of biopsy specimens nonoverlapping microscopic fields were examined until a field containing positive cells was found. Other applications of inverse sampling can be also found in haematological study (Haldane 1945), botanical study of plant diseases (Madden et al. 1996) and case-control study involving a rare exposure of maternal congenital heart disease (Kikuchi 1987). In software engineering, Singh et al. (1997) used the method of inverse sampling to develop a statistical procedure for quantifying the reliability of a piece of software. Their proposed procedure substantially reduced the average number of executions run over the traditional binomial sampling.

Statistical Model and Inference

Suppose that we consider a study under inverse sampling, in which we continue to collect subjects until the predetermined number r (≥ 1) of index subjects with certain attributes of interested are obtained. Let Y be the number of subjects without the attributes of interest finally accumulated in the sample before we obtain the first r index subjects. We denote the proportion p as the corresponding probability that a randomly chosen subject with the attributes of interest, where $0 < p < 1$. As a result, the random variable Y follows a negative binomial distribution with parameters r and p with probability mass function

$$f(y|p) = \Pr(Y = y|p) = \binom{r+y-1}{y} p^r (1-p)^y, \quad y = 0, 1, 2, \cdots.$$

In practice, p is usually the parameter of interest and different methods are proposed to estimate p. Two commonly used estimators of p are the maximum likelihood estimator (MLE) and the uniformly minimum variance unbiased estimator (UMVUE). The MLE of p is given by

$$\hat{p} = \frac{r}{N},$$

where $N = r + Y$ is the total number of trials required to obtain the predetermined number r. It can be shown that the variance of $\hat{p}$ is

$$Var(\hat{p}) = \frac{p^2(1-p)}{r}.$$

Note that the MLE is actually a biased estimator of p and an UMVUE of p can be obtained by

$$\tilde{p} = \frac{r-1}{N-1}.$$

The variance of $\tilde{p}$ can be shown to be

$$Var(\tilde{p}) = (r-1)(1-p)\left\{\sum_{k=1}^{r-1} \frac{(-p)^k}{(1-p)^k(r-k)} - \left(\frac{-p}{1-p}\right)^r \log(p)\right\} - p^2$$

and an unbiased estimator of $Var(\tilde{p})$ for $r > 2$ is given by

$$\widehat{Var}(\tilde{p}) = \frac{\tilde{p}(1-\tilde{p})}{N-2}.$$

For interval estimation of p, there are different types of confidence intervals available in the literature. For example, Wald-type confidence interval based on MLE or UMVUE (Lui 2004), exact confidence interval due to Casella and Berger (1990), confidence interval based on the fact that $2(r+Y)p$ follows a χ^2 distribution with $2r$ degrees of freedom (Bennett 1981), score confidence interval, likelihood ratio based confidence interval saddle-point approximation based confidence interval. For a detail review on different confidence intervals for parameter p and comparative study among these methods, one may refer the Tian et al. (2009).

Extensions to two negative binomial proportions comparison have been recently studied. For instance, Tang et al. (2007) proposed different asymptotic procedures for testing negative binomial proportion ratio. Tang et al. (2008) considered exact unconditional procedures for risk ratio under standard inverse sampling. Tian et al. (2009) derived an asymptotically reliable saddle-point approximate confidence interval for risk ratio under inverse sampling. Tang and Tian (2009, 2010) proposed various asymptotic and approximate confidence intervals for proportion difference under inverse sampling.

About the Authors

For the biography of H. K. T. Ng *see* the entry ►Censoring Methodology.

M. L. Tang is an Associate Professor in the Department of Mathematics at the Hong Kong Baptist University. He received his Ph.D. degree in Biostatistics (1995) from UCLA, USA. He is an elected member of International statistical Institute (ISI). He is currently an Associate Editor of Communications in Statistics, Advances and Applicationos in Statistical Sciences, Journal of Probability and Statistics, and The Open Medical Informatics Journal.

Cross References

►Binomial Distribution
►Estimation
►Geometric and Negative Binomial Distributions
►Proportions, Inferences, and Comparisons
►Saddlepoint Approximations

References and Further Reading

Bennett BM (1981) On the use of the negative binomial in epidemiology. Biometrical J 23:69–72
Casella G, Berger RL (1990) Statistical inference. Duxbury, Belmont
Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York
Haldane JBS (1945) On a method of estimating frequencies, Biometrika 33:222–225
Kikuchi DA (1987) Inverse sampling in case control studies involving a rare exposure. Biometrical J 29:243–246
Lui KJ (2004) Statistical estimation of epidemiological risk. Wiley, New York
Madden LV, Hughes G, Munkvold GP (1996) Plant disease incidence: inverse sampling, sequential sampling, and confidence intervals when observed mean incidence is zero. Crop Prot 15:621–632
Singh B, Viveros R, Parnas DL (1997) Estimating software reliability using inverse sampling. CRL Report 351, McMaster University, Hamilton
Smith PD, Fox CH, Masur H, Winter HS, Alling DW (1994) Quantitative analysis of mononuclear cells expressing human immunodeficiency virus type 1 RNA in esophageal mucosa. J Exp Med 180:1541–1546
Tang ML, Liao Y, Ng HKT, Chan PS (2007) On tests of rate ratio under standard inverse sampling. Comput Meth Prog Bio 89:261–268
Tang ML, Liao Y, Ng HKT (2008) Testing rate ratio under inverse sampling. Biometrical J 49:551–564
Tang ML, Tian M (2009) Asymptotic confidence interval construction for risk difference under inverse sampling. Comput Stat Data An 53:621–631
Tang ML, Tian M (2010) Approximate confidence interval construction for risk difference under inverse sampling. Stat Comput 20:87–98
Tian M, Tang ML, Ng HKT, Chan PS (2008) Confidence interval estimators for risk ratio under inverse sampling. Stat Med 27:3301–3324
Tian M, Tang ML, Ng HKT, Chan PS (2009) A comparative study of confidence intervals for negative binomial proportion. J Stat Comput Sim 33:241–249

Inversion of Bayes' Formula for Events

Kai Wang Ng
Professor and Head
The University of Hong Kong, Hong Kong, China

In standard notation, let $\{H_1, H_2, \cdots, H_m\}$ and $\{A_1, A_2, \cdots, A_n\}$ be two distinct partitions of the sample space, or equivalently two sets of events satisfying three properties: (i) each event is non-void, (ii) events in the same set are mutually exclusive (i.e., $P(H_j \cup H_k) = P(H_j) + P(H_k)$ for $j \neq k$), and (iii) each set is collectively exhaustive, (i.e., $P(\cup_{i=1}^{m} H_j) = 1$). The *Bayes' formula* in general form is, for $j = 1, \cdots, m, i = 1, \cdots, n$,

$$P(H_j|A_i) = \frac{P(A_i|H_j)P(H_j)}{P(A_i)} = \frac{P(A_i|H_j)P(H_j)}{\sum_{k=1}^{m} P(A_i|H_k)P(H_k)}, \quad (1)$$

where the last substitution is by virtue of the so-called *formula of total probability*,

$$P(A_i) = \sum_{k=1}^{m} P(A_i|H_k)P(H_k), \quad i = 1, \cdots, n, \quad (2)$$

which is valid due to properties (i) to (iii).

In Bayesian inference, which is the first paradigm of statistical inference in history, $\{H_1, H_2, \cdots, H_m\}$ are antecedent events viewed as competing hypotheses and $P(A_i|H_j)$ is the probability that the event A_i occurs as an outcome of the jth hypothesis. The investigator assigns $P(H_j)$, called the *prior probability*, to the jth hypothesis based on available information to him/her or in accordance with his/her belief on the odds of the competing hypotheses. Given that A_i occurs, Bayes' formula (1) gives the revised probability, called the *posterior probability*, of the jth competing hypothesis.

We put all the above probabilities in the combined two-way table on the top of next page, (Table 1). Where the posteriors $P_{ij} = P(H_j|A_i)$ and the likelihoods $L_{ij} = P(A_i|H_j)$ are respectively in the upper part and lower part of the (i,j) cell, while the priors $p_j = P(H_j)$ and the Bayes' factors $q_i = P(A_i)$ are in the two margins.

Now consider the question whether we can prescribe the posterior probabilities which we want to get in the end and work out the prior probabilities. In terms of the table, the question is equivalent to finding the values on margins, given all pairs of values in the cells.

A practical need leading to the above question, which seems to be the first, arises in the *Data Augmentation Algorithm* of Tanner and Wong (1987) for Bayesian inference in probability density function (pdf) setting; see also

Inversion of Bayes' Formula for Events. Table 1

$P(H_j\|A_i)$ / $P(A_i\|H_j)$	H_1	⋯	H_j	⋯	H_m	$P(A_i)$
A_1	P_{11} / L_{11}	⋯	P_{1j} / L_{1j}	⋯	P_{1m} / L_{1m}	q_1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	P_{i1} / L_{i1}	⋯	P_{ij} / L_{ij}	⋯	P_{im} / L_{im}	q_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_n	P_{n1} / L_{n1}	⋯	P_{nj} / L_{nj}	⋯	P_{nm} / L_{nm}	q_n
$P(H_j)$	p_1	⋯	p_j	⋯	p_m	1

Tanner (1996, Chap. 5) for more detail. But it was not recognized as an inversion of Bayes' formula until the author of this article revisited the integral equation for which the Algorithm aimed to solve by successive substitution; see Tan et al. (2009, pp. 1–3, 9) for a detailed account and Ng (1995a, 1997b) for original reference. Beyond the context of Bayesian inference, the inversion of Bayes' formula (*IBF*) is mathematically equivalent to de-conditioning (*DC*) in the sense of finding unconditional probabilities given the conditional ones. Note that *IBF* is also used for "*Inverse Bayes Formulae*" as in Ng (1995b, 1997a) in reference to those de-conditioning formulae in pdf setting. If the inversion is through an algorithm, as discussed by Ng (1996) for those cases of the above two-way table where r_{ij} are defined only in certain cells in haphazard patterns, we shall call it an *IBF Algorithm*, or a *DC Algorithm*.

Note that the given values in the cells are assumed to be from an existing set of joint probabilities, $P(A_i \cap H_j)$, which are not known to us. This assumption is called *compatibility* or *consistency* and needs be confirmed in applications; see Arnold et al. (1999). If we have found all p_j or all q_i, we can construct $P(A_i \cap H_j)$ and hence reconstruct all the $P(H_j|A_i)$ and $P(A_i|H_j)$. *If the given P_{ij} and L_{ij} are not identical to the reconstructed conditional probabilities, they are not compatible by contradiction; otherwise we have a constructive proof that they are.* So when an IBF or IBF algorithm is available, checking compatibility is just a natural flow of the aftermath-checking and hence we shall omit this trivial part in the article.

Although A_i and H_j are non-void events, their intersection $A_i \cap H_j$ may be void and hence the conditional probabilities concerning them may be zero. This does not matter at all in Bayes' formula, but it does in its inversion because probability ratios are the key quantities. We summarize in the following lemma some self-explaining results which we shall need in the sequel.

Lemma 1 (Preliminary facts) The following are true if A_i and H_j satisfy the aforesaid properties (i)-(iii) and if $P_{ij} = P(H_j|A_i)$ and $L_{ij} = P(A_i|H_j)$ as assumed in the table.

(a) In each cell of the table, P_{ij} and L_{ij} are simultaneously both zero if and only if $A_i \cap H_j = \emptyset$ and both positive if and only if $A_i \cap H_j \neq \emptyset$. Thus the ratio $r_{ij} = P_{ij}/L_{ij}$ is well-defined in at least one cell in each row and each column, resulting in a system of equations,

$$p_j/q_i = r_{ij} \equiv P_{ij}/L_{ij}, \quad i = 1, 2, \cdots, n;\ j = 1, \cdots, m, \tag{3}$$

for the $m + n - 2$ effective unknowns, namely p_j and q_i subject to constraints $\sum_{j=1}^{m} p_j = 1$ and $\sum_{i=1}^{n} q_i = 1$, in as many number of equations as the number of well-defined r_{ij}.

(b) If $\{p_j\}$ are determined, so are $\{q_i\}$ by (3), and vice versa. Furthermore, we may always swap the roles of rows and columns for purpose of de-conditioning without loss of generality.

(c) The $\{p_j\}$ are uniquely determined if their relative proportions can be obtained; in this case we shall say they are *completely proportionable*. For example, given the complete set of proportions relative to a common denominator p_{j^*}, $(p_1/p_{j^*}, \cdots, p_{(j^*-1)}/p_{j^*}, 1, p_{(j^*+1)}/p_{j^*}, \cdots, p_m/p_{j^*})$, we can express p_j in terms of the proportions:

$$p_j = \frac{p_j}{p_{j^*}} \left\{ \sum_{j=1}^{m} \frac{p_j}{p_{j^*}} \right\}^{-1}, \quad j = 1, 2, \cdots, m. \tag{4}$$

Furthermore, if the complete consecutive ratios $(p_1/p_2, p_2/p_3, \cdots, p_{m-1}/p_m)$ are given, we can obtain

the proportions against a common denominator by chained multiplications; for instance,

$$\frac{p_1}{p_m} = \prod_{j=1}^{m-1} \frac{p_j}{p_{j+1}}, \frac{p_2}{p_m} = \prod_{j=2}^{m-1} \frac{p_j}{p_{j+1}}, \cdots \frac{p_{m-1}}{p_m} = \prod_{j=m-1}^{m-1} \frac{p_j}{p_{j+1}}. \tag{5}$$

Finally, the analogous results concerning $q_i, i = 1, \cdots, n$, are also valid.

(d) If in a particular row i, a subset of the ratios, $r_{ij_1}, r_{ij_2}, \cdots, r_{ij_k}$, are defined, the proportions between the corresponding marginal probabilities, $p_{j_1}, p_{j_2}, \cdots, p_{j_k}$, are readily available and we shall say that $p_{j_1}, p_{j_2}, \cdots,$ and p_{j_k} are *proportionable* in row i; in notation, $[p_{j_1}(i)p_{j_2}(i)\cdots(i)p_{j_k}]$, where the order is immaterial, or $[j_1(i)j_2(i)\cdots(i)j_k]$ for short. For example, the proportions relative to p_{j_k} are:

$$\begin{aligned} p_{j_1}/p_{j_k} &= r_{ij_1}/r_{ij_k}, p_{j_2}/p_{j_k} = r_{ij_2}/r_{ij_k}, \cdots, p_{j_{k-1}}/p_{j_k} \\ &= r_{ij_{k-1}}/r_{ij_k}. \end{aligned} \tag{6}$$

The analogous conclusion about a subset of q_i at a particular column j is also valid.

So the key of inversion of Bayes' formula is to determine the relative proportions of the marginal probabilities. Now if there is one row where all ratios are defined, all $\{p_j\}$ are completely proportionable in that row according to (3), hence determining all p_j. And similarly for the q_i.

Proposition 1 (IBF: completely defined ratios in one row or column) If P_{ij} and L_{ij} are compatible, the following hold:

(a) If there is a particular row, say the i^*th row, where all ratios r_{i^*j} are defined, we have

$$\begin{aligned} p_j &= r_{i^*j}\left\{\sum_{j=1}^{m} r_{i^*j}\right\}^{-1}, \ j = 1, \cdots, m; \\ q_i &= \frac{r_{i^*j}}{r_{ij}}\left\{\sum_{j=1}^{m} r_{i^*j}\right\}^{-1}, i = 1, \cdots, n. \end{aligned} \tag{7}$$

(b) If there is a particular column, say the j^*th column, where all ratios r_{ij^*} are defined, we have

$$\begin{aligned} q_i &= r_{ij^*}^{-1}\left\{\sum_{i=1}^{n} r_{ij^*}^{-1}\right\}^{-1}, \ i = 1, \cdots, n; \\ p_j &= r_{ij}r_{ij^*}^{-1}\left\{\sum_{i=1}^{n} r_{ij^*}^{-1}\right\}^{-1}, \ j = 1, \cdots, m. \end{aligned} \tag{8}$$

(c) Under the so-called **positivity condition** where all r_{ij} are defined, the following are valid:

$$\begin{aligned} p_j &= 1/\sum_{i=1}^{n} r_{ij}^{-1} = r_{ij}\left\{\sum_{j=1}^{m} r_{ij}\right\}^{-1}, \ j = 1, \cdots, m, \\ &\quad \text{regardless any } i = 1, \cdots, n; \end{aligned} \tag{9}$$

$$\begin{aligned} q_i &= 1/\sum_{j=1}^{m} r_{ij} = r_{ij}^{-1}\left\{\sum_{i=1}^{n} r_{ij}^{-1}\right\}^{-1}, \ i = 1, \cdots, n, \\ &\quad \text{regardless any } j = 1, \cdots, m. \end{aligned} \tag{10}$$

Proof For part (a), we have all the proportions $p_j/p_m = r_{i^*j}/r_{i^*m}$ according to Lemma 1(d). By substituting them in (4) and simplifying we get the first identity, and then the second, of (7). For part (b), flip over the ratio in (3) for the particular $j = j^*$, obtaining $q_i/p_{j^*} = r_{ij^*}^{-1}$, and proceed as in the proof for part (a). Since the assumption in (c) implies that (7) is true for every $i^* = 1, \cdots, n$, we immediately have the second identity of (9) and the first identity in (10). The same assumption implies (8) for every $j^* = 1, \cdots, m$, so we have the second identity in (10) and the first identity of (9).

There are situations where each row and each column has one undefined r_{ij}, or more, so that Proposition 1 does not apply. We can still consider making $\{p_j\}$ completely proportionable using more than one row, and all q_i using more than one column. This process, however, would lead to algorithms instead of nice and neat formulae as in Proposition 1. In view of Lemma 1(b), we shall concentrate on finding p_j.

Proposition 2 (IBF: at least one ratio undefined in every row and every column) If P_{ij} and L_{ij} are compatible, the following hold:

(a) Let $(p_{j_1}, p_{j_2}, \cdots, p_{j_m})$ be any permutation of $\{p_j\}$. If p_{j_1} and p_{j_2} are proportionable in row i_1, p_{j_2} and p_{j_3} proportionable in row i_2, $\cdots, p_{j_{(m-1)}}$ and p_{j_m} proportionable in row $i_{(m-1)}$, or in notation,

$$[p_{j_1}(i_1)p_{j_2}(i_2)p_{j_3}(i_3)\cdots p_{j_{(m-1)}}(i_{(m-1)})p_{j_m}], \tag{11}$$

then $\{p_j\}$ are completely proportionable. For example, the consecutive proportions are given by

$$\frac{p_{j_1}}{p_{j_2}} = \frac{r_{i_1j_1}}{r_{i_1j_2}}, \frac{p_{j_2}}{p_{j_3}} = \frac{r_{i_2j_2}}{r_{i_2j_3}}, \cdots, \frac{p_{j_{(m-1)}}}{p_{j_m}} = \frac{r_{i_{(m-1)}j_{(m-1)}}}{r_{i_{(m-1)}j_m}}, \tag{12}$$

and we can then determine $\{p_j\}$ by (5) and (4). In practice, as shown in examples below, a common denominator would suggest itself for less work of finding the proportions.

Inversion of Bayes Formula for Events. Table 2

	p_1	p_2	p_3	p_4	p_5	p_6
q_1				r_{14}		r_{16}
q_2		r_{22}	r_{23}		r_{25}	
q_3						r_{36}
q_4	r_{41}				r_{45}	
q_5			r_{53}			
q_6		r_{62}			r_{65}	
q_7	r_{71}			r_{74}		

Lemma 1(d)	Available proportions
[4(1)6]	$p_6/p_4 = r_{16}/r_{14}$
[2(2)3(2)5]	$p_2/p_5 = r_{22}/r_{25}$, $p_3/p_5 = r_{23}/r_{25}$
[1(4)5]	$p_1/p_5 = r_{41}/r_{45}$
[2(6)5]	Done in row 2
[1(7)4]	$p_4/p_5 = (r_{74}/r_{71})(p_1/p_5)$, $p_6/p_5 = (p_6/p_4)(p_4/p_5)$.

(b) The solution for $\{p_j\}$ is not unique if the following situation happens after going down the table row by row in search for proportionable $\{p_j\}$:

The process ends up with two or more subsets of $\{p_j\}$ whose union equals the whole set of $\{p_j\}$ and which satisfy two conditions: (i) each member of one subset is found not proportionable with any member of another subset and (ii) members within the same subset are proportionable unless the subset is a singleton.

Proof (a) is straightforward. In (b), first consider the case of two subsets. We can assign an arbitrary weight, a $(0 < a < 1)$, to the sum of one subset and $1 - a$ to the sum of the other, producing a solution for $\{p_j\}$. Then we have $\{q_i\}$ by Lemma 1(b). Other cases of more than two subsets are similar.

In the following examples, all well-defined $r_{ij} = P_{ij}/L_{ij}$ are shown, while an empty cell means that the ratio is not defined for a pair of zeros. We need only demonstrate finding $\{p_j\}$.

Example 1 *(IBF for haphazard patterns of well-defined ratios)* For a haphazard pattern, we can apply Lemma 1(d) row by row to accumulate available proportions of $\{p_j\}$, with the aid of a work-sheet as illustrated in Table 2.

The first column on the right side of the table (Table 2) shows proportionable columns by each row. The results in this column suggest using p_5 as the common denominator, because it is directly proportionable to more p_j than any other choice. Then on the next column, we accumulate available proportions relative to p_5. The results from row 1 to row 6 are straightforward. Row 7 provides proportion $p_4/p_1 = r_{74}/r_{71}$, which is multiplied by (p_1/p_5) (of row 4) to yield p_4/p_5. Then p_6/p_4 from row 1 and the newly found p_4/p_5 yields p_6/p_5. At this point, all 5 proportions to p_5 are ready to yield the 6 marginal probabilities by (4) in Lemma 1(c).

Now suppose the last row is dropped and we are dealing with a 6×6 table. The above process stops at the 6th row and there are two subsets, [4(1)6] and [2(2)3(2)5(4)1], as stipulated in part (b) of proposition 2. Although the proportions within subsets are determined as before, there are infinitely many possible proportions, $a : (1 - a), 0 < a < 1$, to be allocated to the two subsets, each being as good as another in reproducing the supposedly compatible P_{ij} and L_{ij} which define r_{ij}.

Example 2 *(IBF for zigzag paths of well-defined ratios)* The two examples (Table 3) are quite straightforward in reaching (11), so we don't need a work-sheet as in the last example.

For the left table, we have $[p_1(3)p_2(3)p_3]$ with consecutive ratios, $p_1/p_2 = r_{31}/r_{32}$ and $p_2/p_3 = r_{32}/r_{33}$. Next, we have $[p_3(5)p_4(5)p_5(5)p_6]$ with consecutive ratios, $p_3/p_4 = r_{53}/r_{54}$, $p_4/p_5 = r_{54}/r_{55}$ and $p_5/p_6 = r_{55}/r_{56}$. So we can plug in (5) and then (4) to get all p_i. An alternative route is $[p_1(3)p_2(3)p_3(4)p_5(5)p_4(5)p_6]$. For the right table, we can abbreviate the situation as $[p_1(2)p_2(3)p_3(4)p_4(5)p_5(6)p_6]$, which obviously provides consecutive ratios of p_1 to p_6 and is thus completely proportionable. Note that an alternative route is $[p_1(2)p_2(3)p_3(4)p_4(6)p_5(6)p_6]$.

The Bayes' formula was developed in a manuscript by Reverend Thomas Bayes and, after his death in 1761,

Inversion of Bayes Formula for Events. Table 3

	p_1	p_2	p_3	p_4	p_5	p_6
q_1						r_{16}
q_2					r_{25}	
q_3	r_{31}	r_{32}	r_{33}			
q_4			r_{43}		r_{45}	
q_5			r_{53}	r_{54}	r_{55}	r_{56}
q_6		r_{62}				
q_7	r_{71}					

	p_1	p_2	p_3	p_4	p_5	p_6
q_1	r_{11}					
q_2	r_{21}	r_{22}				
q_3		r_{32}	r_{33}			
q_4			r_{43}	r_{44}		
q_5				r_{54}	r_{55}	
q_6				r_{64}	r_{65}	r_{66}
q_7						r_{76}

was submitted by his friend to the Royal Society for posthumous publication in 1763. It is still a puzzle as why Bayes, who "was for twenty years a Fellow of the Royal Society" (Fisher 1973, p. 8), did not submit his fine essay. Fisher (1973, p. 9) wrote: "it seems clear that Bayes had recognized that the postulate proposed in his argument (though not used in his central theorem) would be thought disputable by a critical reader, and there can be little doubt that this was the reason why his treatise was not offered for publication in his own lifetime." Stigler (1983) provided another conjecture. The mathematicians at Bayes' time, and of his standing, would usually ponder and explore all possible converses and corollaries of a theorem of importance, especially a theorem of one's own. And the converses as described above are well within Bayes' capacity and do not require any new mathematics invented after his time. It is a conjecture of the author of this article that, after finishing the manuscript, Bayes recognized, or sensed, the inversion of his formula. His prior probabilities, therefore, could be perceived as the results of reverse-engineering. So he had to think about the implications of the argument and needed more time to re-write his essay (hand-written with feather and ink at that time).

About the Author

Kai Wang Ng (http://www.hku.hk/statistics/staff/kaing/) is presently Professor & Head, Department of Statistics & Actuarial Science, The University of Hong Kong. This is his 4th Headship in the same Department and he has been instrumental in the Department's growth of enrolment and other developments since 1991. He switched his early interest in functional analysis to statistical inference and completed his Ph.D. under Prof. D.A.S. Fraser in 1975. His first book in 1990, coauthored with K.T. Fang and S. Kotz, *Symmetric and Related Multivariate Analysis* (No. 36 in Chapman & Hall Monographs on Statistics and Applied probability), has roots in his 1980 paper with D.A.S. Fraser, which shows that the inference of the regression coefficients in a univariate/multivariate linear model with a spherically symmetric error vector/ matrix should be identical with i.i.d. normal errors according to the structural approach of inference. The book has been cited by about 500 articles in more than 140 international journals in diverse fields listed in ISI Science. The Google Scholar search provides more than a thousand citations. In 2009, his book *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation* (with Ming T. Tan and Guo-Liang Tian), was published as No. 32 in Chapman & Hall/CRC Biostatistics Series. The methods introduced in the book are based on his 1995 discovery of the inversion of Bayes' formula in PDF form in the process of obtaining an explicit solution for the underlying integral equation of the Data Augmentation Algorithm of Tanner & Wong – considered as the Bayesian counter-part of the EM Algorithm in likelihood approach for handling incomplete data. This article presents a complete solution of this inversion in the most basic form – events.

Cross References

►Bayes' Theorem
►Bayesian Statistics

References and Further Reading

Arnold BC, Castillo E, Sarabia JM (1999) Conditional specification of statistical models. Springer, New York

Fisher RA (1973) Statistical methods and scientific inference, 3rd edn. Hafner Press, New York

Ng KW (1995a) Explicit formulas for unconditional PDF (Rev. March). Research Report No. 82, Department of Statistics and Actuarial Science, The University of Hong Kong

Ng KW (1995b) On the inversion of Bayes theorem. Paper presented at the 3rd ICSA statistical conference, Beijing, China, 17–20 August 1995 [Presented in a session chaired by Professor Meng XL who made reference to this talk in his discussion on a paper by George Casella, read before the Spanish Statistical Society in September 1996]

Ng KW (1996) Inversion of Bayes formula without positivity assumption. Paper presented at the Sydney international statistical congress 1996, ASC contributed session: topics in statistical inference III, Sydney, Australia, 12 July, 8:30–10:20 am. Abstract #438 in Final Programme

Ng KW (1997a) Inversion of Bayes formula: explicit formulae for unconditional pdf. In: Johnson NL, Balakrishnan N (eds) Advances in the theory and practice of statistics. Chapter 37, Wiley, New York, pp 571–584

Ng KW (1997b) Applications of inverse Bayes formula. In: Proceedings of contemporary multivariate analysis and its applications, I.1–I.10. Hong Kong Baptist University

Stigler SM (1983) Who discovered Bayes's theorem? Am Stat 37: 290–296

Tan M, Tian G, Ng KW (2009) Bayesian missing data problems: EM, data augmentation and noniterative computation. Chapman & Hall/CRC, Boca Raton, USA

Tanner MA (1996) Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions, 3rd edn. Springer, New York

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation (with discussion). J Am Stat Assoc 82:528–540

Itô Integral

Boris L. Rozovskiĭ
Ford Foundation Professor of Applied Mathematics
Brown University, Providence, RI, USA

It was established in the first half of the twentieths century that Brownian motion (Wiener process, see ▶Brownian Motion and Diffusions) $B(s)$ is of fundamental importance for stochastic modeling of many real life processes ranging from diffusion of pollen on a water surface to volatility of financial markets. Further development of stochastic modeling brought up more complicated mathematical tools, including integrals with respect to Brownian motion. The task of constructing such an integral is not trivial. Since Brownian motion is not differentiable at any point and its quadratic variation is infinite, the classical Riemann-Stielltjes integral does not provide an appropriate framework. The first successful construction of stochastic integral

$$I_t(f) = \int_0^t f(s)\, dB(s)$$

with respect to Brownian motion was proposed by N. Wiener. Wiener's integral was defined for deterministic functions $f(s) \in L_2[0,\infty)$. Wiener has also established an important isometry

$$E\left(\int_0^t f(s)\, dB(s)\right)^2 = \int_0^t f^2(s)\, ds. \tag{1}$$

In 1944–1946, K. Itô extended the Wiener integral to a large class, written $\mathcal{J}$, of random functions $f(t)$. The elements of class $\mathcal{J}$ must be non-anticipating. Roughly speaking, the latter means that $f(t)$ must be a reasonably nice (appropriately measurable) function of t and/or the path of $B(s)$ for $s \le t$. In other words, the integrand at point t may depend only on the "past" and "present" of $B(s)$ but not on the "future". In addition, the elements of $\mathcal{J}$ must be square-integrable, which means that $E\left(\int_0^\infty f^2(s)\, ds\right) < \infty$. A function $f(t)$ from class $\mathcal{J}$ is called simple if there exists a partition $0 = s_0 \le s_1 \le \ldots$ such that $f(s) = f(s_i)$ for $s \in [s_i, s_{i+1})$. The Itô integral for a simple function $f(s)$ is defined by

$$I_t(f) := \int_0^t f(s)\, dB(s) = \sum_{s_i \le t} f(s_i)\left(B(s_{i+1}) - B(s_i)\right). \tag{2}$$

It is not difficult to see that if f is a simple function from class $\mathcal{J}$, then

$$E\left|\int_0^\infty f(s)\, dB(s)\right|^2 = \int_0^\infty E|f(s)|^2\, ds. \tag{3}$$

By making use of property (3), Itô extended his integral to any $f(s)$ such that for some sequence f_n of simple functions

$$E\int_0^\infty |f(s) - f_n(s)|^2\, ds \longrightarrow 0.$$

Other important properties of Itô's integral include: (a) $E\int_0^\infty f(s)\, dB(s) = 0$; (b) $I_t(f)$ is a continuous function of t with probability 1; (c) $I_t(f)$ is a continuous martingale.

Note, that in contrast to the Riemann integral, Itô integral is sensitive to shifts of the argument of f in the right hand part of (2). In particular, if $f(s)$ is a

smooth deterministic function, $s_i^* = (s_{i+1} + s_i)/2$, and $\max_i |s_{i+1} - s_i| \longrightarrow 0$, then

$$\sum_{s_i \le t} f\left(B\left(s_i^*\right)\right)\left(B\left(s_{i+1}\right) - B\left(s_i\right)\right) \longrightarrow \int_0^t f\left(B\left(s\right)\right) \circ dB\left(s\right) := \int_0^t f\left(B\left(s\right)\right) dB\left(s\right) + \frac{1}{2}\int_0^t f'\left(B\left(s\right)\right) ds.$$

The chain rule for functions of Brownian motion, usually referred to as the Itô formula, is given by

$$f\left(B(t)\right) - f\left(B(s)\right) = \int_s^t f'\left(B(r)\right) dB\left(r\right) + \frac{1}{2}\int_s^t f''\left(B\left(r\right)\right) dr.$$

Write $X_t = X^0 \exp\{\sigma B\left(t\right) + \mu t - \sigma^2 t/2\}$, where X^0, σ and μ are constants. This process is called geometric Brownian motion. It plays fundamental role in modeling of the dynamics of financial assets. By Itô formula, one can show that X_t is a solution of the following stochastic differential equation:

$$dX_t = \mu X_t dt + \sigma X_t dB\left(t\right). \tag{4}$$

The ratio dX_t/X_t models the rate of return on the asset X_t, μ is the mean rate of return, and σ represents the volatility of the asset. The Black-Scholes option pricing formula is based on Eq. (4).

About the Author

Boris Rozovskiĭ is Ford Foundation Professor of Applied Mathematics at Brown University, Providence. He was previously Director of the Center for Applied Mathematical Sciences at the University of Southern California, Los Angelos. He is Fellow of Institute of Mathematical Statistics. Professor Rozovskiĭ is Editor of Springer's series *Stochastic Modeling and Applied Probability*, member of Advisory Board of the Journal *Asymptotic Analysis* and Associate Editor of the *SIAM Journal on Mathematical Analysis*. In the past he was an Associate Editor of several journals, including *Annals of Probability* and *Stochastic Processes and their Applications*.

Cross References

►Brownian Motion and Diffusions
►Numerical Methods for Stochastic Differential Equations
►Stochastic Differential Equations
►Stochastic Modeling, Recent Advances in

References and Further Reading

Itô K (1944) Stochastic integral. Proc Imp Acad Tokyo 20:519–524
Itô K (1951) One formula concerning stochastic differentials. Nagoya Math J 3:55–65
Karatzas I, Shreve SE (1988) Brownian motion and stochastic calculus. Springer, Berlin
Revuz D, Yor M (1988/1991) Continuous martingales and Brownian motion. Springer, Berlin

J

Jackknife

Yoshihiko Maesono
Professor, Faculty of Mathematics
Kyushu University, Fukuoka, Japan

Introduction

The jackknife method was introduced by Quenouille (1949) for reducing a bias of a correlation estimator. Tukey (1958) extended his idea to make confidence intervals. Further, Efron (1979) proposed the bootstrap method (see ►Bootstrap Methods) as an extension of the jackknife inference. Using these resampling methods, we can make statistical inferences without assuming underlying distribution of the data.

Let $X_1, \ldots, X_n$ be independently and identically distributed random variables with distribution function F_θ. Quenouille (1949) proposed a technique for reducing the bias, splitting the sample into two half-samples, and in 1956 he extended this technique which split the sample into g groups of size h, where $n = gh$. Let $\hat{\theta}_n$ be an estimator of a parameter θ based on n observations, and $\hat{\theta}^{(i)}_{(g-1)h}$ be the corresponding estimator based on $(g-1)h$ observations for $i = 1, \cdots, g$. Let us define

$$\tilde{\theta}^{(i)} = g\hat{\theta}_n - (g-1)\hat{\theta}^{(i)}_{(g-1)h}.$$

Using $\tilde{\theta}^{(i)}$, Quenouille discussed the bias reduction. Tukey (1958) called $\tilde{\theta}^{(i)}$ a pseudo-value and proposed a confidence interval, using the idea that

$$\frac{\sqrt{g}(\tilde{\theta}_n - \theta)}{\sqrt{\sum_{i=1}^{g}(\tilde{\theta}^{(i)} - \tilde{\theta}_n)^2/(g-1)}}$$

should have approximate t-distribution with $g-1$ degrees of freedom, where $\tilde{\theta}_n = \sum_{i=1}^{g}\tilde{\theta}^{(i)}/g$. His idea was based on the conjecture that $\tilde{\theta}^{(i)} (i = 1, \cdots, g)$ could be treated as approximately independent and identically distributed random variables. After their introduction of the jackknife method, many papers discussed applications to statistical inferences. Let us discuss the bias reduction, and variance and higher order moment estimation.

Bias Reduction

Let us consider an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$ of the parameter θ, and

$$\hat{\theta}^{(i)}_{n-1} = \hat{\theta}_{n-1}(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$$

denotes a corresponding estimator based on $n-1$ observations with X_i omitted. Then a bias corrected estimator is given by

$$\tilde{\theta}_J = n\hat{\theta}_n - (n-1)\overline{\theta}_n$$

where $\overline{\theta}_n = \sum_{i=1}^{n}\hat{\theta}^{(i)}_{n-1}/n$. Let us assume that $\hat{\theta}_n$ has the following bias

$$E(\hat{\theta}_n) - \theta = \frac{a_1(F)}{n} + \frac{a_2(F)}{n^2} + O(n^{-3})$$

where $a_1(F)$ and $a_2(F)$ depend on F but not on n. Then we can show that

$$E(\tilde{\theta}_J) - \theta = -\frac{a_2(F)}{n(n-1)} + O(n^{-2}) = O(n^{-2}).$$

Variance Estimation

The most popular jackknife variance estimator of $Var(\hat{\theta}_n)$ is given by

$$V_{J(1)} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}^{(i)}_{n-1} - \hat{\theta}_n\right)^2,$$

which is called a delete-1 jackknife variance estimator. If the estimator $\hat{\theta}_n$ is smooth enough, $V_{J(1)}$ is asymptotically consistent. But if the estimator is not smooth, it may be inconsistent. The best known example is the sample quantile (see Miller (1974)). To recover this inconsistency, Shao (1988) and Shao and Wu (1989) proposed the delete-d jackknife variance estimator. Let d be an integer less than n and $\mathbf{S}_{n,d}$ to be the collection of subsets of $\{1, \ldots, n\}$ with size d. For any $\delta = \{i_1, \ldots, i_d\} \in \mathbf{S}_{n,d}$, let $\hat{\theta}^{(\delta)}_{n-d}$ be the value of $\hat{\theta}_n$ when $X_{i_1}, \ldots, X_{i_d}$ are deleted from the sample. Then the delete-d jackknife variance estimator is given by

$$V_{J(d)} = \frac{n-d}{dN}\sum_{\delta}\left(\hat{\theta}^{(\delta)}_{n-d} - \hat{\theta}_n\right)^2,$$

where $N = \binom{n}{d}$ and $\sum_\delta$ is the summation over all the subsets in $\mathbf{S}_{n,d}$. For the estimators of the sample quantile, Shao

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

and Wu (1989) proved that $V_{J(d)}$ is consistent and asymptotically unbiased when $n^{1/2}/d \to 0$ and $n - d \to \infty$. When the estimator is smooth enough, like ►U-statistics, Maesono (1996) has proved that $v_{J(d-1)} \leq v_{J(d)}$ for any sample point $(x_1, \ldots, x_n)$. Efron and Stein (1981) showed that the jackknife variance estimator $V_{J(1)}$ has a positive bias. Then the bias of the delete-d jackknife variance estimator is at least as large as the bias of the delete-1 estimator in the case of the smooth original estimator.

Higher Order Moment Estimation

Let us consider the following *ANOVA*- or H-decomposition (Hoeffding):

$$\hat{\theta}_n = \theta + n^{-1}\delta + n^{-1}\sum_{i=1}^{n} g_1(X_i) + n^{-2}\sum_{1\le i<j\le n} g_2(X_i, X_j) + n^{-3}\sum_{1\le i<j<k\le n} g_3(X_i, X_j, X_k) + \cdots \quad (1)$$

where $E[g_1(X_1)] = 0$, $E[g_2(X_1, X_2)|X_1] = E[g_3(X_1, X_2, X_3)|X_1, X_2] = 0$ *a.s.* Using the von Mises expansion or H-decomposition, we can show that many statistics satisfy the equation (1). The jackknife variance estimator $nV_{J(1)}$ is consistent to $E\left[g_1^2(X_1)\right]$, which is a main term of an asymptotic variance. Hinkley (1978) and Efron and Stein (1981) studied bias reductions of the jackknife variance estimator $V_{J(1)}$, using an estimator of $E\left[g_2^2(X_1, X_2)\right]$. Lai and Wang (1993) obtained an Edgeworth expansion which includes higher order moments. $n^{-1/2}$ term of the expansion is

$$\kappa_3 = E\left[g_1^3(X_1)\right] + 3E[g_1(X_1)g_1(X_2)g_2(X_1, X_2)].$$

In order to estimate these moments, we need pseudo values of $g_1(X_i)$ and $g_2(X_i, X_j)$. They are given by

$$\hat{g}_1(X_i) = n\hat{\theta}_n - (n-1)\hat{\theta}_{n-1}^{(i)}$$

and

$$\hat{g}_2(X_i, X_j) = (n-2)\left\{n\hat{\theta}_n - (n-1)\left(\hat{\theta}_{n-1}^{(i)} + \hat{\theta}_{n-1}^{(j)}\right) + (n-2)\hat{\theta}_{n-2}^{(i,j)}\right\}$$

where $\hat{\theta}_{n-2}^{(i,j)}$ is a corresponding statistic based on $n-2$ observations with X_i and X_j omitted. For these pseudo values, we can show that

$$\hat{g}_1(X_i) = g_1(X_i) + O_p(n^{-1}) \quad \text{and}$$
$$\hat{g}_2(X_i, X_j) = g_2(X_i, X_j) + O_p(n^{-1})$$

and then we have

$$\frac{2}{n(n-1)}\sum_{1\le i<j\le n}\hat{g}_2^2(X_i, X_j) \xrightarrow{P} E\left[g_2^2(X_i, X_j)\right]$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_1^3(X_i) + \frac{6}{n(n-1)}\sum_{1\le i<j\le n}\hat{g}_1(X_i)\hat{g}_1(X_j) \times \hat{g}_2(X_i, X_j) \xrightarrow{P} \kappa_3.$$

Similarly, a pseudo value of $g_3(X_i, X_j, X_k)$ is given by

$$\hat{g}_3(X_i, X_j, X_k) = (n-4)(n-5)\left\{n\hat{\theta}_n - (n-1) \times \left(\hat{\theta}_{n-1}^{(i)} + \hat{\theta}_{n-1}^{(j)} + \hat{\theta}_{n-1}^{(k)}\right) + (n-2)\left(\hat{\theta}_{n-2}^{(i,j)} + \hat{\theta}_{n-2}^{(j,k)} + \hat{\theta}_{n-2}^{(i,k)}\right) - (n-3)\hat{\theta}_{n-3}^{(i,j,k)}\right\}$$

where $\hat{\theta}_{n-3}^{(i,j,k)}$ is a corresponding statistic based on $n-3$ observations with X_i, X_j and X_k omitted. Using the pseudo values $\hat{g}_1, \hat{g}_2$ and $\hat{g}_3$, we can make consistent estimators of the higher order moments which appear in the fourth cumulant κ_4.

About the Author

Professor Maesono is Editor of *Bulletin of Informatics and Cybernetics* (former Bulletin of Mathematical Statistics). He is also Associate Editor of *Statistics* (2006–), *Annals of the Institute of Mathematical Statistics* (2006–) and *Journal of the Korean Statistical Society* (2006–). He is an elected member of the International Statistical Institute.

Cross References

►Bootstrap Methods
►Stratified Sampling
►Target Estimation: A New Approach to Parametric Estimation

References and Further Reading

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26
Efron B, Stein C (1981) The jackknife estimate of variance. Ann Stat 9:586–596
Hinkley DV (1978) Improving the jackknife with special reference to correlation estimation. Biometrika 65:13–21
Lai TL, Wang JQ (1993) Edgeworth expansion for symmetric statistics with applications to bootstrap methods. Stat Sinica 3: 517–542
Maesono Y (1996) Higher order comparisons of jackknife variance estimators. J Nonparametr Stat 7:35–45
Miller RG (1974) The jackknife: a review. Biometrika 61:1–15
Quenouille M (1949) Approximation tests of correlation in time series. J Roy Stat Soc B 11:18–84
Quenouille M (1956) Notes on bias in estimation. Biometrika 43:353–360

Shao J (1988) Consistency of jackknife estimators of the variances of sample quantiles. Commun Stat-Theor M 17:3017–3028
Shao J, Wu CFJ (1989) A general theory for jackknife variance estimation. Ann Stat 17:1176–1197
Tukey J (1958) Bias and confidence in not quite large samples (Abstract.) Ann Math Stat 29:614

James–Stein Estimator

Christian Heumann
Ludwig Maximilian University, Munich, Germany

James–Stein Estimator in the Original Problem

The James–Stein estimator was developed in the seminal work of Charles Stein in 1956 (Stein 1956), and James and Stein in 1961 (James and Stein 1961). They showed that the ordinary least squares estimator (in a special situation as described below) is dominated by the James–Stein estimator if the dimension of the parameter vector is greater than two. Many statisticians were first not believing this phenomenon often called Stein's phenomenon or Stein's paradox. One has to realize that the James–Stein estimator is a nonlinear estimator and therefore falls outside the class of linear estimators (linear in the data y). Second it is a biased estimator and domination of the ordinary least squares is defined in terms of the scalar mean squared error criterion (MSE). The situation that was studied by James and Stein is the following: consider a parameter vector $\beta = (\beta_1, \ldots, \beta_p)$ of dimension p which is the mean of a mutivariate normal distribution (see ▶Multivariate Normal Distributions) with $p \times p$ covariance matrix $\sigma^2 I_p$, where I is the identity matrix of dimension p (ones on the diagonal, zero elsewhere). Now consider a sample y of size $n = 1$ of that multivariate normal distribution:

$$y \sim N_p(\beta, \sigma^2 I_p)\,. \tag{1}$$

It is then clear that the components of y, y_j, $j = 1, \ldots, p$, are independent normally distributed random variables with mean β_j and known homoscedastic variance σ^2. The ordinary least squares estimator, which equals the maximum likelihood estimator in the normal case, is simply

$$\hat{\beta}_{ML} = y\,. \tag{2}$$

To see this, simply write the problem as a regression model:

$$y = I_p\beta + \epsilon\,, \tag{3}$$

where $\epsilon \sim N(0, \sigma^2 I_p)$ and apply the usual ordinary least squares formula:

$$\hat{\beta}_{OLS} = \left(I_p' I_p\right)^{-1} I_p' y = y\,. \tag{4}$$

Now consider the scalar mean squared error of *all* components of an estimator $\hat{\beta}$ of β. It is defined as

$$MSE(\beta, \hat{\beta}) = E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) = \sum_{j=1}^{p} E(\hat{\beta}_j - \beta_j)^2\,. \tag{5}$$

It is easily seen that for the OLS (or ML) estimator, we get

$$MSE(\beta, \hat{\beta}) = \sum_{j=1}^{p} E(y_j - \beta_j)^2 = p\sigma^2\,. \tag{6}$$

James and Stein now showed that the estimator

$$\hat{\beta}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{y'y}\right) y = \left(1 - \frac{(p-2)\sigma^2}{\sum_{j=1}^{p} y_j^2}\right) y \tag{7}$$

dominates the OLS (ML) with respect to the MSE criterion above which means that it always has lower MSE than the OLS independent of the true β if $p > 2$. Therefore the OLS is inadmissible in that case. We can write the JS estimator also as a function of the OLS or ML estimator:

$$\left(1 - \frac{(p-2)\sigma^2}{\hat{\beta}_{ML}' \hat{\beta}_{ML}}\right) \hat{\beta}_{ML}\,. \tag{8}$$

Note, that if $(p-2)\sigma^2 < \sum_{j=1}^{p} y_j^2$, the JS estimator shrinks the estimator y towards the origin 0. That it why it is sometimes called a shrinkage estimator. James and Stein then showed that, if $p > 2$,

$$MSE(\hat{\beta}_{JS}, \beta) = p\sigma^2 - (p-2)\sigma^2 \exp\left(-\frac{\beta'\beta}{2\sigma^2}\right) \cdot F\,, \tag{9}$$

where $F = F(\beta, \sigma^2, p)$ is some positive complicated sum. It is easily seen that $MSE(\hat{\beta}_{JS}, \beta)$ is lower than $MSE(\hat{\beta}_{OLS}, \beta)$ if $p > 2$. Since F is zero if $\beta = 0$, the $MSE(\hat{\beta}_{JS}, \beta)$ reaches its minimum at $\beta = 0$ with the value $p\sigma^2 - (p-2)\sigma^2 = 2\sigma^2$. In fact, a number of modifications are possible. For example the Stein type estimator

$$\hat{\beta}_{JS} = \left(1 - \frac{c\sigma^2}{y'y}\right) y = \left(1 - \frac{c\sigma^2}{\sum_{j=1}^{p} y_j^2}\right) \tag{10}$$

can be shown to dominate the OLS as long as $0 < c < 2(p-2)$ for $p > 2$. Note, that for $p \leq 2$, the OLS is admissible. Another modification is to use some "guess" vector for β, lets say, μ. Then the JS estimator can be formulated as

$$\hat{\beta}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\sum_{j=1}^{p} (y_j - \mu_j)^2}\right) (y - \mu) + \mu\,. \tag{11}$$

Then the improvement of the JS estimator is small if $(\beta - \mu)'(\beta - \mu)$ is big and big if our guess was good, i.e., $(\beta - \mu)'(\beta - \mu)$ is small.

Summarizing the result of James and Stein we can say that, when there are three or more unrelated measurements of parameters, the MSE as defined above can be reduced by using a combined estimator as the JS estimator. But this may not be true for each single component of β. That is, the estimate may be bad for a single component and only better if we look at all components as a whole, i.e., if we look at the sum of all single MSE measures. This is an important aspect to be considered when applying such type of estimators. For each single component the OLS is admissible. A second aspect is that σ^2 is assumed to be known in all formulas. Third, the JS estimator itself is inadmissible.

The James–Stein Estimator as an Empirical Bayes Estimator

In the following we think of β as random variable with prior density

$$\beta \sim N(0, \tau^2 I_p)\,, \tag{12}$$

where τ^2 is a scalar value. The Bayes estimator can then be derived as

$$\hat{\beta}_B = \frac{\tau^2}{\tau^2 + \sigma^2} y = \left(1 - \frac{\sigma^2}{\tau^2 + \sigma^2}\right) y\,. \tag{13}$$

Now consider the case that τ is unknown. But instead of estimating τ, one can estimate the ratio

$$\frac{\sigma^2}{\tau^2 + \sigma^2} \tag{14}$$

as a whole. It can be shown that (again assuming σ^2 to be known)

$$\frac{(p-2)\sigma^2}{y'y} \tag{15}$$

is an unbiased estimator of the ratio above. If this estimator is substituted into (13), the JS estimator is obtained.

Further Modifications

Further modifications have been made by M. E. Bock in 1975 (Bock 1975) for the case that $y \sim N(\beta, \Sigma)$ where Σ is an arbitrary $p \times p$ symmetric positive definite covariance matrix. The JS estimator in this case is

$$\hat{\beta} = \left(1 - \frac{p^* - 2}{y'\Sigma^{-1}y}\right) y\,, \tag{16}$$

where p^* is defined as

$$p^* = \frac{trace(\Sigma)}{\lambda_{max}(\Sigma)}\,, \tag{17}$$

with $\lambda_{max}(\Sigma)$ the maximum eigenvalue of Σ.

Another modification is the positive–part JS estimator introduced by Baranchik in 1964 (Baranchik 1964). It is given by

$$\hat{\beta}_{JS+} = \left(1 - \frac{(p-2)\sigma^2}{y'y}\right)^+ y \tag{18}$$

where the + sign is defined as

$$z^+ = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}\,. \tag{19}$$

Thus, the effect that the factor in brackets can be negative if $y'y$ is small can be avoided. The JS+ estimator dominates the JS estimator but is itself inadmissible since it is not a smooth estimator. The other mentioned modifications can also be applied to this estimator.

Stein Type Estimators in Regression

In the usual regression setup

$$y = X\beta + \epsilon\,, \tag{20}$$

where X is a $T \times K$ matrix of covariates, Stein-type estimators are defined as

$$\hat{\beta}_{ST} = \left(1 - \frac{cs^2}{b'X'Xb}\right) b\,, \tag{21}$$

where b is the ordinary least squares estimator

$$b = (X'X)^{-1}X'y\,, \tag{22}$$

s^2 is the usual sum of squared residuals

$$s^2 = (y - Xb)'(y - Xb)\,, \tag{23}$$

and c is some constant. See e.g., Judge et al. (1985). It can be shown that the estimator is minimax if

$$0 \leq c \leq \frac{2}{T - K + 2}\left\{trace(X'X)^{-1} \cdot \frac{1}{\lambda_{max}[(X'X)^{-1}]} - 2\right\}\,. \tag{24}$$

Cross References

►Bayesian Statistics
►Bootstrap Asymptotics

References and Further Reading

Baranchik AJ (1964) Multiple regression and estimation of the mean of multivariate normal distribution. Technical Report 51, Department of Statistics, Stanford University, Stanford, California

Bock ME (1975) Minimax estimators of the mean of a multivariate normal distribution. Ann Stat 3(1):209–218

James W, Stein C (1961) Estimation with quadratic loss. In: Proceedings of the 4th Berkley symposium on mathematical statistics

and probability, vol 1. University of California Press, Berkeley, California, pp 361–379
Judge GG, Griffiths WE, Hill RC, Lee T-C (1985) The theory and practice of econometrics, 2nd edn. Wiley, New York
Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the 3rd Berkley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, California, pp 197–206

Jarque-Bera Test

Carlos M. Jarque
Inter American Development Bank, Paris, France

Testing Normality of Observations

In statistical analysis the assumption of data coming from a normal distribution is often made. In fact, up until the end of the nineteenth century, many people were convinced that there was no need for curves other than the normal distribution. Later, by the beginning of the twentieth century, most informed opinion had accepted that populations might be non-normal. This led to the development of other distribution functions and very importantly to normality testing (see Jarque and Bera (1980), and Ord (1972)). Presently, testing the normality of observations has become a standard feature in statistical work. The Jarque-Bera test is a goodness-of-fit test of departure from normality, based on the sample skewness and kurtosis.

Consider having $v_1, \ldots, v_N$ observations and the wish to test if they come from a normal distribution. The test statistic JB is defined as

$$JB = \frac{N}{6}\left(W^2 + \frac{(K-3)^2}{4}\right)$$

where N is the number of observations, W is the sample skewness, and K is the sample kurtosis, defined as

$$W = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\hat{\mu}_3}{(\hat{\sigma}^2)^{3/2}}$$

$$K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\hat{\mu}_4}{(\hat{\sigma}^2)^2}.$$

In this formula $\hat{\mu}_3$ and $\hat{\mu}_4$ are the estimates of the third and fourth central moments, respectively, and $\bar{v}$ is the sample mean:

$$\hat{\mu}_3 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \bar{v})^3$$

$$\hat{\mu}_4 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \bar{v})^4.$$

In turn, $\hat{\sigma}^2$ is the estimate of the second central moment, i.e., the sample variance

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(v_i - \bar{v})^2.$$

The hypothesis tested is if skewness is 0 and kurtosis is 3, which are the values for a normal distribution. The test statistic JB follows asymptotically ▶Chi-Square distribution with two degrees of freedom. Normality is rejected if JB is large. It has maximum asymptotic local power. For small sample sizes, significance points for $\alpha = 0.10$ and $\alpha = 0.05$ are given in the table below or may be obtainedas a routine in most statistical computer softwares (Table 1).

Testing Normality of Unobserved Regression Residuals

Now consider the linear regression model (see ▶Linear Regression Models) $y_i = x_i'\beta + u_i$ for $i = 1, \ldots, N$, where y_i is the dependent variable, x_i' a 1 by K vector of observations on K fixed regressors $(X_1, X_2, \ldots, X_K)$, u_i is the ith *unobservable* residual or disturbance, and β is a K by 1 vector of unknown parameters. Assume the model contains a constant term so $X_1 = 1$ for all i.

A traditional assumption is statistical and econometric work is that u_i is normally distributed (with zero mean and constant variance σ^2). This model has wide, everyday application. It is used to model phenomena in many sectors: economic, financial, social, technological, natural sciences, etc. The consequences of u_i not being normally

Jarque-Bera Test. Table 1 Normality of observations; significance points for JB normality test

N	$\alpha = 0.10$	$\alpha = 0.05$	N	$\alpha = 0.10$	$\alpha = 0.05$
20	2.13	3.26	200	3.48	4.43
30	2.49	3.71	250	3.54	4.51
40	2.70	3.99	300	3.68	4.60
50	2.90	4.26	400	3.76	4.74
75	3.09	4.27	500	3.91	4.82
100	3.14	4.29	800	4.32	5.46
125	3.31	4.34	∞	4.61	5.99
150	3.43	4.39			

Source: Jarque and Bera (1987)

distributed are many, in terms of validity of inferential processes (e.g., traditional t and F-tests are sensitive to non-normality); efficiency of estimates used (the ordinary least squares estimator of β may be very sensitive to non-normality, particularly in long tailed distributions); and efficiency in forecasting techniques. This may lead to research and studies that arrive at wrong conclusions and to incorrect decision making, be it a policy prescription or a scientific finding.

As shown in Jarque and Bera (1980), formulating the distribution of u_i as a member of the Pearson Family of Distributions and applying the Lagrange Multiplier principle to test for normality of u_i within this overarching family of distributions, then the *JB* statistic is obtained. In this case, the test would be computed using the OLS residuals $\hat{u}_1, \ldots, \hat{u}_N$ (i.e., substitute v_i for $\hat{u}_i$ in *JB*), where $\hat{u}_i = y_i - x_i' b$ and where b in the OLS estimate of β

$$b = \left(X'X\right)^{-1} X'y \text{ and with } y = (y_1, \ldots, y_N)' \text{ and } X' = (x_1, \ldots, x_N).$$

The *JB* test applied with OLS residuals is simple to compute and has maximum local asymptotic power. It is asymptotically distributed as Chi-Square with two degrees of freedom. For small samples, in any regression model, the approximation to the finite sample distribution of the test can be easily obtained by computer simulation. Regression residual normality is rejected for large values of *JB*. Naturally the test may also be applied even if the distribution of u_i is outside the Pearson Family of Distributions. Due to its simplicity and good power properties it has wide use in statistical regression analysis.

About the Author

Carlos M. Jarque graduated in 1976 with a degree in Actuarial Science, at Anáhuac University, Mexico. He then obtained a Posgraduate Diploma and a Master's Degree from the London School of Economics and Political Science. He also undertook graduate studies in planning and economic policy at the University of Oslo. He holds a Doctorate in Economics from the Australian National University, and obtained a posdoctorate in Econometrics at Harvard University. Dr. Jarque has taught at the Australian National University, and at Harvard University. Dr. Jarque was Director and Vice-President of the International Statistical Institute. He was also elected Chairman of the United Nations Statistical Commission, and President of the U.N. Cartographic Conference. He was the Minister of Social Development in Mexico, Secretary for the National Development Plan of Mexico (1995–2000), and President of the National Institute of Statistics, Geography and Informatics of Mexico. From January 2001 to December 2005, he was Manager of the Sustainable Development Department of the Inter-American Development Bank, and from December 2005 to August 2007, the Secretary of the IDB. He is currently the IDB Representative in Europe and Israel, and Principal Advisor of the IDB President. Dr. Jarque is the author of over a hundred academic articles on economics, social development, planning and technology. He is well known in statistics and econometrics, among other contributions, for his paper (with Anil K. Bera) *Efficient tests for normality, homoscedasticity and serial independence of regression residuals* (*Economic Letters*, Volume 6, Issue 3, 1980, pp. 255–259), where a test was proposed, that is today known as Jarque-Bera test. He has received numerous distinctions and honors, including Mexico's National Award in Science and Technology, the President Benito Juárez Medal of Merit, the Henri Willen Methorst Medal, and the Adolf Quetelet Medal.

Cross References

►Normality Tests
►Normality Tests: Power Comparison
►Omnibus Test for Departures from Normality
►Residuals

References and Further Reading

Jarque CM, Bera AK (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ Lett 6(3):255–259

Jarque CM, Bera AK (1987) A test for normality of observations and regression residuals. Int Stat Rev 55(2):163–172

Ord JK (1972) Families of frequency distributions, Griffin's statistical monographs and courses 30. Griffin, London

Jump Regression Analysis

Peihua Qiu
Professor
University of Minnesota, Minneapolis, MN, USA

Nonparametric regression analysis provides statistical tools for estimating regression curves or surfaces from noisy data. Conventional nonparametric regression procedures, however, are only appropriate for estimating continuous regression functions. When the underlying regression function has jumps, functions estimated by the conventional procedures are not statistically consistent at

the jump positions. Recently, regression analysis for estimating jump regression functions is under rapid development (Qiu 2005), which is briefly introduced here.

1-D Jump Regression Analysis

In one-dimensional (1-D) cases, the *jump regression analysis (JRA)* model has the form

$$y_i = f(x_i) + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{1}$$

where $\{y_i,\ i = 1, 2, \ldots, n\}$ are observations of the response variable y at design points $\{x_i,\ i = 1, 2, \ldots, n\}$, f is an unknown regression function, and $\{\varepsilon_i,\ i = 1, 2, \ldots, n\}$ are random errors. For simplicity, we assume that the design interval is $[0, 1]$. In (1), f is assumed to have the expression

$$f(x) = g(x) + \sum_{j=1}^{p} d_j I(x > s_j), \text{ for } x \in [0, 1], \tag{2}$$

where g is a continuous function in the entire design interval, p is the number of jump points, $\{s_j,\ j = 1, 2, \ldots, p\}$ are the jump positions, and $\{d_j,\ j = 1, 2, \ldots, p\}$ are the corresponding jump magnitudes. If $p = 0$, then f is continuous in the entire design interval. In (2), the function g is called the *continuity part* of f, and the summation $\sum_{j=1}^{p} d_j I(x > s_j)$ is called the *jump part* of f. The major goal of JRA is to estimate g, p, $\{s_j,\ j = 1, 2, \ldots, p\}$ and $\{d_j,\ j = 1, 2, \ldots, p\}$ from the observed data $\{(x_i, y_i),\ i = 1, 2, \ldots, n\}$.

A natural jump detection criterion is

$$M_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i K_1\left(\frac{x_i - x}{h_n}\right) - \frac{1}{nh_n} \sum_{i=1}^{n} Y_i K_2\left(\frac{x_i - x}{h_n}\right), \tag{3}$$

where h_n is a positive bandwidth parameter, K_1 and K_2 are two density kernel functions with supports $[0, 1]$ and $[-1, 0)$, respectively. Obviously, $M_n(x)$ is a difference of two *one-sided* kernel estimators. The first kernel estimator in equation (3) is right-sided; it is a weighted average of the observations in the right-sided neighborhood $[x, x + h_n]$. Similarly, the second kernel estimator in (3) is left-sided; it is a weighted average of the observations in the left-sided neighborhood $[x-h_n, x)$. Intuitively, $M_n(x)$ would be large if x is a jump point, and small otherwise. So, if we know that there is only one jump point (i.e., $p = 1$) in the design interval $[0, 1]$, then the jump point s_1 can be estimated by the maximizer of $|M_n(x)|$ over $x \in [h_n, 1 - h_n]$, denoted as $\widehat{s}_1$, and d_1 can be estimated by $M_n(\widehat{s}_1)$. In cases when $p > 1$ and p is known, the jump positions $\{s_j, j = 1, 2, \ldots, p\}$ and the jump magnitudes $\{d_j, j = 1, 2, \ldots, p\}$ can be estimated in a similar way. Let s_j^* be the maximizer of $|M_n(x)|$ over the range

$$x \in [h_n, 1 - h_n] \setminus \left(\bigcup_{\ell=1}^{j-1} [s_\ell^* - h_n, s_\ell^* + h_n] \right)$$

for $j = 1, 2, \ldots, p$. The ▶order statistics of $\{s_j^*, j = 1, 2, \ldots, p\}$ are denoted by $s_{(1)}^* \le s_{(2)}^* \le \ldots \le s_{(p)}^*$. Then we define $\widehat{s}_j = s_{(j)}^*$ and $\widehat{d}_j = M_n\left(s_{(j)}^*\right)$, for $j = 1, 2, \ldots, p$.

When the number of jumps p is unknown, people often use a threshold value u_n and flag all design points in $\{x_i : |M_n(x_i)| \ge u_n\}$ as candidate jumps. Then, certain deceptive candidate jumps need to be deleted using a modification procedure (cf., Qiu 2005, Sect. 3.3.3). An alternative approach is to perform a series of hypothesis tests for $H_0 : p = j$ versus $H_1 : p > j$, for $j = 0, 2, \ldots$, until the first "fail to reject H_0" (cf., Qiu 2005, Sect. 3.3.2).

The jump detection criterion $M_n(x)$ in (3) can be regarded as an estimator of the first-order derivative $f'(x)$ of f. It is based on local constant kernel estimation of the one-sided limits $f_-(x)$ and $f_+(x)$. Alternative jump detection criteria, based on other estimators of $f'(x)$ or based on estimators of both the first-order and the second-order derivatives of f, also exist. See Joo and Qiu (2009) for a recent discussion on this topic and on estimation of the continuity part g after jump points being detected.

2-D Jump Regression Analysis

In two-dimensional (2-D) cases, the regression model becomes

$$Z_i = f(x_i, y_i) + \varepsilon_i,\ i = 1, 2, \ldots, n, \tag{4}$$

where n is the sample size, $\{(x_i, y_i), i = 1, 2, \ldots, n\}$ are the design points in the design space, f is the 2-D regression function, $\{Z_i, i = 1, 2, \ldots, n\}$ are n observations of the response variable Z, and $\{\varepsilon_i, i = 1, 2, \ldots, n\}$ are random errors. For simplicity, we assume that the design space is the unit square $[0, 1] \times [0, 1]$. In such cases, jump positions of f are curves in the design space, which are called the *jump location curves (JLCs)*. Because jumps are an important structure of f, 2-D JRA is mainly for estimating JLCs and for estimating f with the jumps at the JLCs preserved, which are referred to as *jump detection* and *jump-preserving surface estimation*, respectively, in the literature (cf., Qiu 2005, Chaps. 4 and 5).

Early 2-D jump detection methods assume that the number of JLCs is known; they are usually the generalized versions of their 1-D counterparts, based on estimation of certain first-order directional derivatives of f. In Qiu and Yandell (1997), Qiu and Yandell describe the JLCs as *a pointset* in the design space, and suggest estimating the JLCs by another pointset in the same design space. Since points in a pointset need not form curves, the connection among the points of a pointset is much more flexible than the connection among the points on curves, which makes detection of arbitrary JLCs possible. For instance, Qiu and

Yandell (1997) suggest flagging a design point as a candidate jump point if the estimated gradient magnitude of f at this point is larger than a threshold. In that paper, we also suggest two modification procedures to remove certain deceptive jump candidates. Various other jump detection procedures, based on estimation of the first-order derivatives of f, or the second-order derivatives of f, or both, have been proposed in the literature. See Sun and Qiu (2007) for a recent discussion on this topic.

In the literature, there are two types of jump-preserving surface estimation methods. Methods of the first type usually estimate the surface after jumps are detected (Qiu 1998). Around the detected jumps, the surface estimator at a given point is often defined by a weighted average of the observations whose design points are located on the same side of the estimated JLC as the given point in a neighborhood of the point. Potential jumps can thus be preserved in the estimated surface. The second type of methods estimates the surface without detecting the jumps explicitly, using the so-called adaptive local smoothing. Adaptive local smoothing procedures obtain certain evidence of jumps from the observed data directly, and adapt to such evidence properly to preserve jumps while removing noise (Gijbels et al. 2006).

2-D Jump Regression Analysis and Image Processing

Model (4) can be used in cases with arbitrary 2-D design points. In certain applications (e.g., image processing), design points are regularly spaced in the 2-D design space. In such cases, a simpler model would be

$$Z_{ij} = f(x_i, y_j) + \varepsilon_{ij},\ i = 1, 2, \ldots, n_1;\ j = 1, 2, \ldots, n_2, \quad (5)$$

where $\{Z_{ij},\ i = 1, 2, \ldots, n_1;\ j = 1, 2, \ldots, n_2\}$ are observations of the response variable Z observed at design points $\{(x_i, y_j),\ i = 1, 2, \ldots, n_1;\ j = 1, 2, \ldots, n_2\}$, and $\{\varepsilon_{ij},\ i = 1, 2, \ldots, n_1;\ j = 1, 2, \ldots, n_2\}$ are random errors.

Model (5) is ideal for describing a monochrome digital image. In the setup of a monochrome digital image, x_i denotes the ith row of pixels, y_j denotes the jth column of pixels, f is the image intensity function, $f(x_i, y_j)$ is the true image intensity level at the (i, j)th pixel, ε_{ij} denotes the noise at the (i, j)th pixel, and Z_{ij} is the observed image intensity level at the (i, j)th pixel. The image intensity function f often has jumps at the outlines of objects. Therefore, 2-D JRA can provide a powerful statistical tool for image processing. In the image processing literature, positions at which f has jumps are called *step edges*, and positions at which the first-order derivatives of f have jumps are called *roof edges* (cf., Qiu 2005, Chap. 6). Edge detection and edge-preserving image restoration are two major problems in image processing, which are essentially the same problems as jump detection and jump-preserving surface estimation in 2-D JRA. See Qiu (2007) for a recent discussion about the connections and differences between the two areas.

About the Author

Dr. Peihua Qiu is a Professor, School of Statistics, University of Minnesota, USA. He is an Elected fellow of the American Statistical Association, an Elected fellow of the Institute of Mathematical Statistics, and an Elected member of the International Statistical Institute. He has authored and co-authored more than 70 papers and book chapters. His research monograph titled *Image Processing and Jump Regression Analysis* (John Wiley & Sons, 2005) won the inaugural Ziegel prize sponsored by Technometrics. Currently, he is an Associate editor for *Journal of the American Statistical Association* (2006–present) and *Technometrics* (2007–present).

Cross References

►Linear Regression Models
►Nonparametric Regression Based on Ranks
►Nonparametric Regression Using Kernel and Spline Methods

References and Further Reading

Gijbels I, Lambert A, Qiu P (2006) Edge-preserving image denoising and estimation of discontinuous surfaces. IEEE Trans Pattern Anal Mach Intell 28:1075–1087

Joo J, Qiu P (2009) Jump detection in a regression curve and its derivative. Technometrics 51:289–305

Qiu P (1998) Discontinuous regression surfaces fitting. Ann Stat 26:2218–2245

Qiu P (2005) Image processing and jump regression analysis. Wiley, New York

Qiu P (2007) Jump surface estimation, edge detection, and image restoration. J Am Stat Assoc 102:745–756

Qiu P, Yandell B (1997) Jump detection in regression surfaces. J Comput Graph Stat 6:332–354

Sun J, Qiu P (2007) Jump detection in regression surfaces using both first-order and second-order derivatives. J Comput Graph Stat 16:289–311

K

Kalman Filtering

Mohinder S. Grewal
Professor
California State University, Fullerton, CA, USA

Theoretically, a Kalman filter is an estimator for what is called the linear quadratic Gaussian (*LQG*) problem, which is the problem of estimating the instantaneous "state" of a linear dynamic system perturbed by Gaussian white noise, by using measurements linearly related to the state, but corrupted by Gaussian white noise. The resulting estimator is statistically optimal with respect to any quadratic function of estimation error. R. E. Kalman introduced the "filter" in 1960 (Kalman 1960).

Practically, the Kalman filter is certainly one of the greater discoveries in the history of statistical estimation theory, and one of the greatest discoveries in the twentieth century. It has enabled humankind to do many things that could not have been done without it, and it has become as indispensable as silicon in the makeup of many electronic systems. The Kalman filter's most immediate applications have been for the control of complex dynamic systems, such as continuous manufacturing processes, aircraft, ships, spacecraft, and satellites.

In order to control a dynamic system, one must first know what the system is doing. For these applications, it is not always possible or desirable to measure every variable that one wants to control. The Kalman filter provides a means for inferring the missing information from indirect (and noisy) measurements. In such situations, the Kalman filter is used to estimate the complete state vector from partial state measurements and is called an observer. The Kalman filter is also used to predict the outcome of dynamic systems that people are not likely to control, such as the flow of rivers during flood conditions, the trajectories of celestial bodies, or the prices of traded commodities.

Kalman filtering is an algorithm made from mathematical models. The Kalman filter makes it easier to solve a problem, but it does not solve the problem all by itself. As with any algorithm, it is important to understand its use and function before it can be applied effectively.

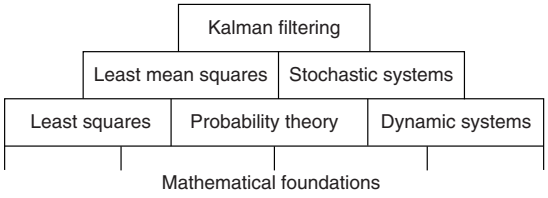
The Kalman filter is a recursive algorithm. It has been called "ideally suited to digital computer implementation," in part because it uses a finite representation of the estimation problem-by a finite number of variables (Gelb et al. 1974). It does, however, assume that these variables are real numbers with infinite precision. Some of the problems encountered in its use arise from the distinction between finite dimension and finite information, and the distinction between finite and manageable problem sizes. These are all issues on the practical side of Kalman filtering that must be considered along with the theory.

It is a complete statistical characterization of an estimation problem. The Kalman filter is much more than an estimator, because it propagates the entire probability distribution of the variables it is tasked to estimate. This is a complete characterization of the current state of knowledge of the dynamic system, including the influence of all past measurements. These probability distributions are also useful for statistical analysis and predictive design of sensor systems.

In a limited context, the Kalman filter is a learning process. It uses a model of the estimation problem that distinguishes between phenomena (what we are able to observe), noumena (what is really going on), and the state of knowledge about the noumena that we can deduce from the phenomena. That state of knowledge is represented by probability distributions. To the extent that those probability distributions represent knowledge of the real world, and the cumulative processing of knowledge is learning, this is a learning process. It is a fairly simple one, but quite effective in many applications. Figure 1 depicts the essential subjects forming the foundations for Kalman filtering theory. Although this shows Kalman filtering as the apex of a pyramid, it is but part of the foundations of another discipline-modern control theory-and a proper subset of statistical decision theory (Grewal and Andrews 2008).

Applications of Kalman filtering encompass many fields. As a tool, the algorithm is used almost exclusively for estimation and performance analysis of estimators and as observers for control of a dynamical system. Except for a few fundamental physical constants, there is hardly anything in the universe that is truly constant. The orbital parameters of the asteroid Ceres are not constant, and

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,



Kalman Filtering. Fig. 1 Foundational concepts in Kalman filtering

even the "fixed" stars and continents are moving. Nearly all physical systems are dynamic to some degree. If we want very precise estimates of their characteristics over time, then we must take their dynamics into consideration.

We do not always know the dynamics very precisely. Given this state of partial ignorance, the best we can do is express ignorance more precisely-using probabilities. The Kalman filter allows us to estimate the state of such systems with certain types of random behavior by using such statistical information. A few examples of common estimation problems are shown in Table 1. The third column lists some sensor types that we might use to estimate the state of the corresponding dynamic systems. The objective of design analysis is to determine how best to use these sensor types for a given set of design criteria. These criteria are typically related to estimation accuracy and system cost.

Because the Kalman filter uses a complete description of the probability distribution of its estimation errors to determine the optimal filtering gains, this probability distribution may be used to assess its performance as a function of the design parameters of an estimation system, such as the types of sensors to be used, the locations and orientations of the various sensor types with respect to the system to be estimated, the allowable noise characteristics of the sensors, the prefiltering methods for smoothing sensor noise, the data sampling rates for the various sensor types, and the level of model simplification to reduce implementation requirements.

This analytical capability of the Kalman filter enables system designers to assign "error budgets" to subsystems of an estimation system and to trade off the budget allocations to optimize cost or other measures of performance while achieving a required level of estimation accuracy. Furthermore, it acts like an observer by which all the states not measured by the sensors can be constructed for use in the control system applications.

Linear Estimation

Linear estimation addresses the problem of estimating the state of a linear stochastic system by using measurements or sensor outputs that are linear functions of the state. We suppose that the stochastic systems can be represented by the types of plant and measurement models (for continuous and discrete time) shown as equations in Table 2, with dimensions of the vector and matrix quantities. The measurement and plant noise v_k and w_k, respectively, are assumed to be zero-mean ▶Gaussian processes, and the initial value x_o is a Gaussian random variable with known mean x_0 and known covariance matrix P_0. Although the noise sequences w_k and v_k are assumed to be uncorrelated, this restriction can be removed, modifying the estimator equations accordingly.

A summary of equations for the discrete-time Kalman estimator are shown in Table 3, where Q_k, R_k are process and measurement noise covariances, Φ_k is the state transition matrix, H_k is the measurement sensitivity matrix, $\overline{K}_k$ is the Kalman gain. $P_k(-), P_k(+)$ are covariances before and after measurement updates.

Implementation Methods

The Kalman filter's theoretical performance has been characterized by the covariance matrix of estimation uncertainty, which is computed as the solution of a matrix Riccati differential and difference equation. A relationship between optimal deterministic control and optimal estimation problems has been described via the separation principle.

Soon after the Kalman filter was first implemented on computers, it was discovered that the observed mean-square estimation errors were often much larger than the values predicted by the covariance matrix, even with simulated data. The variances of the filter estimation errors were observed to diverge from their theoretical values, and the solutions obtained for the Riccati equations were observed to have negative variances. Riccati equations should have positive or zero variances.

Current work on the Kalman filter primarily focuses on development of robust and numerically stable implementation methods. Numerical stability refers to robustness against roundoff errors. Numerically stable implementation methods are called square root filtering because they use factors of the covariance matrix of estimation uncertainty or its inverse, called the information matrix.

Numerical solution of the Riccati equation tends to be more robust against roundoff errors if Cholesky factors of a symmetrical nonnegative definite matrix P is a matrix C such that $CC^T = P$. Cholesky decomposition algorithms solve for C that is either upper triangular or lower triangular. Another method is modified Cholesky decomposition. Here, algorithms solve for diagonal factors and either a lower triangular factor L or an upper triangular

Kalman Filtering. Table 1 Examples of estimation problems

Application	Dynamic system	Sensor types
Process control	Chemical plant	Pressure, temperature, flow rate, gas analyzer
Flood prediction	River system	Water level, rain gauge, weather radar
Tracking	Spacecraft	Radar, imaging system
Navigation	Ships	Sextant
	Aircraft, missiles	Log
	Smart bombs	Gyroscope
	Automobiles	Accelerometer
	Golf carts	Global Positioning System (GPS) receiver
	Satellites	GPS receiver
	Space shuttle	GPS receiver, Inertial Navig. Systems (INS)

Kalman Filtering. Table 2 Linear plant and measurement models

Model	Continuous time		Discrete time	
Plant	$x(t) = F(t)x(t) + w(t)$		$x_k = \Phi_{k-1}x_{k-1} + w_{k-1}$	
Measurement	$z(t) = H(t)x(t) + v(t)$		$z_k = H_k x_k + v_k$	
Plant noise	$E\langle w(t)\rangle = 0$ $E\langle w(t)w^T(s)\rangle = \delta(t-s)Q(t)$		$E\langle w_k\rangle = 0$ $E\left\langle w_k w_i^T\right\rangle = \Delta(k-i)Q_k$	
Observation noise	$E\langle v(t)\rangle = 0$ $E\langle v(t)v^T(s)\rangle = \delta(t-s)R(t)$		$E\langle v_k\rangle = 0$ $E\left\langle v_k v_i^T\right\rangle = \Delta(k-i)R_k$	
(Linear model)	Symbol	Dimensions	Symbol	Dimensions
Dimensions of vectors and matrices	x, w	$n \times 1$	Φ, Q	$n \times n$
	z, v	$\ell \times 1$	H	$\ell \times n$
	R	$\ell \times \ell$	Δ, δ	Scalar

factor U such that $P = UD_uU^T = LD_LL^T$ where D_L and D_u are diagonal factors with nonnegative diagonal elements. Another implementation method uses square root information filters that use a symmetric product factorization of the information matrix P^{-1}. Another implementation with improved numerical properties is the "sigmaRho filter." Individual terms of the covariance matrix can be interpreted as $P_{ij} = \sigma_i\sigma_j\rho_{ij}$ where P_{ij} is the ijth of the covariance matrix, σ_i is the standard deviation of the estimate of the ith state component, and ρ_{ij} is the correlation coefficient between ith and jth state component (Grewal and Kain 2010).

Alternative Kalman filter implementations use these factors of the covariance matrix (or its inverse) in three types of filter operations: (1) temporal updates, (2) observation updates, and (3) combined updates (temporal and observation). The basic algorithm methods used in these alternative Kalman filter implementations fall into four general categories. The first three of these categories are concerned with decomposing matrices into triangular factors and maintaining the triangular form of the factors through all the Kalman filtering operation. The fourth category includes standard matrix operations (multiplication, inversion, etc.) that have been specialized for triangular

Kalman Filtering. Table 3 Discrete-time Kalman filter equations

System dynamic model	$x_k = \Phi_{k-1}x_{k-1} + w_{k-1}, \quad w_k \sim N(0, Q_k)$
Measurement model	$z_k = H_k x_k = v_k, \quad v_k \sim N(0, R_k)$
Initial conditions	$E\langle x_0\rangle = \widehat{x}_0, \quad E\left\langle \widetilde{x}_0 \widetilde{x}_0^T \right\rangle = P_0$
Independence assumption	$E\left\langle w_k v_j^T \right\rangle = 0$ for all k and j
State estimate extrapolation	$\widehat{x}_k(-) = \Phi_{k-1}\widehat{x}_{k-1}(+)$
Error covariance extrapolation	$P_k(-) = \Phi_{k-1}P_{k-1}(+)\Phi_{k-1}^T + Q_{k-1}$
State estimate observational update	$\widehat{x}(+) = \widehat{x}_k(-) + \overline{K}_k[z_k - H_k\widehat{x}_k(-)]$
Error covariance update	$P_k(+) = [I - \overline{K}_k H_k]P_k(-)$
Kalman gain matrix	$\overline{K}_k = P_k(-)H_k^T\left[H_k P_k(-)H_k^T + R_k\right]^{-1}$

matrices. These implementation methods have succeeded where the conventional Kalman filter implementations have failed (Grewal and Andrews 2008).

Even though uses are being explored in virtually every discipline, research is particularly intense on successful implementation of Kalman filtering to global positioning systems (GPS), inertial navigation systems (INS), and guidance and navigation. GPS is a satellite-based system that has demonstrated unprecedented levels of positioning accuracy, leading to its extensive use in both military and civil arenas. The central problem for GPS receivers is the precise estimation of position, velocity, and time, based on noisy observations of satellite signals. This provides an ideal setting for the use of Kalman filtering. GPS technology is used in automobile, aircraft, missiles, ships, agriculture, and surveying. Currently, the Federal Aviation Agency (FAA) is sponsoring research on the development of wide-area augmentation system (WAAS) for precision landing and navigation of commercial aircraft (Grewal et al. 2007).

Kalman filters are used in bioengineering, traffic systems, photogrammetry, and myriad process controls. The Kalman filter is observer, parameter identifier in modeling, predictor, filter, and smoother in a wide variety of applications. It has become integral to twenty-first century technology (Grewal and Kain 2010; Grewal et al. 2007).

About the Author

Dr. Mohinder S. Grewal, P.E., coauthored *Kalman Filtering: Theory & Practice Using MATLAB* (with A.P. Andrews, 3rd edition, Wiley & Sons 2008) and *Global Positioning Systems, Inertial Navigation, & Integration* (with L.R. Weill and A.P. Andrews, 2nd edition, Wiley & Sons 2007). Dr. Grewal has consulted with Raytheon Systems, Geodetics, Boeing Company, Lockheed-Martin, and Northrop on application of Kalman filtering. He has published over 60 papers in IEEE and ION refereed journals and proceedings, including the Institute of Navigation's Redbook. Currently, Dr. Grewal is Professor of Electrical Engineering at California State University, Fullerton, in Fullerton, California, where he received the 2009 Outstanding Professor award. He is an architect of the GEO Uplink Subsystem (GUS) for the Wide Area Augmentation System (WAAS), including the GUS clock steering algorithms, and holds two patents in this area. His current research interest is in the area of application of GPS, INS integration to navigation. Dr. Grewal is a member of the Institute of Navigation, Senior Member of IEEE, and a Fellow of the Institute for the Advancement of Engineering.

Cross References

►Conditional Expectation and Probability
►Intervention Analysis in Time Series
►Statistical Inference for Stochastic Processes
►Structural Time Series Models

References and Further Reading

Gelb A et al (1974) Applied optimal estimation. MIT Press, Cambridge

Grewal MS, Andrews AP (2008) Kalman filtering theory and practice using MATLAB, 3rd edn. Wiley, New York

Grewal MS, Kain J (September 2010) Kalman filter implementation with improved numerical properties, Transactions on automatic control, vol 55(9)

Grewal MS, Weill LR, Andrews AP (2007) Global positioning systems, inertial navigation, & integration, 2nd edn. Wiley, New York

Kalman RE (1960) A new approach to linear filtering and prediction problems. ASME J Basic Eng 82:34–45

Kaplan-Meier Estimator

Irène Gijbels
Professor, Head of Statistics Section
Katholieke Universiteit Leuven, Leuven, Belgium

The Kaplan-Meier estimator estimates the distribution function of a lifetime T based on a sample of randomly right censored observations. In survival analysis the lifetime T is a nonnegative random variable describing the time until a certain event of interest happens. In medical applications examples of such events are the time till death of a patient suffering from a specific disease, the time till recovery of a disease after the start of the treatment, or the time till remission after the curing of a patient. A typical difficulty in survival analysis is that the observations might be incomplete. For example, when studying the time till death of a patient with a specific disease, the patient might die from another cause. As a consequence the lifetime of this patient is not observed, and is only known to be larger than the time till the patient was "censored" by the other cause of death. Such a type of censoring mechanism is called right random censorship. Other areas of applications in which one encounters this type of data are reliability in industry and analysis of duration data in economics, to just name a few.

Let $T_1, T_2, \cdots, T_n$ denote n independent and identically distributed random variables, all having the same distribution as the lifetime T. Denote by $F(t) = P\{T \le t\}$ the cumulative distribution function of T. Due to the right random censoring, the lifetime T might be censored by a censoring time C, having cumulative distribution function G. Associated at each lifetime T_i there is a censoring time C_i. Under a right random censorship model the observations consist of the pairs

$$(Z_i, \delta_i) \quad \text{where} \quad Z_i = \min(T_i, C_i) \quad \text{and} \quad \delta_i = I\{T_i \le C_i\} \qquad i = 1, \cdots, n.$$

The indicator random variable $\delta = I\{T \le C\}$ takes value 1 when the lifetime T is observed, and is 0 when the censoring time is observed instead. A crucial assumption in this model is that the lifetime T_i (also often called survival time) is independent of the censoring time C_i for all individuals. An observation (Z_i, δ_i) is called uncensored when $\delta_i = 1$ and hence the survival time T_i for individual i has been observed. When $\delta_i = 0$ the observed time is the censoring time C_i and one only has the incomplete observation that $T_i > C_i$.

Kaplan and Meier (1958) studied how to estimate the survival function $S(t) = 1 - F(t) = P\{T > t\}$, based on observations $(Z_1, \delta_1), \cdots, (Z_n, \delta_n)$ from n patients. The estimation method does not make any assumptions about a specific form of the cumulative distribution functions F and G, and is therefore a nonparametric estimate. When there are no tied observations the estimate is defined as

$$\widehat{S}(t) = \begin{cases} \prod_{j:Z_{(j)} \le t} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}} & \text{if} \quad t < Z_{(n)} \\ 0 & \text{if} \quad t \ge Z_{(n)}, \end{cases}$$

where $Z_{(1)} \le Z_{(2)} \cdots \le Z_{(n)}$ denote the ordered Z_i's, and $\delta_{(i)}$ is the indicator variable associated with $Z_{(i)}$. In case of a tie between a censored and an uncensored observation, the convention is that the uncensored observation happened just before the censored observation. An equivalent expression, for $t < Z_{(n)}$, is

$$\widehat{S}(t) = \prod_{j:Z_{(j)} \le t} \left(1 - \frac{\delta_{(j)}}{n-j+1} \right).$$

In case of n complete observations, $\delta_{(i)} = 1$ for all individuals, and the Kaplan-Meier estimate reduces to $S_n(t) = 1 - \#\{j : Z_j \le t\}/n$, i.e., one minus the empirical cumulative distribution function. The latter estimate is a decreasing step function with downward jumps of size $1/n$ at each observation $Z_j = T_j$.

Suppose now that there are tied observations, and that there are only r distinct observations. Denote by $Z_{(1)} \le Z_{(2)} \le Z_{(r)}$ the r ordered different observations, and by d_j the number of times that $Z_{(j)}$ has been observed. Then, for $t < Z_{(r)}$, the Kaplan-Meier estimate is defined as

$$\prod_{j:Z_{(j)} \le t} \left(1 - \frac{d_j}{n_j} \right)^{\delta_{(j)}},$$

where n_j denotes the number of individuals in the sample that are at risk at time point $Z_{(j)}$, i.e., the set of individuals that are still "alive" just before the time point $Z_{(j)}$. The Kaplan-Meier estimate is also called the product-limit estimate.

In studies of life tables (see ▶Life Table), the actuarial estimate for the survival function was already around much earlier. One of the first references for the product-limit estimate, obtained as a limiting case of the actuarial estimate, is Bähmer (1912).

The Kaplan-Meier estimate of the survival function $S = 1 - F$ is a decreasing step function, which jumps at the uncensored observations but remains constant when passing a censored observation. In contrast to the empirical estimate S_n based on a complete sample of size n, the sizes of the jumps are random.

The construction of the Kaplan-Meier estimate $\widehat{S}(t)$ also has a very simple interpretation due to Efron (1967). The mass $1/n$ that is normally attached to each observation in the empirical estimate for S, is now for a censored observation redistributed equally over all observations that are larger than the considered one.

Kaplan and Meier (1958) give the maximum likelihood derivation of the product-limit estimate and discuss mean and variance properties of it. An estimate for the variance of $\widehat{S}(t)$ was already established by Greenwood (1926). Greenwood's formula estimates the variance of $\widehat{S}(t)$ by

$$\widehat{\mathrm{Var}}\left(\widehat{S}(t)\right) = \left(\widehat{S}(t)\right)^2 \sum_{j:Z_{(j)} \le t} \frac{d_j}{n_j(n_j - d_j)}.$$

This estimate can be used to construct confidence intervals.

The theoretical properties of the Kaplan-Meier estimate have been studied extensively. For example, weak convergence of the process $\sqrt{n}\left(\widehat{S}(t) - S(t)\right)$ to a Gaussian process was established by Breslow and Crowley (1974), and uniform strong consistency of the Kaplan-Meier estimate was proven by Winter et al. (1978).

The Kaplan-Meier estimate is implemented in most statistical software packages, and is a standard statistical tool in survival analysis.

About the Author

Professor Irène Gijbels is Head of the Statistics Section, Department of Mathematics, of the Katholieke Universiteit Leuven, and is the Chair of the Research Commission of the Leuven Statistics Research Center. She is Fellow of the Institute of Mathematical Statistics and Fellow of the American Statistical Association. She has (co-)authored over 70 articles in internationally reviewed scientific journals.

Cross References

►Astrostatistics
►Life Table
►Mean Residual Life
►Modeling Survival Data
►Survival Data

References and Further Reading

Böhmer PE (1912) Theorie der unabhängigen Wahrscheinlichkeiten. Rapports Mémoires et Procès-verbaux de Septième Congrès International d'Actuaries. Amsterdam, vol 2. pp 327–343

Breslow N, Crowley J (1974) A large sample study of the life table and product limit estimates under random censorship. Ann Stat 2:437–453

Efron B (1967) The two sample problem with censored data. In: Proceedings of the 5th Berkeley Symposium, vol 4. pp 831–853

Greenwood M (1926) The natural duration of cancer. In: Reports on public health and medical subjects, vol 33. His Majesty's Stationery Office, London

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Winter BB, Földes A, Rejtö L (1978) Glivenko-Cantelli theorems for the product limit estimate. Probl Control Inform 7:213–225

Kappa Coefficient of Agreement

Tarald O. Kvålseth
Professor Emeritus
University of Minnesota, Minneapolis, MN, USA

Introduction

When two (or more) observers are independently classifying items or observations into the same set of k mutually exclusive and exhaustive categories, it may be of interest to use a measure that summarizes the extent to which the observers agree in their classifications. The Kappa coefficient first proposed by Cohen (1960) is one such measure.

In order to define this measure, let p_{ij} be the proportion of items assigned to category i by Observer 1 and to category j by Observer 2. Furthermore, let p_{i+} be the proportion of items assigned to category i by Observer 1 and p_{+j} the proportion of items assigned to category j by Observer 2. If these proportions or sample probabilities are represented in terms of a two-way contingency table with k rows and k columns, then p_{ij} becomes the probability in the cell corresponding to row i and column j. With row i being the same as column i for $i = 1, \ldots, k$, the diagonal of this table with probabilities p_{ii} $(i = 1, \ldots, k)$ represents the agreement probabilities, whereas the off-diagonal entries represent the disagreement probabilities.

The observed probability of agreement $P_{ao} = \sum_{i=1}^{k} p_{ii}$ could, of course, be used as an agreement measure. However, since there may be some agreement between the two observers based purely on chance, it seems reasonable that a measure of interobserver agreement should also account for the agreement expected by chance. By defining chance-expected agreement probability as $P_{ac} = \sum_{i=1}^{k} p_{i+}p_{+i}$ and based on independent classifications between the two observers, Cohen (1960) introduced the Kappa coefficient as

$$K = \frac{P_{ao} - P_{ac}}{1 - P_{ac}} \qquad (1)$$

where $K \le 1$, with $K = 1$ in the case of perfect agreement, $K = 0$ when the observed agreement probability equals that due to chance, and $K < 0$ if the observed agreement probability is less than the chance-expected one.

Kappa can alternatively be expressed in terms of the observed probability of disagreement P_{do} and the chance-expected probability of disagreement P_{dc} as

$$K = 1 - \frac{P_{do}}{P_{dc}}; \quad P_{do} = \sum_{\substack{i=1}}^{k}\sum_{\substack{j=1 \\ i \neq j}}^{k} p_{ij}, \quad P_{dc} = \sum_{\substack{i=1}}^{k}\sum_{\substack{j=1 \\ i \neq j}}^{k} p_{i+}p_{+j} \quad (2)$$

The form of K in (1) is the most frequently used one. However, it should be pointed out that the normalization used in (1), i.e., using the dominator $1 - P_{ac}$ such that $K \leq 1$, is not unique. In fact, there are infinitely many such normalizations. Thus, for any given marginal probability distributions $\{p_{i+}\}$ and $\{p_{+j}\}$, one could, for example, instead of the denominator in (1), use $\sum_{i=1}^{k}\left(\frac{1}{2}p_{i+}^{\alpha} + \frac{1}{2}p_{+i}^{\alpha}\right)^{1/\alpha} - P_{ac}$ for any real value of α. For $\alpha \to -\infty$, this alternative denominator would become $\sum_{i=1}^{k}\min\{p_{i+}, p_{+i}\} - P_{ac}$. No such non-uniqueness issue would arise by using the form of K in (2). This K also has the simple interpretation of being the proportional difference between the chance and observed disagreement probabilities, i.e., the relative extent to which P_{do} is less than P_{dc}.

Weighted Kappa

In the case when the $k > 2$ categories are ordinal, or also possibly in some cases involving nominal categories, some disagreements may be considered more serious than others. Consequently, the weighted Kappa (K_w) was introduced (Cohen 1968). In terms of the set of nonnegative agreement weights $v_{ij} \in [0,1]$ and disagreement weights $w_{ij} \in [0,1]$ for all i and j, K_w can be defined as

$$K_w = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}}{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij}p_{i+}p_{+j}} \quad (3)$$

$$= 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}} \quad (4)$$

where $w_{ij} = 1 - v_{ij}, w_{ij} = w_{ji}$ for all i and j, and $w_{ij} = 0$ for all $i = j$. Of course, when w_{ij} is the same for all $i \neq j$, K_w reduces to K in (1) – (2). From (4), which seems to be the most intuitive and preferred form of K_w, it is clear that $K_w \leq 1$, with $K_w = 1$ if, and only if, $p_{ij} = 0$ for all $i \neq j$ (i.e., if all disagreement cells have zero probability), $K_w = 0$ under independence (i.e., $p_{ij} = p_{i+}p_{+j}$ for all i and j), and K_w may also take on negative values. Unless there are particular justifications to the contrary, the most logical choice of weights would seem to be $w_{ij} = |i-j|/(k-1)$ or $w_{ij} = (i-j)^2/(k-1)^2$ for all i and j.

Specific Category Kappa

Besides measuring the overall agreement between two observers, it may be of interest to determine their extent of agreement on specific categories. As first proposed by Spitzer et al. (1967) (see also (Fleiss et al. 2003)), such measurement required the original $k \times k$ table to be collapsed into 2×2 tables, one for each specific category. Thus, to measure the agreement on a specific category s, the original $k \times k$ table would need to be collapsed into a 2×2 table with one category being the original s category and the other category being "all others". The agreement measurement K_s was then obtained by computing the value of K in (1) based on the collapsed 2×2 table.

As an alternative way of obtaining the agreement K_s on the specific category s, without the need to collapse the original $k \times k$ table, Kvålseth (1989) proposed the specific – category Kappa as

$$K_s = \frac{p_{ss} - p_{s+}p_{+s}}{\bar{p}_s - p_{s+}p_{+s}}, \quad \bar{p}_s = \frac{p_{s+} + p_{+s}}{2} \quad (5)$$

The K_s can alternatively be expressed as

$$K_s = 1 - \frac{\sum\sum_{D_s} p_{ij}}{\sum\sum_{D_s} p_{i+}p_{+j}} \quad (6)$$

where $\sum\sum_{D_s}$ denotes the summation over all disagreement cells for category s, i.e.,

$$D_s = \{(s,j) \text{ for all } j \neq s \text{ and } (i,s) \text{ for all } i \neq s\} \quad (7)$$

From (6), K_s is the proportional difference between the chance - expected disagreement and the observed disagreement for the specific category s. Note that K in (1) and (2) are weighted arithmetic means of K_s in (5) and (6), respectively, for $s = 1, \ldots, k$, with the weights being based on the denominators in (6) – (7).

When K_s is expressed as in (6), an extension to the case when disagreements should be unequally weighted is rather obvious. Thus, for disagreement weights $w_{ij} \in [0,1]$ for all i and j, with $w_{ij} = 0$ for all $i = j$, the following weighted specific – category Kappa has been proposed (Kvålseth 2003):

$$K_{ws} = 1 - \frac{\sum\sum_{D_s} w_{ij}p_{ij}}{\sum\sum_{D_s} w_{ij}p_{i+}p_{+j}} \quad (8)$$

where D_s is the set of disagreement cells in (7). When w_{ij} is the same for all $(i,j) \in D_s$, (8) reduces to (6). Note also that

K_w in (4) is a weighted arithmetic mean of the K_{ws} for $s = 1, \ldots, k$, with the weights based on the denominator in (8).

The possible values of K_s and K_{ws} range from 1 (when the disagreement probabilities for category s are all zero), through 0 (under the independence $p_{sj} = p_{s+}p_{+j}$ for all j and $p_{is} = p_{i+}p_{+s}$ for all i), and to negative values when observed disagreement exceeds chance - expected disagreement for category s.

Conditional and Asymmetric Kappa

Light (1971) considered the agreement between two observers for only those items (observations) that Observer 1 assigned to category i, with the conditional Kappa defined as

$$K_{2|1}^{(i)} = \frac{p_{ii}/p_{i+} - p_{+i}}{1 - p_{+i}} \tag{9}$$

This measure can also be expressed as

$$K_{2|1}^{(i)} = 1 - \frac{\sum_{\substack{j=1 \\ j \neq i}}^{k} p_{ij}}{\sum_{\substack{j=1 \\ j \neq i}}^{k} p_{i+}p_{+j}} \tag{10}$$

which immediately suggests the following weighted form (Kvålseth 1985):

$$K_{2|1,w}^{(i)} = 1 - \frac{\sum_{\substack{j=1 \\ j \neq i}}^{k} w_{ij}p_{ij}}{\sum_{\substack{j=1 \\ j \neq i}}^{k} w_{ij}p_{i+}p_{+j}} \tag{11}$$

Whereas K in (1) – (2) and K_w in (3) – (4) treat the two observers equivalently, i.e., these measures are effectively symmetric, asymmetric versions of Kappa may be defined in terms of the weighted means of the measures in (9) – (11) as

$$\overline{K}_{2|1} = \sum_{i=1}^{k} p_{i+}K_{2|1}^{(i)}, \; \overline{K}_{2|1,w} = \sum_{i=1}^{k} p_{i+}K_{2|1,w}^{(i)} \tag{12}$$

Such measures as in (12) may be appropriate if Observer 1 is to be designated as the "standard" against which classifications by Observer 2 are to be compared (Kvålseth 1991).

Statistical Inferences

Consider now that the above Kappa coefficients are estimates (and estimators) based on sample probabilities (proportions) $p_{ij} = n_{ij}/N$ for $i = 1, \ldots, k$ and $j = 1, \ldots, k$ and sample size $N = \sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}$, with $\{\pi_{ij}\}$ being the corresponding population probabilities. It may then be of interest to make statistical inferences, especially confidence - interval construction, about the corresponding $\{\pi_{ij}\}$ – based population coefficients or measures. Such approximate inferences can be made based on the *delta method* (Bishop et al. 1975).

Consequently, under multinomial sampling and when N is reasonably large, the various Kappa coefficients introduced above are approximately normally distributed with means equal to the corresponding population coefficients and with variances that can be determined as follows. Since those Kappa coefficients can all be expressed in terms of K_w in (4) by special choices among the set of weights $\{w_{ij}\}$, it is sufficient to determine the variance of (the estimator) K_w. For instance, in the case of K_s in (6) and K_{ws} in (8), one can simply set $w_{ij} = 0$ in (4) for all cells that do not belong to D_s in (7). Thus, the estimated variance of K_w has been given in (Kvålseth 2003) as

$$\mathrm{Var}(K_w) = \left(NF_w^2\right)^{-1}\left\{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}E_{ij}^2 - \left[K_w - (1 - F_w)(1 - K_w)\right]^2\right\} \tag{13a}$$

where

$$E_{ij} = 1 - w_{ij} - (2 - \overline{w}_{i+} - \overline{w}_{+j})(1 - K_w) \tag{13b}$$

$$\overline{w}_{i+} = \sum_{j=1}^{k} w_{ij}p_{+j}, \; \overline{w}_{+j} = \sum_{i=1}^{k} w_{ij}p_{i+}, \; F_w = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}p_{i+}p_{+j}. \tag{13c}$$

Note that F_w is the dominator of K_w in (4).

Example As an example of this inference procedure, consider the (fictitious) probabilities (proportions) in Table 1.

In the case of category 1, e.g., it follows from (5) or (6) and Table 1 that the interobserver agreement $K_1 = 0.72$.

Kappa Coefficient of Agreement. Table 1 Results from two observers' classifications with three categories

	Observer 2			
Observer 1	1	2	3	Total
1	0.40	0.07	0.01	0.48
2	0.04	0.20	0.06	0.30
3	0.02	0.05	0.15	0.22
Total	0.46	0.32	0.22	1.00

If, however, the categories in Table 1 are ordinal and the weights $w_{ij} = |i - j|/(k - 1)$ are used, it is found from (7) – (8) and Table 1, with D_1 consisting of the cells (1,2), (1,3), (2,1), and (3,1), that $K_{w1} = 0.76$. Similarly, $K_2 = 0.49, K_3 = 0.59$, $K_{w2} = 0.49$, and $K_{w3} = 0.69$, whereas, from (1) – (4), $K = 0.61$ and $K_w = 0.67$.

In order to construct a confidence interval for the population equivalent of K_{w1}, (13) can be used by (a) setting $w_{ij} = 0$ for those cells that do not belong to D_1 in (7), i.e., the cells (2,2), (2,3), (3,2) and (3,3) and (b) replacing K_w and F_w with K_{ws} and F_{ws} (the denominator of K_{ws}). Thus, with $K_{w1} = 1 - 0.0850/0.3526 = 0.7589$ (and $F_{w1} = 0.3526$), it is found from (13b) – (13c) that $E_{11} = 0.6986$, $E_{12} = 0.1673, \ldots, E_{33} = 0.7444$ so that, from (13a), if the data in Table 1 are based on sample size $N = 100$, it is found that $\mathrm{Var}(K_{w1}) = 0.0042$. Consequently, an approximate 95% confidence interval for the population equivalent of K_{w1} becomes $0.76 \pm 1.96\sqrt{0.0042}$, or $(0.63, 0.89)$. By comparison, setting $w_{12} = w_{13} = w_{21} = w_{31} = 1$ and all other $w_{ij} = 0$, it is found in (Kvålseth 2003) that a 95% confidence interval for the population equivalent of the unweighted K_1 is (0.58, 0.86).

Concluding Comment

While the overall Kappa and its weighted form in (1) – (4) are the most popular measures of interobserver agreement, they are not without some criticism or controversy. In particular, they depend strongly on the marginal distributions $\{p_{i+}\}$ and $\{p_{+j}\}$ so that, when those distributions are highly uneven (non-uniform), values of Kappa tend to be unreasonably small. Also, since the p_{ii} $(i = 1, \ldots, k)$ are included in the marginal distributions, the agreement probabilities enter into both the overall probability of agreement as observed and as expected by chance.

About the Author

For biography *see* the entry ►Entropy.

Cross References

►Measures of Agreement
►Psychiatry, Statistics in

References and Further Reading

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis, Ch. 14. MIT Press, Cambridge
Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46
Cohen J (1968) Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70:213–220
Fleiss JL, Levin B, Paik MC (2003) Statistical methods for rates and proportions, Ch. 18, 3rd edn. Wiley, Hoboken
Kvålseth TO (1985) Weighted conditional Kappa. B Psychonomic Soc 23:503–505
Kvålseth TO (1989) Note on Cohen's Kappa. Psychol Rep 65:223–226
Kvålseth TO (1991) A coefficient of agreement for nominal scales: an asymmetric version of Kappa. Educ Psychol Meas 51:95–101
Kvålseth TO (2003) Weighted specific – category Kappa measure of interobserver agreement. Psychol Rep 93:1283–1290
Light RJ (1971) Measure of response agreement for qualitative data: some generalizations and alternatives. Psychol Bull 76:365–377
Spitzer RL, Cohen J, Fleiss JL, Endicott J (1967) Quantification of agreement in psychiatric diagnosis. Arch Gen Psychiat 17:83–87

Kendall's Tau

Llukan Puka
Professor
University of Tirana, Tirana, Albania

K

Kendall's *Tau* is a nonparametric measure of the degree of correlation. It was introduced by Maurice Kendall in 1938 (Kendall 1938).

Kendall's *Tau* measures the strength of the relationship between two ordinal level variables. Together with Spearman's rank correlation coefficient, they are two widely accepted measures of rank correlations and more popular rank correlation statistics.

It is required that the two variables, X and Y, are paired observations. Then, provided both variables are at least ordinal, it would be possible to calculate the correlation between them. In general, application of the product-moment correlation coefficient is limited by the requirement that the trend must be linear. A less restrictive measure of correlation is based on the probabilistic notion that the correlation between variables X and Y is strong if on average, there is a high probability that an increase in X will be accompanied by an increase in Y (or decrease in Y). Then the only limitation imposed on the trend line is that it should be either continually increasing or continually decreasing.

One of the properties of coefficients that adopt this notion of correlation, like *Kendall's Tau* coefficient, is that the definition of the correlation depends only on the ranks of the data values and not on the numerical values. To this end, they can be applied either to data from scaled variables that has been converted to ranks, or to ordered categorical variables.

Formula for Calculation of Kendall's Tau Coefficient, (Hollander and Wolfe 1998)

For any sample of n paired observations of a bivariate variables (X, Y), there are $m = \frac{n(n-1)}{2}$ possible comparisons of points (X_i, Y_i) and (X_j, Y_j). A pair of observation data set (X_i, Y_i), (X_j, Y_j) is called concordant if $X_j - X_i$ and $Y_j - Y_i$ has the same sign. Otherwise, if they have opposite signs, the pair is called discordant. If $X_i = X_j$, or $Y_i = Y_j$ or both, the comparison is called a "tie." Ties are not counted as concordant or discordant.

If C is the number of pairs that are concordant and D is the number of pairs that are discordant, then the value *Tau* of Kendall's *Tau* is

$$Tau = \frac{C - D}{m}$$

The quantity $S = C - D$ is known as Kendall S. A predominance of concordant pairs resulting in a large positive value of S indicates a strong positive relationship between X and Y; a predominance of discordant pairs resulting in a large negative value of S indicates a strong negative relationship between X and Y.

The denominator m is a normalizing coefficient such that the *Kendall's Tau* coefficient can assume values between −1 and +1: $-1 \le Tau \le 1$.

The interpretation of *Kendall's Tau* value is similar as for the other correlation coefficients: when the value is +1, then the two rankings are the same (the concordance between two variables is perfect); when the value is −1, the discordance is perfect (the ranking of one of variables is reverse to the other); and finally any other value between −1 and +1 is interpreted as a sign of the level of relationship, a positive relationship (*Kendall's Tau* > 0, both variables increase together), or a negative relationship (*Kendall's Tau* < 0, the rank of one variable increases, the other one decreases); the value 0 is an indication for non relationship.

If there are a large number of ties, then the dominator has to be corrected by $\sqrt{(m - n_x)(m - n_y)}$ where n_x is the number of ties involving X and n_y is the number of ties involving Y.

For inferential purposes, *Kendall's Tau* coefficient is used to test the hypothesis that X and Y are independent, $Tau = 0$, against one of the alternatives: $Tau \neq 0$, $Tau > 0$, $Tau < 0$. Critical values are tabulated, Daniel (1990), Abdi (2007). The problem of ties is considered also by Sillitto (1947), Burr (1960).

In large samples, the statistic

$$\frac{3 \times Kendall's\ Tau \times \sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

has approximately a normal distribution with mean 0 and standard deviation 1, and therefore can be used as a test statistic for testing the null hypothesis of zero correlation. It can be used also to calculate the confidence intervals (Noether 1967).

Kendall's *Tau* and Spearman's *Rho*

Kendall's *Tau* is equivalent to Spearman's *Rho*, with regard to the underlying assumptions. But they differ in their underlying logic and also computational formulas are quite different. The relationship between the two measures is given by

$$-1 \le \{(3 \times \text{Kendall's } Tau) - (2 \times \text{Spearman's } Rho)\} \le +1.$$

Their values are very similar in most cases, and when discrepancies occur, it is probably safer to interpret the lower value. More importantly, Kendall's *Tau* and Spearman's *Rho* imply different interpretations. Spearman's *Rho* is considered as the regular Pearson's correlation coefficient in terms of the proportion of variability accounted for, whereas Kendall's *Tau* represents a probability, i.e., the difference between the probabilities that the observed data are in the same order versus the probability that the observed data are not in the same order.

The distribution of Kendall's *Tau* has better statistical properties. In most of the situations, the interpretations of Kendall's *Tau* and Spearman's rank correlation coefficient are very similar and thus invariably lead to the same inferences. In fact neither statistics has any advantage in being easier to apply (since both are freely available in statistical packages) or easier to interpret. However Kendall's statistics structure is much simpler than that of the Spearman coefficient and has the advantage that it can be extended to explore the influence of a third variable on the relationship.

There are two different variations of Kendall's *Tau* that make adjustment for ties: *Tau b* and *Tau c*. These measures differ only as to how tied ranks are handled.

Kendall's *Tau-b*

Kendall's *Tau-b* is a nonparametric measure of correlation for ordinal or ranked variables that take ties into account. The sign of the coefficient indicates the direction of the relationship, and its absolute value indicates the strength, with larger absolute values indicating stronger relationships. Possible values ranges from −1 to 1. The calculation formula for *Kendall's Tau-b* is given by the following:

$$Tau - b = \frac{C - D}{\sqrt{(C + D + X_0)(C + D + Y_0)}}$$

where X_0 is the number of pairs tied only on the X variable, Y_0 is the number of pairs tied only on the Y variable. When

there are no ties, the values of Kendall's *Tau* and Kendall's *Tau b* are identical.

The *Kendall's Tau-b* has properties similar to the properties of the *Spearman Rho*. Because it does estimate a population parameter, many statisticians prefer the Kendall's *Tau-b* to the Spearman rank correlation coefficient.

Kendall's *Tau-c*

Kendall's *Tau-c*, is a variant of *Tau-b* used for situations of unequal-sized sets of ordered categories. It equals the excess of concordant over discordant pairs, multiplied by a term representing an adjustment for the size of the table. It is also called *Kendall–Stuart Tau-c* (or Stuart's Tau-c) and is calculated by formula

$$Tau - c = \frac{2m \times (C - D)}{n^2(m-1)}$$

where m is the smaller of the number of rows and columns, and n is the sample size.

Kendall's *Tau-b* and Kendall's *Tau-c* are superior to other measures of ordinal correlation when a test of significance is required.

About the Author

Dr. Llukan Puka is a Professor, Department of Mathematics, Faculty of Natural Science, University of Tirana, Albania. He was the Head of Statistics and Probability Section at the Faculty, Head of Mathematics Department too. During 1997–2007, he was the Dean of the Faculty. Professor Puka is an Elected Member of the ISI, associated to the IASC. He has authored and coauthored more than 30 papers and 15 textbooks, university level and high school level, mainly concerning probability, statistics, and random processes. He was the Head of National Council of Statistics (2001–2005). Professor Puka is the President of the Albanian Actuarial Association.

Cross References

►Copulas: Distribution Functions and Simulation
►Frailty Model
►Measures of Dependence
►Nonparametric Statistical Inference
►Sequential Ranks
►Statistics on Ranked Lists
►Tests of Independence
►Validity of Scales

References and Further Reading

Abdi H (2007) The Kendall rank correlation coefficient. In: Salkin NJ (ed) Encyclopedia of measurement and statistics. Sage, Thousand Oaks

Burr EJ (1960) The distribution of Kendall's score *S* for a pair of tied rankings. Biometrika 47:151–171

Daniel WW (1990) Applied nonparametric statistics, 2nd edn. PWS-KENT Publishing Company, Boston

Hollander M, Wolfe DA (1998) Nonparametric statistical methods, 2nd edn. Wiley, New York

Kendall M (1938) A new measure of rank correlation. Biometrika 30:81–89

Noether GE (1967) Elements of nonparametric statistics. Wiley, New York

Sillitto GP (1947) The distribution of Kendall's coefficient of rank correlation in rankings containing ties. Biometrika 34:36–40

Khmaladze Transformation

Hira L. Koul, Eustace Swordson
Professor and Chair, President of the Indian Statistical Association
Michigan State University, East Lansing, MI, USA

Background

Consider the problem of testing the null hypothesis that a set of random variables $X_i, i = 1, \ldots, n$, is a random sample from a specified continuous distribution function (d.f.) F. Under the null hypothesis, the empirical d.f.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \le x\}$$

must "agree" with F. One way to measure this agreement is to use omnibus test statistics from the empirical process (see ►Empirical Processes)

$$v_n(x) = \sqrt{n}(F_n(x) - F(x)).$$

The time transformed uniform empirical process

$$u_n(t) = v_n(x), \quad t = F(x)$$

is an empirical process based on random variables $U_i = F(X_i), i = 1, \ldots, n$, that are uniformly distributed on $[0,1]$ under the null hypothesis. Hence, although the construction of u_n depends on F, the null distribution of this process does not depend on F any more (Kolmogorov (1933), Doob (1949)). From this sprang a principle, universally accepted in goodness of fit testing theory, that one should choose tests of the above hypothesis based on statistics $A(v_n, F)$ which can be represented as statistics $B(u_n)$ just from u_n. Any such statistic, like, for example, weighted Cramér-von Mise statistics $\int v_n^2(x)\alpha(F(x))dF(x)$, or Kolmogorov-Smirnov statistics $\max_x |v_n(x)|/\alpha(F(x))$, will have a null distribution free

from F, and hence this distribution can be calculated once and used for many different F – still a very desirable property in present times, in spite of great advantages in computational power. It is called the distribution free property of the test statistic.

However, as first clarified by Gikhman (1954) and Kac et al. (1955), this property is lost even asymptotically as soon as one is fitting a family of parametric d.f.'s. More precisely, suppose one is given a parametric family of d.f.'s F_θ, θ a k-dimensional Euclidean parameter, and one wishes to test the hypothesis that $X_i, i = 1, \dots, n$, is a random sample from some F_θ. Denoting $\hat{\theta}_n$ a $n^{1/2}$-consistent estimator of θ, the relevant process here is the parametric empirical process

$$\hat{v}_n(x) = \sqrt{n}(F_n(x) - F_{\hat{\theta}_n}(x)).$$

To describe the effect of estimation of θ on $\hat{v}_n$, let $\dot{F}_\theta(x) = \partial F_\theta(x)/\partial\theta$ and y^T denote the transpose of a k-vector y. Under simple regularity conditions,

$$\begin{aligned}\hat{v}_n(x) &= \sqrt{n}(F_n(x) - F_{\hat{\theta}_n}(x)) \\ &= v_n(x) - \dot{F}_\theta(x)^T\sqrt{n}(\hat{\theta}_n - \theta) + o_P(1).\end{aligned}$$

If additionally, for a k-vector of square integrable functions ψ,

$$\sqrt{n}(\hat{\theta}_n - \theta) = \int \psi dv_n + o_P(1),$$

then $\hat{v}_n$ converges weakly to a mean zero Gaussian process $\hat{v}$, different from the weak limit of v_n, with a covariance function that depends on the unknown parameter θ via F_θ and ψ in a complicated fashion (Durbin (1973), Khmaladze (1979)). Critical values of any test based on this process are difficult to find even for large samples. Thus the goodness of fit testing theory was in danger of being fragmented into large number of particular cases and becoming computationally heavy and complex.

Khmaladze Transformation

To overcome this shortcoming, Khmaladze devised a transformation of $\hat{v}_n$ whose asymptotic null distribution under the parametric null hypothesis is distribution free while at the same time this transformed process stays in one-to-one correspondence with the process $\hat{v}_n$ without the loss of any "statistical information."

To describe this transformation, let f_θ denote density of F_θ and $\psi_\theta = \partial \log f_\theta/\partial\theta$ and let v denote the limit in distribution of empirical process v_n. Equip the process $\hat{v}$ with filtration $\mathcal{H} = \{\mathcal{H}_x, -\infty < x < \infty\}$, where each σ-field $\mathcal{H}_x = \sigma\{v(y), y \le x, \int \psi_\theta dv\}$ is generated not only by the "past" of v but also $\int \psi_\theta dv$, which contains a "little bit of a future" as well. This filtration is not an intrinsic part of the testing problem as it is usually formulated in statistics. Nevertheless, Khmaladze (1981) suggested to use it, because then it is natural to speak about martingale part $\{w, \mathcal{H}\}$ of the resulting semi-martingale $\{\hat{v}, \mathcal{H}\}$. Let $h_\theta^T(x) = (1, \psi_\theta(x))$ be "extended" score function and let $\Gamma_{x,\theta}$ be covariance matrix of $\int_x h_\theta dv$. Then this martingale part has the form

$$w(x) = v(x) - \int^x h_\theta(y)\Gamma_{y,\theta}^{-1}\int_y h_\theta dv\, dF_\theta(y). \qquad (1)$$

The change of time $t = F_\theta(x)$ will transform it to a standard Brownian motion (see ▶Brownian Motion and Diffusions) on $[0,1]$ – a convenient limiting process, with the distribution independent from F_θ. The substitution of $\hat{v}_n$ in (1) produces a version of empirical process w_n, which, basically, is the Khmaladze transform (KhT hence forth). It was shown to possess the following asymptotic properties: it will not change, regardless of which function ψ, or which estimator $\hat{\theta}_n$, was used in $\hat{v}_n$; it stays in one-to-one correspondence with $\hat{v}_n$, if $\hat{\theta}_n$ is the maximum likelihood estimator; and also the centering of empirical distribution function F_n in empirical process is unnecessary. Hence, the final form of KhT for parametric hypothesis is

$$w_{n,\theta}(x) = \sqrt{n}\left[F_n(x) - \int^x h_\theta(y)\Gamma_{y,\theta}^{-1}\int_y h_\theta dF_n\, dF_\theta(y)\right].$$

If the hypothesis is true, after time transformation $t = F_\theta(x)$, the processes $w_{n,\theta}$ and $w_{n,\hat{\theta}_n}$ converge weakly to standard Brownian motion. Consequently a class of tests based on time transformed $w_{n,\hat{\theta}_n}$ are asymptotically distribution free.

A slightly different point of view on $w_{n,\theta}$ is that its increment

$$dw_{n,\theta}(x) = \sqrt{n}\left[dF_n(x) - h_\theta(x)\Gamma_{x,\theta}^{-1}\int_x h_\theta dF_n\, dF_\theta(x)\right]$$

is (normalized) difference between $dF_n(x)$ and its linear regression on $F_n(x)$ and $\int_x h_\theta dF_n$.

If θ is known, i.e., if the hypothesis is simple, then $w_{n,\theta}$ reduces to what is called in the theory of empirical processes the basic martingale (see, e.g., Shorack and Wellner (1986)).

It is well known that the analog of Kolmogorov test is not distribution free when fitting a multivariate d.f. Khmaladze (1988, 1993) developed an analog of KhT in this case also, using the notion of so called scanning martingales.

Tsigroshvili (1998), and in some cases Khmaladze and Koul (2009), show that the KhT is well defined even if the matrix $\Gamma_{x,\theta}$ is not of full rank.

Some power properties of tests based on the $w_{n,\hat{\theta}_n}$ were investigated in a number of publications, including Janssen & Ünlü (2008), Koul and Sakhanenko (2005) and Nikitin (1995).

The specific form of $w_{n,\theta}$ and the practicality of its use for some particular parametric families was studied, e.g., in Koul and Sakhanenko (2005) and Haywood and Khmaladze (2007).

KhT for Counting Processes

If $N(t), t \geq 0$, is a point process (see ►Point Processes) then Aalen (1978) used an appropriate filtration and the corresponding random intensity function $\lambda(t)$ to create the martingale

$$M(t) = N(t) - \int_0^t \lambda(s)ds.$$

This in turn gave rise to broad and successful theory, especially in survival analysis with randomly censored observations, as explained in the monograph by Andersen et al. (1993). However, if the $\lambda = \lambda_\theta$ depends on unspecified parameter, which needs to be estimated using N itself, then the process $\hat{M}(t)$ is not a martingale any more and suffers from the same problems as the process $\hat{v}_n$.

Again, by including the estimator $\hat{\theta}$ in the filtration used, the KhT for $\hat{M}(t)$ was constructed in Maglaperidze et al. (1998), Nikabadze and Stute (1997), and later in O'Quigley (2003), Sun et al. (2001) and Scheike and Martinussen (2004).

KhT in Regression

The transformation was taken into new direction of the quantile regression problems in Koenker and Xiao (2002), where some additional problems were resolved. The practicality of the approach was demonstrated by the software, created by Roger Koenker and his colleagues. Recent extension to the case of autoregression is presented in discussion paper Koenker and Xiao (2006).

In the classical mean regression set up with covariate X and response Y, $Y = \mu(X) + \epsilon$, where error ϵ is independent of X, $E\epsilon = 0$, and $\mu(x) = E(Y|X = x)$. Let (X_i, Y_i), $i = 1, \cdots, n$, be a random sample from this model.

Here the two testing problems are of interest. One is the goodness-of-fit of an error d.f. and the second is the problem of lack-of-fit of a parametric regression function $m_\theta(x)$. In parametric regression model, tests for the first problem are based on the residual empirical process $\hat{v}_n(x)$ of the residuals $\hat{\epsilon}_i = Y_i - m_{\hat{\theta}_n}(X_i)$, $i = 1, \cdots, n$, where $\hat{\theta}_n$ is a $n^{1/2}$-consistent estimator of θ. Khmaladze and Koul (2004) develops the KhT of $\hat{v}_n$. Similar results were obtained for nonparametric regression models in Khmaladze and Koul (2009). It is shown, somewhat unexpectedly, that in nonparametric regression models, KhT not only leads to an asymptotically distribution free process, but also tests based on it have larger power than the tests based on $\hat{v}_n$ with non-parametric residuals $Y_i - \hat{m}_n(X_i)$.

Tests of lack-of-fit are typically based on the partial sum processes of the form

$$\sum_{i=1}^{n} g(\hat{\epsilon}_i) I\{X_i \leq x\},$$

for some known function g. However, again their limiting distribution depend on the form of the regression function, on the estimator $\hat{\theta}_n$ used and on the particular value of the parameter. Starting with Stute et al. (1998) this tradition was changed and KhT was introduced for these partial sum processes, which again, led to the process converging to standard Brownian motion. Khmaladze and Koul (2004) studied the analog of KhT for partial sum process when design variable is multi-dimensional.

Extension to some time series models are discussed in Koul and Stute (1999), Bai (2003) and Koul (2006). Koul and Song (2008, 2009, 2010), Dette and Hetzler (2008, 2009) illustrate use of KhT in some other problems in the context of interval censored data, Berkson measurement error regression models and fitting a parametric model to the conditional variance function.

About the Author

Dr. Hira Lal Koul is Professor and Chair, Department of Statistics and Probability, Michigan State University. He was President of the International Indian Statistical Association (2005–2006). He was awarded a Humboldt Research Award for Senior Scientists (1995). He is a Fellow of the American Statistical Association, Institute of Mathematical Statistics, and Elected member of the International Statistical Institute. He is Co-Editor in Chief of *Statistics and Probability Letters*, Associate Editor for *Applicable Analysis and Discrete Mathematics*, and Co-Editor of *The J. Mathematical Sciences* he has (co-)authored about 110 papers, and several monographs and books. He is joint editor of the text *Frontiers in Statistics* (with Jianqing Fan, dedicated to Prof. Peter J Bickel in honor of his 65th birthday, Imperial College Press, London, UK, 2006).

Cross References

►Empirical Processes
►Martingales
►Point Processes

References and Further Reading

Aalen OO (1978) Nonparametric inference for a family of counting processes. Ann Stat 6:701–726

Anderson TW, Darling DA (1952) Asymptotic theory of certain "goodness of fit" criteria based on stochastic porcesses. Ann Math Stat 23:193–212

Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Bai J (2003) Testing parametric conditional distributions of dynamic models. Rev Econ Stat 85:531–549

Dette H, Hetzler B (2008) A martingale-transform goodness-of-fit test for the form of the conditional variance. http://arXiv.org/abs/0809.4914?context=stat

Dette H, Hetzler B (2009) Khmaladze transformation of integrated variance processes with applications to goodness-of-fit testing. Math Meth Stat 18:97–116

Doob JL (1949) Heuristic approach to the Kolmogorov-Smirnov theorems. Ann Math Stat 20:393–403

Durbin J (1973) Weak convergence of the sample distribution function when parameters are estimated. Ann Statist 1:279–290

Gikhman II (1954) On the theory of ω^2 test. Math Zb Kiev State Univ 5:51–59

Haywood J, Khmaladze EV (2007) On distribution-free goodness-of-fit testing of exponentiality. J Econometrics 143:5–18

Janssen A, Ünlü H (2008) Regions of alternatives with high and low power for goodness-of-fit tests. J Stat Plan Infer 138:2526–2543

Kac M, Kiefer J, Wolfowitz J (1955) On tests of normality and other tests of goodness of fit based on distance methods. Ann Math Stat 26:189–211

Khmaladze EV (1979) The use of $\omega 2$ tests for testing parametric hypotheses. Theor Probab Appl 24(2):283–301

Khmaladze EV (1981) Martingale approach in the theory of goodness-of-fit tests. Theor Probab Appl 26:240–257

Khmaladze EV (1988) An innovation approach to goodness-of-fit tests in R^m. Ann Stat 16:1503–1516

Khmaladze EV (1993) Goodness of fit problem and scanning innovation martingales. Ann Stat 21:798–829

Khmaladze EV, Koul HL (2004) Martingale transforms goodness-of-fit tests in regression models. Ann Stat 32:995–1034

Khmaladze EV, Koul HL (2009) Goodness of fit problem for errors in non-parametric regression: distribution free approach. Ann Stat 37:3165–3185

Koenker R, Xiao Zh (2002) Inference on the quantile regression process. Econometrica 70:1583–1612

Koenker R, Xiao Zh (2006) Quantile autoregression. J Am Stat Assoc 101:980–990

Kolmogorov A (1933) Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari 4: 83–91

Koul HL, Stute W (1999) Nonparametric model checks for time series. Ann Stat 27:204–236

Koul HL, Sakhanenko L (2005) Goodness-of-fit testing in regression: A finite sample comparison of bootstrap methodology and Khmaladze transformation. Stat Probabil Lett 74: 290–302

Koul HL (2006) Model diagnostics via martingale transforms: a brief review. In: Frontiers in statistics, Imperial College Press, London, pp 183–206

Koul HL, Yi T (2006) Goodness-of-fit testing in interval censoring case 1. Stat Probabil Lett 76(7):709–718

Koul HL, Song W (2008) Regression model checking with Berkson measurement errors. J Stat Plan Infer 138(6):1615–1628

Koul HL, Song W (2009) Model checking in partial linear regression models with berkson measurement errors. Satistica Sinica

Koul HL, Song W (2010) Conditional variance model checking. J Stat Plan Infer 140(4):1056–1072

Maglaperidze NO, Tsigroshvili ZP, van Pul M (1998) Goodness-of-fit tests for parametric hypotheses on the distribution of point processes. Math Meth Stat 7:60–77

Nikabadze A, Stute W (1997) Model checks under random censorship. Stat Probabil Lett 32:249–259

Nikitin Ya (1995) Asymptotic efficiency of nonparametric tests. Cambridge University Press, Cambridge, pp xvi+274

O'Quigley J (2003) Khmaladze-type graphical evaluation of the proportional hazards assumption. Biometrika 90(3): 577–584

Scheike TH, Martinussen T (2004) On estimation and tests of time-varying effects in the proportional hazards model. Scand J Stat 31:51–62

Shorack GR, Wellner JA (1986) Empirical processes with application to statistics. Wiley, New York

Stute W, Thies S, Zhu Li-Xing (1998) Model checks for regression: an innovation process approach. Ann Stat 26:1916–1934

Sun Y, Tiwari RC, Zalkikar JN (2001) Goodness of fit tests for multivariate counting process models with applications. Scand J Stat 28:241–256

Tsigroshvili Z (1998) Some notes on goodness-of-fit tests and innovation martingales (English. English, Georgian summary). Proc A Razmadze Math Inst 117:89–102

Kolmogorov-Smirnov Test

Raul H. C. Lopes
Research Fellow
Brunel University, Uxbridge, UK
Professor of Computer Science at UFES, Vitoria
Brazil

Applications of Statistics are frequently concerned with the question of whether two sets of data come from the same distribution function, or, alternatively, of whether a probabilistic model is adequate for a data set. As an example, someone might be interested in evaluating the quality of a computer random numbers generator, by testing if the sample is uniformly distributed. A test like that is generally called a goodness-of-fit test. Examples of it are the χ^2 test and the Kolmogorov-Smirnov test.

Generally given a sample $X = x_0, x_1, \ldots, x_{n-1}$ and a probability distrbution function $P(x)$ the target would be

to test the Null Hypothesis H_0 that P is the sample's distribution function. When testing with two data sets the Null Hypothesis H_0 states that they both have the same distribution functions.

The choice of a statistical test must take into account at least two factors: (1) whether the data is continuous or discrete, (2) and if the comparison to be performed uses two data sets or a one set against a fitting probability model. Testing that a set of computer generated pseudo-random real numbers follows a uniformly distributed model is an example of testing a continuous data set against a probabilistic model, while comparing the amount of Vitamin C in two different brands of orange juice would fit a comparison of two continuous data sets.

The χ^2 test was designed to test discrete data sets against a probability model. However, it could be applied in the test of the computer random numbers generator by discretising the sample. Given the set $X = x_0, x_1, \ldots, x_{n-1}$ of generated numbers, a set of k intervals (bins)

$$(-\infty, z_1), (z_1, z_2), \ldots, (z_{k-1}, \infty)$$

could be used to define a discrete function

$$X_j = i, \text{ when } x_j \in (z_{i-1}, z_i).$$

Kolmogorov (1933) and Smirnov (1948) proved a result, also Schmid (1958), that is the basis for a much more efficient goodness-of-fit test when continuous data is involved. The test starts with the definition of a function $F_{X,n}(x)$ that gives the fraction of points $x_i, i \in (0, \ldots, n-1)$, in a sample X that are below x as follows (E.W. Dijkstra's uniform notation for quantifiers is used, with $\#\, i : P(x_i)$ denoting the number of elements in the set satisfying the property $P(x_i)$, for all possible i.):

$$F_{X,n}(x) = \frac{\#i : x_i \leq x}{n}$$

Assuming that another sample $Y = y_0, y_1, \ldots, y_{m-1}$ is given, then its function can be defined:

$$F_{Y,m}(y) = \frac{\#i : y_i \leq y}{m}$$

And any statistic could be used to measure the difference between X and Y, by measuring the difference between $F_{X,n}(x)$ and $F_{Y,m}(x)$. Even the area between the curves defined by these functions could be used. The Kolmogorov-Smirnov distance, is defined as the maximum absolute value of the difference between $F_{X,n}(x)$ and $F_{Y,m}(x)$ for all possible values of x:

$$D = max\, x : -\infty < x < \infty : F_{X,n}(x) - F_{Y,m}(x)$$

In a test trying to fit one sample with a probabilistic model defined by the function $P(x)$, the distance, also called Kolmogorov-Smirnov statistic, would be defined as

$$D = max\, x : -\infty < x < \infty : F_{X,n}(x) - P(x)$$

The distribution of the Kolmogorov-Smirnov statistic in the case of a Null Hypothesis test can be computed, giving a significance level for the observed value. For that purpose, let $D^\star$ be the following function of the observed value:

$$D^\star(d) = \left[\sqrt{n_e} + 0.12 + 0.11/\sqrt{n_e}\right] d$$

In the definition of $D^\star$, the quantity n_e is defined as follows:

- n_e is the number of points in the sample, when doing a one-sample test.
- $n_e = \frac{n*m}{n+m}$, in the case of a two-sample test, with n and m being the sizes of the samples.

The significance level can then be computed using the function Q below (Stephens 1970):

$$Q(d) = 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2d^2}$$

Given a d, computed by the Kolmogorov-Smirnov distance, the significance level of d, which comes to be the probability that the null hypotheses (that the two distributions are the same) is invalid, is given by

$$\text{Probability}(D > d) = Q(D^\star(d))$$

The Kolmogorov-Smirnov test offers several advantages over the χ^2 test:

- It can be applied to continuous data.
- The distribution of its statistic is invariant under re-parametrisation and it can be easily implemented by computers.
- It can be extended to multivariate data.

Several statistics packages implement the Kolmogorov-Smirnov test. The package **R** (Crawley 2007), freely available (Software and documentation from http://www.r-project.org) for most operating systems, offers a Kolmogorov-Smirnov test in the function *ks.test*.

Adapting goodness-of-fit tests to multivariate data is considered a challenge. In particular, tests based on binning suffer from what has been described as the "curse of multi-dimensionality": the multi-dimensional space is essentially empty and binning tests tend to be ineffective even with large data sets.

Peacock in (Peacock 1983) introduced an extension of the Kolmogorov-Smirnov test to multivariate data. The idea consists in taking into account the distribution function of the two samples in all possible orderings, $2^d - 1$ orderings when d dimensional data is being considered. Given n points, in a two-dimensional space, Peacock proposed to compute the distribution functions in the $4n^2$ quadrants of the plane defined by all pairs (x_i, y_i), x_i and y_i being coordinates of all points of two given samples. This gives an algorithm of $\Omega(n^3)$ complexity. Fasano e Franceschini introduced in (Fasano and Franceschini 1987) an approximation of the Peacock's test that computes the statistic over all quadrants centred in each point of the given samples. Their test can be computed in time $\Omega(n^2)$. Lopes et alii introduced an algorithm (Available, under GPL license, from http://www.inf.ufes.br/ raul/cern.2dks.tar.bz2) based on range-counting trees that computes this last statistic in $O(n \lg n)$, which is a lower-bound for the test (Lopes et al. 2008).

Cross References

►Chi-Square Tests
►Cramér-Von Mises Statistics for Discrete Distributions
►Normality Tests
►Normality Tests: Power Comparison
►Parametric and Nonparametric Reliability Analysis
►Tests of Fit Based on The Empirical Distribution Function

References and Further Reading

Crawley MJ (2007) The R book. Wiley, Chichester
Fasano G, Franceschini A (1987) A multi-dimensional version of the Kolmogorov-Smirnov test. Monthly Notices of the Royal Astronomy Society 225:155–170
Kolmogorov AN (1933) Sulla determinazione empirica di una legge di distribuzione. Giornale dell' Instituto Italiano degli Attuari 4:83–91
Lopes RHC, Hobson PR, Reid I (2008) Computationally efficient algorithms for the two-dimensional Kolmogorov-Smirnov test. Conference Series, Journal of Physics, p 119
Peacock JA (1983) Two-dimensional goodness-of-fit in Astronomy. Monthly Notices of the Royal Astronomy Society 202: 615–627
Schmid P (1958) On the Kolmogorov and Smirnov limit theorems for discontinuous distribution functions. Ann Math Stat 29(4):1011–1027
Smirnov N (1948) Table for estimating the goodness of fit of empirical distributions. Ann Math Stat 19(2):279–281
Stephens MA (1970) Use of the Kolmogorov-Smirnov, Cramér-von Mises and related statistics without extensive tables. J R Stat Soc 32:115–122

Kullback-Leibler Divergence

James M. Joyce
Chair and Professor of Philosophy and of Statistics
University of Michigan, Ann Arbor, MI, USA

Kullback-Leibler divergence (Kullback and Leibler 1951) is an information-based measure of disparity among probability distributions. Given distributions P and Q defined over X, with Q absolutely continuous with respect to P, the *Kullback-Leibler divergence* of Q from P is the P-expectation of $-\log_2\{P/Q\}$. So, $D_{KL}(P,Q) = -\int_X \log_2(Q(x)/P(x))dP$. This quantity can be seen as the difference between the *cross-entropy for Q on P*, $H(P,Q) = -\int_X \log_2(Q(x))dP$, and the *self-entropy* (Shannon 1948) of P, $H(P) = H(P,P) = -\int_X \log_2(P(x))dP$. Since $H(P,Q)$ is the P-expectation of the number of bits of information, beyond those encoded in Q, that are needed to identify points in X, $D_{KL}(P,Q) = H(P) - H(P,Q)$ is the expected difference, from the perspective of P, between the information encoded in P and the information encoded in Q.

D_{KL} has a number of features that make it plausible as a measure of probabilistic divergence. Here are some of its key properties:

Premetric. $D_{KL}(P,Q) \geq 0$, with identity if and only if $P = Q$ a.e. with respect to P.

Convexity. $D_{KL}(P,Q)$ is convex in both P and Q.

Chain Rule. Given joint distributions $P(x,y)$ and $Q(x,y)$, define the KL-divergence conditional on x as $D_{KL}(P(y|x), Q(y|x)) = \int_X D_{KL}(P(y|x), Q(y|x))dP_x$ where P_x is P's x-marginal. Then,

$$D_{KL}(P(x,y), Q(x,y)) = D_{KL}(P_x, Q_x) + D_{KL}(P(y|x), Q(y|x)).$$

Independence. When X and Y are independent in both P and Q the Chain Rule assumes the simple form $D_{KL}(P(x,y), Q(x,y)) = D_{KL}(P_x, Q_x) + D_{KL}(P_y, Q_y)$, which reflects the well-known idea that independent information is additive.

It should be emphasized that KL-divergence is not a genuine metric: it is not symmetric and fails the triangle inequality. Thus, talk of Kullback-Leibler "distance" is misleading. While one can create a symmetric divergence measure by setting $D^*_{KL}(P,Q) = \frac{1}{2}D_{KL}(P,Q) + \frac{1}{2}D_{KL}(Q,P)$, this still fails the triangle inequality.

There is a close relationship between KL-divergence and a number of other statistical concepts. Consider, for example, *mutual information*. Given a joint distribution

$P(x,y)$ on $X \times Y$ with marginals P_X and P_Y, the mutual information of X and Y with respect to P is defined as $I_P(X,Y) = -\int_{X\times Y} \log_2(P(x,y)/[P_X(x)\cdot P_Y(y)])dP$. If we let $P_\perp(x,y) = P_X(x)\cdot P_Y(y)$ be the factorization of P, then $I_P(X,Y) = D(P,P_\perp)$. Thus, according to *KL*-divergence, mutual information measures the dissimilarity of a joint distribution from its factorization.

There is also a connection between *KL*-divergence and maximum likelihood estimation. Let $l_x(\theta) = p(x|\theta)$ be a likelihood function with parameter $\theta \in \Theta$, and imagine that enough data has been collected to make a certain empirical distribution $f(x)$ seem reasonable. In MLE one often hopes to find an estimate for θ that maximizes *expected log-likelihood* relative to one's data, i.e., we seek $\theta^\wedge = \text{argmax}_\theta E_f[\log_2(p(x|\theta)]$. To find this quantity it suffices to minimize the *KL*-divergence between $f(x)$ and $p(x|\theta^\wedge)$ since

$$\begin{aligned}
&\text{argmin}_\theta\, D_{KL}(f, p(\cdot|\theta^\wedge)) \\
&\quad = \text{argmin}_\theta - \int_X f(x)\cdot\log_2(p(x|\theta^\wedge)/f(x))dx \\
&\quad = \text{argmin}_\theta[H(f,f) - H(f,p(\cdot|\theta^\wedge))] \\
&\quad = \text{argmax}_\theta H(f,p(\cdot|\theta^\wedge)) \\
&\quad = \text{argmax}_\theta E_f[\log_2(p(x|\theta))].
\end{aligned}$$

In short, MLE minimizes Kullback-Leibler divergence from the empirical distribution.

Kullback-Leibler also plays a role in ▶model selection. Indeed, Akaike (1973) uses D_{KL} as the basis for his "information criterion" (AIC). Here, we imagine an unknown true distribution $P(x)$ over a sample space X, and a set Π_θ of models each element of which specifies a parameterized set of distributions $\pi(x|\theta)$ over X. The models in Π_θ are meant to approximate P, and the aim is to find the best approximation in light of data drawn from P. For each π and θ, $D_{KL}(P,\pi(x|\theta))$ measures the information lost when $\pi(x|\theta)$ is used to approximate P. If θ were known, one could minimize information loss by choosing π to minimize $D_{KL}(P,\pi(x|\theta))$. But, since θ is unknown one must estimate. For each body of data y and each π, let $\theta^\wedge_y$ be the MLE estimate for θ given y, and consider $D_{KL}(P,\pi(x|\theta^\wedge_y))$ as a random variable of y. Akaike maintained that one should choose the model that minimizes the expected value of this quantity, so that one chooses π to minimize $E_y[D_{KL}(P,\pi(x|\theta^\wedge_y))] = E_y[H(P,P) - H(P,\pi(\cdot|\theta^\wedge_y))]$. This is equivalent to maximizing $E_yE_x[\log_2(\pi(x|\theta^\wedge_y))]$. Akaike proved that $2k - \log_2(l_x(\theta^\wedge))$ is an unbiased estimate of this quantity for large samples, where $\theta^\wedge$ is the MLE estimate of θ and k is the number of estimated parameters. In this way, some have claimed, the policy of minimizing *KL*-divergence leads one to value simplicity in models since the "$2k$" term functions as a kind of penalty for complexity. (see Sober 2002).

KL-divergence also figures prominently in Bayesian approaches experimental design, where it is treated as a utility function. The objective in such work is to design experiments that maximize *KL*-divergence between the prior and posterior. The results of such experiments are interpreted as having a high degree of informational content. Lindley (1956) and De Groot (1962) are essential references here.

Bayesians have also appealed to *KL*-divergence to provide a rationale for Bayesian conditioning and related belief update rules, e.g., the probability kinematics of Jeffrey (1965). For example, Diaconis and Zabell (1982) show that the posterior probabilities prescribed by Bayesian conditioning or by probability kinematics minimize *KL*-divergence from the perspective of the prior. Thus, in the sense of information divergence captured by D_{KL}, these forms of updating introduce the least amount of new information consistent with the data received.

About the Author

For biography *see* the entry the ▶St. Petersburg paradox.

Cross References

▶Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
▶Chernoff Bound
▶Entropy
▶Entropy and Cross Entropy as Diversity and Distance Measures
▶Estimation
▶Information Theory and Statistics
▶Measurement of Uncertainty
▶Radon–Nikodým Theorem
▶Statistical Evidence

References and Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Proceedings of the international symposium on information theory. Budapest, Akademiai Kiado

De Groot M (1962) Uncertainty, information, and sequential experiments. Ann Math Stat 33:404–419

Diaconis P, Zabell S (1982) Updating subjective probability. J Am Stat Assoc 77:822–830

Jeffrey R (1965) The logic of decision. McGraw-Hill, New York

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86

Lindley DV (1956) On the measure of information provided by an experiment. Ann Stat 27:985–1005
Shannon CE (1948) A mathematical theory of communication. AT&T Tech J 27(379–423):623–656
Sober E (2002) Instrumentalism, parsimony, and the Akaike framework. Philos Sci 69:S112–S123

Kurtosis: An Overview

Edith Seier
Professor
East Tennessee State University, Johnson City, TN, USA

Pearson (1905) defined $\beta_2 = m_4/m_2^2$ (where m_i is the ith moment with respect to the mean) to compare other distributions to the normal distribution, for which $\beta_2 = 3$. He called $\eta = \beta_2 - 3$ the "degree of kurtosis" and mentioned that it "measures whether the frequency towards the mean is emphasized more or less than that required by the Gaussian law." In Greek, *kurtos* means convex, and *kurtosis* had been previously used to denote curvature both in mathematics and medicine. Pearson's development of the idea of kurtosis during the years previous to 1905 is examined by Fiori and Zenga (2009). "Coefficient of kurtosis" is the name usually given to β_2.

A sample estimator of β_2 is $b_2 = (\sum(x-\bar{x})^4/n)/s^4$. Statistical software frequently include an adjusted version of the estimator of η:

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\frac{\left[\sum(x-\bar{x})^4\right]}{s^4} - \frac{3(n-1)(n-1)}{(n-2)(n-3)}.$$

The adjustment reduces the bias, at least in the case of nearly normal distributions. Byers (2000) proved that $b_2 \leq n - 2 + 1/(n-1)$. Simulation results indicate that when β_2 is large for the population of origin, b_2 will be small on average if the sample size is small.

Currently the word kurtosis is understood in a broader sense, not limited to β_2. Balanda and MacGillivray (1988) conclude that kurtosis is best defined as "the location- and scale-free movement of probability mass from the shoulders of a distribution into its center and tails." which can be formalized in many ways. Kurtosis is associated to both, the center and the tails of a distribution. Kurtosis is invariant under linear transformations or change of units of the variable. High kurtosis is linked to high concentration of mass in the center and/or the tails of the distribution. Heavy tails is a topic of interest in the analysis of financial data.

Several kurtosis measures have been defined. L-kurtosis (Hosking 1992) is popular in the field of hydrology. There are other measures defined in terms of distances between quantiles, ratios of spread measures, comparisons of sum of distances to the median, and expected values of functions of the standardized variable other than the fourth power that corresponds to β_2.

Ruppert (1987) proposed the use of the influence function to analyze kurtosis measures and points out that even those defined with the intention of measuring peakedness or tail weight alone, end up measuring both. There are measures that are more sensitive to the tails of the distribution than others: β_2 gives high importance to the tails because it is defined in terms of the fourth power of the deviations from the mean. For example, the value of β_2 is 1.8 for the uniform distribution and 3.53, 4.51, 36.2 and 82.1 for the $SU(0,\delta)$ distribution with $\delta = 3, 2, 1, 0.9$ respectively. For the same distributions, the values of L-kurtosis are 0,0.143,0.168, 0.293 and 0.329. The upper bound for L-kurtosis is 1, while β_2 is unbounded. The estimator b_2 is sensitive to ▶outliers; one single outlier can dramatically change its value.

Another approach to the study of kurtosis is the comparison of cumulative distribution functions. Van Zwet (1964) defined the convexity criterion ($\prec_S$): two symmetric distributions with cumulative distribution functions F and G are ordered and $F \prec_S G$ if $G^{-1}(F(x))$ is convex to the right of the common point of symmetry. If $F \prec_S G$, the value of β_2 for F is not larger than its value for G. The following distributions are ordered according to the convexity criterion:

U-shaped $\prec_S$ Uniform $\prec_S$ Normal $\prec_S$ Logistic $\prec_S$ Laplace.

Some families of distributions are ordered according to the convexity criterion, with the order associated (either directly or inversely) to the value of their parameter. Among those families are *beta*(α,α), *Tukey*(λ), Johnson's $SU(0,\delta)$, and the symmetric two-sided power family $TSP(\alpha)$. Balanda and MacGillivray (1990) defined the spread-spread functions to compare non-necessarily symmetric distributions. Additional ordering criteria have been defined. Any new measure of kurtosis that is defined needs to order distributions in agreement with some ordering based on distribution functions. The numerical value of a kurtosis measure can be obtained for most distributions but not all distributions are ordered according to a CDF based ordering criterion. For example, the *Laplace* and *t-Student*$_{(6)}$ distributions have known values for β_2 (6 and 9 respectively). However, they are not $\prec_S$ ordered because $G^{-1}(F(x))$ is neither convex, nor concave for $x > 0$. In particular, not all the distributions are ordered with respect

to the normal distribution according to the convexity criterion; but uniform $\prec_S$ unimodal distributions.

There are several ways of measuring kurtosis, there is also more than one way of thinking about peak and tails. One simple way of visualizing peak and tails in a unimodal probability distribution is to superimpose, on $f(x)$, a uniform density function with the same median and variance (Kotz and Seier 2008).

High kurtosis affects the behavior of inferential tools. Van Zwet (1964) proved that, when working with symmetric distributions, the median is more efficient than the mean as estimator of the center when the distribution has very high kurtosis. The variance of the sample variance is related to β_2. Simulations indicate that the power of some tests for the equality of variances diminishes (for small samples) when the distribution of the variable has high kurtosis.

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Jarque-Bera Test
►Normality Tests
►Normality Tests: Power Comparison
►Statistical Distributions: An Overview

References and Further Reading

Balanda KP, MacGillivray HL (1988) Kurtosis: a critical review. Am Stat 42:111–119
Balanda KP, MacGillivray HL (1990) Kurtosis and spread. Can J Stat 18:17–30
Byers RH (2000) On the maximum of the standardized fourth moment. InterStat January #2 http://interstat.statjournals.net/YEAR/2000/articles/0001002.pdf
Fiori A, Zenga M (2009) Karl Pearson and the origin of Kurtosis. Int Stat Rev 77:40–50
Hosking JRM (1992) Moments or *L* moments? An example comparing two measures of distributional shape. Am Stat 46:186–199
Kotz S, Seier E (2008) Visualizing peak and tails to introduce kurtosis. Am Stat 62:348–352
Pearson K (1905) Skew variation, a rejoinder. Biometrika 4:169–212
Ruppert D (1987) What is Kurtosis? An influence function approach. Am Stat 41:1–5
Van Zwet WR (1964) Convex transformations of random variables. Mathematics Centre, Tract 7. Mathematisch Centrum Amsterdam, Netherlands

L

Large Deviations and Applications

Francis Comets
Professor
Université Paris Diderot, Paris, France

Large deviations is concerned with the study of rare events and of small probabilities. Let $X_i, 1 \le i \le n$, be independent identically distributed (i.i.d.) real random variables with expectation m, and $\bar{X}_n = (X_1 + \ldots + X_n)/n$ their empirical mean. The law of large numbers shows that, for any Borel set $A \subset \mathbb{R}$ not containing m in its closure, $P(\bar{X}_n \in A) \to 0$ as $n \to \infty$, but does not tell us how fast the probability vanishes. Large deviations theory gives us the rate of decay, which is exponential in n. Cramér's theorem states that,

$$P(\bar{X}_n \in A) = \exp\big(-n\big(\inf\{I(x); x \in A\} + o(1)\big)\big)$$

as $n \to \infty$, for all interval A. The rate function I can be computed as the Legendre conjugate of the logarithmic moment generating function of X,

$$I(x) = \sup\{\lambda x - \ln E \exp(\lambda X_1); \lambda \in \mathbb{R}\},$$

and is called the Cramér transform of the common law of the X_i's. The natural assumption is the finiteness of the ►moment generating function in a neighborhood of the origin, i.e., the property of exponential tails. The function $I : \mathbb{R} \to [0, +\infty]$ is convex with $I(m) = 0$.

- In the Gaussian case $X_i \sim \mathcal{N}(m, \sigma^2)$, we find $I(x) = (x-m)^2/(2\sigma^2)$.
- In the Bernoulli case $P(X_i = 1) = p = 1 - P(X_i = 0)$, we find the entropy function $I(x) = x \ln(x/p) + (1-x) \ln(1-x)/(1-p)$ for $x \in [0,1]$, and $I(x) = +\infty$ otherwise.

To emphasize the importance of rare events, let us mention a consequence, the Erdös–Rényi law: consider an infinite sequence $X_i, i \ge 1$, of Bernoulli i.i.d. variables with parameter p, and let R_n denote the length of the longest consecutive run, contained within the first n tosses, in which the fraction of 1s is at least a ($a > p$). Erdös and Rényi proved that, almost surely as $n \to \infty$,

$$R_n / \ln n \longrightarrow I(a)^{-1},$$

with the function I from the Bernoulli case above. Though it may look paradoxical, large deviations are at the core of this event of full probability. This result is the basis of ►bioinformatics applications like sequence matching, and of statistical tests for sequence randomness.

The theory does not only apply to independent variables, but allows for many variations, including weakly dependent variables in a general state space, Markov or ►Gaussian processes, large deviations from ►ergodic theorems, non-asymptotic bounds, asymptotic expansions (Edgeworth expansions), etc.

Here is the formal definition. Given a Polish space (i.e., a separable complete metric space) $\mathcal{X}$, let $\{\mathbb{P}_n\}$ be a sequence of Borel probability measures on $\mathcal{X}$, let a_n be a positive sequence tending to infinity, and finally let $I : \mathcal{X} \to [0, +\infty]$ be a lower semicontinuous functional on X. We say that the sequence $\{\mathbb{P}_n\}$ satisfies a large deviation principle with speed a_n and rate I, if for each measurable set $E \subset X$

$$-\inf_{x \in E^\circ} I(x) \le \underline{\lim_n}\, a_n^{-1} \ln \mathbb{P}_n(E) \le \overline{\lim_n}\, a_n^{-1} \ln \mathbb{P}_n(E) \le -\inf_{x \in \bar{E}} I(x)$$

where $\bar{E}$ and E° denote respectively the closure and interior of E. The rate function can be obtained as

$$I(x) = -\lim_{\delta \searrow 0} \lim_{n \to \infty} a_n^{-1} \ln \mathbb{P}_n(B(x, \delta)),$$

with $B(x, \delta)$ the ball of center x and radius δ.

Sanov's theorem and sampling with replacement: let μ be a probability measure on a set Σ that we assume finite for simplicity, with $\mu(y) > 0$ for all $y \in \Sigma$. Let $Y_i, i \ge 1$, an i.i.d. sequence with law μ, and N_n the score vector of the n-sample,

$$N_n(y) = \sum_{i=1}^{n} \mathbf{1}_y(Y_i).$$

By the law of large numbers, $N_n/n \to \mu$ almost surely. From the ►multinomial distribution, one can check that, for all ν such that $n\nu$ is a possible score vector for the n-sample,

$$(n+1)^{-|\Sigma|} e^{-nH(\nu|\mu)} \le P(n^{-1}N_n = \nu) \le e^{-nH(\nu|\mu)},$$

where $H(\nu|\mu) = \sum_{y \in \Sigma} \nu(y) \ln \frac{\nu(y)}{\mu(y)}$ is the relative entropy of ν with respect to μ. The large deviations theorem holds

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

for the empirical distribution of a general n-sample, with speed n and rate $I(\nu) = H(\nu|\mu)$ given by the natural generalization of the above formula. This result, due to Sanov, has many consequences in information theory and statistical mechanics (Dembo and Zeitouni 1998; den Hollander 2000), and for exponential families in statistics. Applications in statistics also include point estimation (by giving the exponential rate of convergence of M-estimators) and for hypothesis testing (Bahadur efficiency) (Kester 1985), and concentration inequalities (Dembo and Zeitouni 1998).

The *Freidlin–Wentzell theory* deals with diffusion processes with small noise,

$$dX_t^\epsilon = b\left(X_t^\epsilon\right) dt + \sqrt{\epsilon}\, \sigma\left(X_t^\epsilon\right) dB_t\,, \qquad X_0^\epsilon = y.$$

The coefficients b, σ are uniformly lipshitz functions, and B is a standard Brownian motion (see ►Brownian Motion and Diffusions). The sequence X^ϵ can be viewed as $\epsilon \searrow 0$ as a small random perturbation of the ordinary differential equation

$$dx_t = b(x_t)dt\,, \qquad x_0 = y.$$

Indeed, $X^\epsilon \to x$ in the supremum norm on bounded time-intervals. Freidlin and Wentzell have shown that, on a finite time interval $[0, T]$, the sequence X^ϵ with values in the path space obeys the LDP with speed ϵ^{-1} and rate function

$$I(\phi) = \frac{1}{2}\int_0^T \sigma(\phi(t))^{-2}\left(\dot{\phi}(t) - b(\phi(t))\right)^2 dt$$

if ϕ is absolutely continuous with square-integrable derivative and $\phi(0) = y$; $I(\phi) = \infty$ otherwise. (To fit in the above formal definition, take a sequence $\epsilon = \epsilon_n \searrow 0$, and for $\mathbb{P}_n$ the law of X^{ϵ_n}.)

The Freidlin–Wentzell theory has applications in physics (metastability phenomena) and engineering (tracking loops, statistical analysis of signals, stabilization of systems, and algorithms) (Freidlin and Wentzell 1998; Dembo and Zeitouni 1998; Olivieri and Vares 2005).

About the Author

Francis Comets is Professor of Applied Mathematics at University of Paris - Diderot (Paris 7), France. He is the Head of the team "Stochastic Models" in the CNRS laboratory "Probabilité et modèles aléatoires" since 1999. He is the Deputy Director of the Foundation Sciences Mathématiques de Paris since its creation in 2006. He has coauthored 50 research papers, with a focus on random medium, and one book ("Calcul stochastique et modèles de diffusions" in French, Dunod 2006, with Thierry Meyre). He has supervised more than 10 Ph.D thesis, and served as Associate Editor for *Stochastic Processes and their Applications* (1997–2002).

Cross References

►Asymptotic Relative Efficiency in Testing
►Chernoff Bound
►Entropy and Cross Entropy as Diversity and Distance Measures
►Laws of Large Numbers
►Limit Theorems of Probability Theory
►Moderate Deviations
►Robust Statistics

References and Further Reading

Dembo A, Zeitouni O (1998) Large deviations techniques and applications. Springer, New York
den Hollander F (2000) Large deviations. Am Math Soc, Providence, RI
Freidlin MI, Wentzell AD (1998) Random perturbations of dynamical systems. Springer, New York
Kester A (1985) Some large deviation results in statistics. CWI Tract, 18. Centrum voor Wiskunde en Informatica, Amsterdam
Olivieri E, Vares ME (2005) Large deviations and metastability. Cambridge University Press, Cambridge

Laws of Large Numbers

Andrew Rosalsky
Professor
University of Florida, Gainesville, FL, USA

The *laws of large numbers* (LLNs) provide bounds on the fluctuation behavior of sums of random variables and, as we will discuss herein, lie at the very foundation of statistical science. They have a history going back over 300 years. The literature on the LLNs is of epic proportions, as this concept is indispensable in probability and statistical theory and their application.

Probability theory, like some other areas of mathematics such as geometry for example, is a subject arising from an attempt to provide a rigorous mathematical model for real world phenomena. In the case of probability theory, the real world phenomena are chance behavior of biological processes or physical systems such as gambling games and their associated monetary gains or losses.

The probability of an event is the abstract counterpart to the notion of the long-run relative frequency of the

occurence of the event through infinitely many replications of the experiment. For example, if a quality control engineer asserts that the probability is 0.98 that a widget produced by her production team meets specifications, then she is asserting that in the long-run, 98% of those widgets meet specifications. The phrase "in the long-run" requires the notion of *limit* as the sample size approaches infinity. The long-run relative frequency approach for describing the probability of an event is natural and intuitive but, nevertheless, it raises serious mathematical questions. Does the limiting relative frequency always exist as the sample size approaches infinity and is the limit the same irrespective of the sequence of experimental outcomes? It is easy to see that the answers are negative. Indeed, in the above example, depending on the sequence of experimental outcomes, the proportion of widgets meeting specifications could fluctuate repeatedly from near 0 to near 1 as the number of widgets sampled approaches infinity. So in what sense can it be asserted that the limit exists and is 0.98? To provide an answer to this question, one needs to apply a LLN.

The LLNs are of two types, viz., *weak* LLNs (WLLNs) and *strong* LLNs (SLLNs). Each type involves a different mode of convergence. In general, a WLLN (resp., a SLLN) involves convergence in probability (resp., convergence almost surely (a.s.)). The definitions of these two modes of convergence will now be reviewed.

Let $\{U_n, n \geq 1\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and let $c \in \mathbb{R}$. We say that U_n *converges in probability* to c (denoted $U_n \xrightarrow{P} c$) if

$$\lim_{n\to\infty} P(|U_n - c| > \varepsilon) = 0 \text{ for all } \varepsilon > 0.$$

We say that U_n *converges a.s.* to c (denoted $U_n \to c$ a.s.) if

$$P(\{\omega \in \Omega : \lim_{n\to\infty} U_n(\omega) = c\}) = 1.$$

If $U_n \to c$ a.s., then $U_n \xrightarrow{P} c$; the converse is not true in general.

The celebrated Kolmogorov SLLN (see, e.g., Chow and Teicher [1997], p. 125) is the following result. Let $\{X_n, n \geq 1\}$ be a sequence of independent and identically distributed (i.i.d.) random variables and let $c \in \mathbb{R}$. Then

$$\frac{\sum_{i=1}^n X_i}{n} \to c \text{ a.s. if and only if } EX_1 = c. \tag{1}$$

Using statistical terminology, the sufficiency half of (1) asserts that the *sample mean* converges a.s. to the *population mean* as the *sample size n* approaches infinity provided the population mean exists and is finite. This result is of fundamental importance in statistical science. It follows from (1) that

$$\text{if } EX_1 = c \in \mathbb{R}, \text{ then } \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{P} c; \tag{2}$$

this result is the Khintchine WLLN (see, e.g., Petrov [1995], p. 134).

Next, suppose $\{A_n, n \geq 1\}$ is a sequence of independent events all with the same probability p. A special case of the Kolmogorov SLLN is the limit result

$$\hat{p}_n \to p \text{ a.s.} \tag{3}$$

where $\hat{p}_n = \sum_{i=1}^n I_{A_i}/n$ is the proportion of $\{A_1, \ldots, A_n\}$ to occur, $n \geq 1$. (Here I_{A_i} is the indicator function of $A_i, i \geq 1$.) This result is the first SLLN ever proved and was discovered by Emile Borel in 1909. Hence, with probability 1, the *sample proportion* $\hat{p}_n$ approaches the *population proportion* p as the sample size $n \to \infty$. It is this SLLN which thus provides the theoretical justification for the long-run relative frequency approach to interpreting probabilities. Note, however, that the convergence in (3) is not pointwise on Ω but, rather, is pointwise on some subset of Ω *having probability 1.* Consequently, any interpretation of $p = P(A_1)$ via (3) necessitates that one has *a priori* an intuitive understanding of the notion of an event having probability 1.

The SLLN (3) is a key component in the proof of the Glivenko–Cantelli theorem (see ▶Glivenko-Cantelli Theorems) which, roughly speaking, asserts that with probability 1, a population distribution function can be uniformly approximated by a sample (or empirical) distribution function as the sample size approaches infinity. This result is referred to by Rényi (1970, p. 400) as the *fundamental theorem of mathematical statistics* and by Loève (1977, p. 20) as the *central statistical theorem*.

In 1689, Jacob Bernoulli (1654–1705) proved the first WLLN

$$\hat{p}_n \xrightarrow{P} p. \tag{4}$$

Bernoulli's renowned book *Ars Conjectandi* (*The Art of Conjecturing*) was published posthumously in 1713, and it is here where the proof of his WLLN was first published. It is interesting to note that there is over a 200 year gap between the WLLN (4) of Bernoulli and the corresponding SLLN (3) of Borel.

An interesting example is the following modification of one of Stout (1974, p. 9). Suppose that the quality control engineer referred to above would like to estimate the proportion p of widgets produced by her production team

that meet specifications. She estimates p by using the proportion $\hat{p}_n$ of the first n widgets produced that meet specifications and she is interested in knowing if there will ever be a point in the sequence of examined widgets such that with probability (at least) a specified large value, $\hat{p}_n$ will be within ε of p *and stay* within ε of p as the sampling continues (where $\varepsilon > 0$ is a prescribed tolerance). The answer is affirmative since (3) is equivalent to the assertion that for a given $\varepsilon > 0$ and $\delta > 0$, there exists a positive integer $N_{\varepsilon,\delta}$ such that

$$P\left(\cap_{n=N_{\varepsilon,\delta}}^{\infty}[|\hat{p}_n - p| \le \varepsilon]\right) \ge 1 - \delta.$$

That is, the probability is arbitrarily close to 1 that $\hat{p}_n$ will be arbitrarily close to p simultaneously for all n beyond some point. If one applied instead the WLLN (4), then it could only be asserted that for a given $\varepsilon > 0$ and $\delta > 0$, there exists a positive integer $N_{\varepsilon,\delta}$ such that

$$P(|\hat{p}_n - p| \le \varepsilon) \ge 1 - \delta \text{ for all } n \ge N_{\varepsilon,\delta}.$$

There are numerous other versions of the LLNs and we will discuss only a few of them. Note that the expressions in (1) and (2) can be rewritten, respectively, as

$$\frac{\sum_{i=1}^n X_i - nc}{n} \to 0 \text{ a.s. and } \frac{\sum_{i=1}^n X_i - nc}{n} \xrightarrow{P} 0$$

thereby suggesting the following definitions. A sequence of random variables $\{X_n, n \ge 1\}$ is said to obey a general *SLLN* (resp., *WLLN*) with *centering sequence* $\{a_n, n \ge 1\}$ and *norming sequence* $\{b_n, n \ge 1\}$ (where $0 < b_n \to \infty$) if

$$\frac{\sum_{i=1}^n X_i - a_n}{b_n} \to 0 \text{ a.s.} \left(\text{resp., } \frac{\sum_{i=1}^n X_i - a_n}{b_n} \xrightarrow{P} 0\right).$$

A famous result of Marcinkiewicz and Zygmund (see, e.g., Chow and Teicher (1997), p. 125) extended the Kolmogorov SLLN as follows. Let $\{X_n, n \ge 1\}$ be a sequence of i.i.d. random variables and let $0 < p < 2$. Then

$$\frac{\sum_{i=1}^n X_i - nc}{n^{1/p}} \to 0 \text{ a.s. for some } c \in \mathbb{R} \text{ if and only if } E|X_1|^p < \infty.$$

In such a case, necessarily $c = EX_1$ if $p \ge 1$ whereas c is arbitrary if $p < 1$.

Feller (1946) extended the Marcinkiewicz–Zygmund SLLN to the case of a more general norming sequence $\{b_n, n \ge 1\}$ satisfying suitable growth conditions.

The following WLLN is ascribed to Feller by Chow and Teicher (1997, p. 128). If $\{X_n, n \ge 1\}$ is a sequence of i.i.d. random variables, then there exist real numbers $a_n, n \ge 1$ such that

$$\frac{\sum_{i=1}^n X_i - a_n}{n} \xrightarrow{P} 0 \tag{5}$$

if and only if

$$nP(|X_1| > n) \to 0 \text{ as } n \to \infty. \tag{6}$$

In such a case, $a_n - nE(X_1 I_{[|X_1| \le n]}) \to 0$ as $n \to \infty$.

The condition (6) is weaker than $E|X_1| < \infty$. If $\{X_n, n \ge 1\}$ is a sequence of i.i.d. random variables where X_1 has probability density function

$$f(x) = \begin{cases} \frac{c}{x^2 \log|x|}, & |x| \ge e \\ 0, & |x| < e \end{cases}$$

where c is a constant, then $E|X_1| = \infty$ and the SLLN $\sum_{i=1}^n X_i/n \to c$ a.s. fails for every $c \in \mathbb{R}$ but (6) and hence the WLLN (5) hold with $a_n = 0, n \ge 1$.

Klass and Teicher (1977) extended the Feller WLLN to the case of a more general norming sequence $\{b_n, n \ge 1\}$ thereby obtaining a WLLN analog of Feller's (1946) extension of the Marcinkiewicz–Zygmund SLLN.

Good references for studying the LLNs are the books by Révész (1968), Stout (1974), Loève (1977), Chow and Teicher (1997), and Petrov (1995). While the LLNs have been studied extensively in the case of independent summands, some of the LLNs presented in these books involve summands obeying a dependence structure other than that of independence.

A large literature of investigation on the LLNs for sequences of Banach space valued random elements has emerged beginning with the pioneering work of Mourier (1953). See the monograph by Taylor (1978) for background material and results up to 1978. Excellent references are the books by Vakhania, Tarieladze, and Chobanyan (1987) and Ledoux and Talagrand (1991). More recent results are provided by Adler et al. (1991), Cantrell and Rosalsky (2004), and the references in these two articles.

About the Author

Professor Rosalsky is Associate Editor, *Journal of Applied Mathematics and Stochastic Analysis* (1989–present), and Associate Editor, *International Journal of Mathematics and Mathematical Sciences* (1994–present). He has collaborated with research workers from 15 different countries. At the University of Florida, he twice received a Teaching Improvement Program Award and on five occasions was named an Anderson Scholar Faculty Honoree, an honor reserved for faculty members designated by undergraduate honor students as being the most effective and inspiring.

Cross References

►Almost Sure Convergence of Random Variables
►Borel–Cantelli Lemma and Its Generalizations

►Chebyshev's Inequality
►Convergence of Random Variables
►Ergodic Theorem
►Estimation: An Overview
►Expected Value
►Foundations of Probability
►Glivenko-Cantelli Theorems
►Probability Theory: An Outline
►Random Field
►Statistics, History of
►Strong Approximations in Probability and Statistics

References and Further Reading

Adler A, Rosalsky A, Taylor RL (1991) A weak law for normed weighted sums of random elements in Rademacher type *p* Banach spaces. J Multivariate Anal 37:259–268

Cantrell A, Rosalsky A (2004) A strong law for compactly uniformly integrable sequences of independent random elements in Banach spaces. Bull Inst Math Acad Sinica 32:15–33

Chow YS, Teicher H (1997) Probability theory: independence, interchangeability, martingales, 3rd edn. Springer, New York

Feller W (1946) A limit theorem for random variables with infinite moments. Am J Math 68:257–262

Klass M, Teicher H (1977) Iterated logarithm laws for asymmetric random variables barely with or without finite mean. Ann Probab 5:861–874

Ledoux M, Talagrand M (1991) Probability in Banach spaces: isoperimetry and processes. Springer, Berlin

Loève M (1977) Probability theory, vol I, 4th edn. Springer, New York

Mourier E (1953) Eléments aléatoires dans un espace de Banach. Annales de l' Institut Henri Poincaré 13:161–244

Petrov VV (1995) Limit theorems of probability theory: sequences of independent random variables. Clarendon, Oxford

Rényi A (1970) Probability theory. North-Holland, Amsterdam

Révész P (1968) The laws of large numbers. Academic, New York

Stout WF (1974) Almost sure convergence. Academic, New York

Taylor RL (1978) Stochastic convergence of weighted sums of random elements in linear spaces. Lecture notes in mathematics, vol 672. Springer, Berlin

Vakhania NN, Tarieladze VI, Chobanyan SA (1987) Probability distributions on Banach spaces. D. Reidel, Dordrecht, Holland

Learning Statistics in a Foreign Language

Khidir M. Abdelbasit
Sultan Qaboos University, Muscat, Sultanate of Oman

Background

The Sultanate of Oman is an Arabic-speaking country, where the medium of instruction in pre-university education is Arabic. In Sultan Qaboos University (SQU) all sciences (including Statistics) are taught in English. The reason is that most of the scientific literature is in English and teaching in the native language may leave graduates at a disadvantage. Since only few instructors speak Arabic, the university adopts a policy of no communication in Arabic in classes and office hours. Students are required to achieve a minimum level in English (about 4.0 IELTS score) before they start their study program. Very few students achieve that level on entry and the majority spends about two semesters doing English only.

Language and Cultural Problems

It is to be expected that students from a non-English-speaking background will face serious difficulties when learning in English especially in the first year or two. Most of the literature discusses problems faced by foreign students pursuing study programs in an English-speaking country, or a minority in a multi-cultural society (see for example Coutis P. and Wood L., Hubbard R, Koh E). Such students live (at least while studying) in an English-speaking community with which they have to interact on a daily basis. These difficulties are more serious for our students who are studying in their own country where English is not the official language. They hardly use English outside classrooms and avoid talking in class as much as they can.

My SQU Experience

Statistical concepts and methods are most effectively taught through real-life examples that the students appreciate and understand. We use the most popular textbooks in the USA for our courses. These textbooks use this approach with US students in mind. Our main problems are:

- Most of the examples and exercises used are completely alien to our students. The discussions meant to maintain the students' interest only serve to put ours off. With limited English they have serious difficulties understanding what is explained and hence tend not to listen to what the instructor is saying. They do not read the textbooks because they contain pages and pages of lengthy explanations and discussions they cannot follow. A direct effect is that students may find the subject boring and quickly lose interest. Their attention then turns to the art of passing tests instead of acquiring the intended knowledge and skills. To pass their tests they use both their class and study times looking through examples, concentrating on what formula to use and where to plug the numbers they have to get the answer. This way they manage to do the mechanics fairly well, but the concepts are almost entirely missed.

- The problem is worse with introductory probability courses where games of chance are extensively used as illustrative examples in textbooks. Most of our students have never seen a deck of playing cards and some may even be offended by discussing card games in a classroom.

The burden of finding strategies to overcome these difficulties falls on the instructor. Statistical terms and concepts such as parameter/statistic, sampling distribution, unbiasedness, consistency, sufficiency, and ideas underlying hypothesis testing are not easy to get across even in the students' own language. To do that in a foreign language is a real challenge. For the Statistics program to be successful, all (or at least most of the) instructors involved should be up to this challenge. This is a time-consuming task with little reward, other than self satisfaction. In SQU the problem is compounded further by the fact that most of the instructors are expatriates on short-term contracts who are more likely to use their time for personal career advancement, rather than time-consuming community service jobs.

What Can Be Done?

For our first Statistics course we produced a manual that contains very brief notes and many samples of previous quizzes, tests, and examinations. It contains a good collection of problems from local culture to motivate the students. The manual was well received by the students, to the extent that students prefer to practice with examples from the manual rather than the textbook.

Textbooks written in English that are brief and to the point are needed. These should include examples and exercises from the students' own culture. A student trying to understand a specific point gets distracted by lengthy explanations and discouraged by thick textbooks to begin with. In a classroom where students' faces clearly indicate that you have got nothing across, it is natural to try explaining more using more examples. In the end of semester evaluation of a course I taught, a student once wrote "The instructor explains things more than needed. He makes simple points difficult." This indicates that, when teaching in a foreign language, lengthy oral or written explanations are not helpful. A better strategy will be to explain concepts and techniques briefly and provide plenty of examples and exercises that will help the students absorb the material by osmosis. The basic statistical concepts can only be effectively communicated to students in their own language. For this reason textbooks should contain a good glossary where technical terms and concepts are explained using the local language.

I expect such textbooks to go a long way to enhance students' understanding of Statistics. An international project can be initiated to produce an introductory statistics textbook, with different versions intended for different geographical areas. The English material will be the same; the examples vary, to some extent, from area to area and glossaries in local languages. Universities in the developing world, naturally, look at western universities as models, and international (western) involvement in such a project is needed for it to succeed. The project will be a major contribution to the promotion of understanding Statistics and excellence in statistical education in developing countries. The international statistical institute takes pride in supporting statistical progress in the developing world. This project can lay the foundation for this progress and hence is worth serious consideration by the institute.

Cross References

►Online Statistics Education
►Promoting, Fostering and Development of Statistics in Developing Countries
►Selection of Appropriate Statistical Methods in Developing Countries
►Statistical Literacy, Reasoning, and Thinking
►Statistics Education

References and Further Reading

Coutis P, Wood L (2002) Teaching statistics and academic language in culturally diverse classrooms. http://www.math.uoc.gr/~ictm2/Proceedings/pap172.pdf

Hubbard R (1990) Teaching statistics to students who are learning in a foreign language. ICOTS 3

Koh E (1994) Teaching statistics to students with limited language skills using MINITAB http://archives.math.utk.edu/ICTCM/VOL07/C012/paper.pdf

Least Absolute Residuals Procedure

Richard William Farebrother
Honorary Reader in Econometrics
Victoria University of Manchester, Manchester, UK

For $i = 1, 2, \ldots, n$, let $\{x_{i1}, x_{i2}, \ldots, x_{iq}, y_i\}$ represent the ith observation on a set of $q + 1$ variables and suppose that we wish to fit a linear model of the form

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iq}\beta_q + \epsilon_i$$

to these n observations. Then, for $p > 0$, the L_p-norm fitting procedure chooses values for $b_1, b_2, \ldots, b_q$ to minimise the L_p-norm of the residuals $[\sum_{i=1}^{n} |e_i|^p]^{1/p}$ where, for $i = 1, 2, \ldots, n$, the ith residual is defined by

$$e_i = y_i - x_{i1}b_1 - x_{i2}b_2 - \ldots - x_{iq}b_q.$$

The most familiar L_p-norm fitting procedure, known as the ►least squares procedure, sets $p = 2$ and chooses values for $b_1, b_2, \ldots, b_q$ to minimize the sum of the squared residuals $\sum_{i=1}^{n} e_i^2$.

A second choice, to be discussed in the present article, sets $p = 1$ and chooses $b_1, b_2, \ldots, b_q$ to minimize the sum of the absolute residuals $\sum_{i=1}^{n} |e_i|$.

A third choice sets $p = \infty$ and chooses $b_1, b_2, \ldots, b_q$ to minimize the largest absolute residual $max_{i=1}^{n}|e_i|$.

Setting $u_i = e_i$ and $v_i = 0$ if $e_i \geq 0$ and $u_i = 0$ and $v_i = -e_i$ if $e_i < 0$, we find that $e_i = u_i - v_i$ so that the least absolute residuals (*LAR*) fitting problem chooses $b_1, b_2, \ldots, b_q$ to minimize the sum of the absolute residuals

$$\sum_{i=1}^{n}(u_i + v_i)$$

subject to

$$x_{i1}b_1 + x_{i2}b_2 + \cdots + x_{iq}b_q + U_i - v_i = y_i \quad \text{for } i = 1, 2, \ldots, n$$

$$\text{and} \quad U_i \geq 0, v_i \geq 0 \quad \text{for } i = 1, 2, \ldots, n.$$

The *LAR* fitting problem thus takes the form of a linear programming problem and is often solved by means of a variant of the dual simplex procedure.

Gauss has noted (when $q = 2$) that solutions of this problem are characterized by the presence of a set of q zero residuals. Such solutions are robust to the presence of outlying observations. Indeed, they remain constant under variations in the other $n - q$ observations provided that these variations do not cause any of the residuals to change their signs.

The *LAR* fitting procedure corresponds to the maximum likelihood estimator when the ϵ-disturbances follow a double exponential (Laplacian) distribution. This estimator is more robust to the presence of outlying observations than is the standard least squares estimator which maximizes the likelihood function when the ϵ-disturbances are normal (Gaussian). Nevertheless, the *LAR* estimator has an asymptotic normal distribution as it is a member of Huber's class of M-estimators.

There are many variants of the basic *LAR* procedure but the one of greatest historical interest is that proposed in 1760 by the Croatian Jesuit scientist Rugjer (or Rudjer) Josip Bošković (1711–1787) (Latin: Rogerius Josephus Boscovich; Italian: Ruggiero Giuseppe Boscovich). In his variant of the standard *LAR* procedure, there are two explanatory variables of which the first is constant $x_{i1} = 1$ and the values of b_1 and b_2 are constrained to satisfy the adding-up condition $\sum_{i=1}^{n}(y_i - b_1 - x_{i2}b_2) = 0$ usually associated with the least squares procedure developed by Gauss in 1795 and published by Legendre in 1805. Computer algorithms implementing this variant of the *LAR* procedure with $q \geq 2$ variables are still to be found in the literature.

For an account of recent developments in this area, see the series of volumes edited by Dodge (1987, 1992, 1997, 2002). For a detailed history of the *LAR* procedure, analyzing the contributions of Bošković, Laplace, Gauss, Edgeworth, Turner, Bowley and Rhodes, see Farebrother (1999). And, for a discussion of the geometrical and mechanical representation of the least squares and *LAR* fitting procedures, see Farebrother (2002).

About the Author

Richard William Farebrother was a member of the teaching staff of the Department of Econometrics and Social Statistics in the (Victoria) University of Manchester from 1970 until 1993 when he took early retirement (he has been blind since 1993). From 1993 until 2001 he was an Honorary Reader in Econometrics in the Department of Economic Studies of the same University. He has published three books: *Linear Least Squares Computations* in 1988, *Fitting Linear Relationships: A History of the Calculus of Observations 1750–1900* in 1999, and *Visualizing statistical Models and Concepts* in 2002. He has also published more than 140 research papers in a wide range of subject areas including econometric theory, computer algorithms, statistical distributions, statistical inference, and the history of statistics. Dr. Farebrother was also visiting Associate Professor (1982) at Monash University, Australia, and Visiting Professor (1990) at the Institute for Advanced Studies in Vienna, Austria.

Cross References

►Least Squares
►Residuals

References and Further Reading

Dodge Y (ed) (1987) Statistical data analysis based on the L_1-norm and related methods. North-Holland, Amsterdam
Dodge Y (ed) (1992) L_1-statistical analysis and related methods. North-Holland, Amsterdam
Dodge Y (ed) (1997) L_1-statistical procedures and related topics. Institute of Mathematical Statistics, Hayward
Dodge Y (ed) (2002) Statistical data analysis based on the L_1-norm and related methods. Birkhäuser, Basel

Farebrother RW (1999) Fitting linear relationships: a history of the calculus of observations 1750–1900. Springer, New York
Farebrother RW (2002) Visualizing statistical models and concepts. Marcel Dekker, New York

Least Squares

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

Least Squares (LS) problem involves some algebraic and numerical techniques used in "solving" overdetermined systems $F(x) \approx b$ of equations, where $b \in R^n$ while $F(x)$ is a column of the form

$$F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \dots \\ f_{m-1}(x) \\ f_m(x) \end{bmatrix}$$

with entries $f_i = f_i(x)$, $i = 1, \dots n$, where $x = (x_1, \dots, x_p)^T$. The LS problem is *linear* when each f_i is a linear function, and *nonlinear* – if not.

Linear LS problem refers to a system $Ax = b$ of linear equations. Such a system is overdetermined if $n > p$. If $b \notin range(A)$ the system has no proper solution and will be denoted by $Ax \approx b$. In this situation we are seeking for a solution of some optimization problem. The name "Least Squares" is justified by the l_2-norm commonly used as a measure of imprecision.

The LS problem has a clear statistical interpretation in regression terms. Consider the usual regression model

$$y_i = f_i(x_{i1}, \dots, x_{ip}; \beta_1, \dots, \beta_p) + e_i \text{ for } i = 1, \dots, n \qquad (1)$$

where x_{ij}, $i = 1, \dots n, j = 1, \dots p$, are some constants given by experimental design, f_i, $i = 1, \dots n$, are given functions depending on unknown parameters β_j, $j = 1, \dots, p$, while y_i, $i = 1, \dots, n$, are values of these functions, observed with some random errors e_i. We want to estimate the unknown parameters β_i on the basis of the data set $\{x_{ij}, y_i\}$.

In linear regression each f_i is a linear function of type $f_i = \sum_{j=1}^{p} c_{ij}(x_1, \dots, x_n)\beta_j$ and the model (1) may be presented in vector-matrix notation as

$$y = X\beta + e,$$

where $y = (y_1, \dots, y_n)^T$, $e = (e_1, \dots, e_n)^T$ and $\beta = (\beta_1, \dots, \beta_p)^T$, while X is a $n \times p$ matrix with entries x_{ij}. If $e_1, \dots, e_n$ are not correlated with mean zero and a common (perhaps unknown) variance then the problem of Best Linear Unbiased Estimation (BLUE) of β reduces to finding a vector $\widehat{\beta}$ that minimizes the norm $\left\| y - X\widehat{\beta} \right\|_2 = (y - X\widehat{\beta})^T (y - X\widehat{\beta})$

Such a vector is said to be the *ordinary* LS solution of the overparametrized system $X\beta \approx y$. On the other hand the last one reduces to solving the consistent system

$$X^T X\beta = X^T y$$

of linear equations called normal equations. In particular, if $rank(X) = p$ then the system has a unique solution of the form

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

For linear regression $y_i = \alpha + \beta x_i + e_i$ with one regressor x the BLU estimators of the parameters α and β may be presented in the convenient form as

$$\widehat{\beta} = \frac{ns_{xy}}{ns^2 x} \quad \text{and} \quad \widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x},$$

where

$$ns_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n},$$

$$ns_x^2 = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}, \quad \bar{x} = \frac{\sum_i x_i}{n} \quad \text{and} \quad \bar{y} = \frac{\sum_i y_i}{n}$$

For its computation we only need to use a simple pocket calculator.

Example The following table presents the number of residents in thousands (x) and the unemployment rate in % (y) for some cities of Poland. Estimate the parameters β and α.

x_i	131	87	56	312	185	252	157
y_i	8.7	10.2	9.9	6.3	6.1	5.2	11.0

In this case $\sum_i x_i = 1,180$, $\sum_i y_i = 57.4$, $\sum_i x_i^2 = 247,588$ and $\sum_i x_i y_i = 8,713$. Therefore $ns_x^2 = 48,673.71$ and $ns_{xy} = -963$. Thus $\widehat{\beta} = -0.02$ and $\widehat{\alpha} = 11.77$ and hence $f(x) = -0.02x + 11.77$.

If the variance–covariance matrix of the error vector e coincides (except a multiplicative scalar σ^2) with a positive definite matrix V then the Best Linear Unbiased estimation reduces to the minimization of $(y - X\widehat{\beta})^T V(y - X\widehat{\beta})$, called the *weighed* LS problem. Moreover, if $rank(X) = p$ then its solution is given in the form

$$\widehat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y.$$

It is worth to add that a nonlinear LS problem is more complicated and its explicit solution is usually not known. Instead of this some algorithms are suggested.

Total least squares problem. The problem has been posed in recent years in numerical analysis as an alternative for the LS problem in the case when all data are affected by errors.

Consider an overdetermined system of n linear equations $Ax \approx b$ with k unknown x. The TLS problem consists in minimizing the Frobenius norm

$$\left\| [A, b] - \left[\widehat{A}, \widehat{b}\right] \right\|_F$$

for all $\widehat{A} \in R^{n \times k}$ and $\widehat{b} \in range(\widehat{A})$, where the Frobenius norm is defined by $\|(a_{ij})\|_F = \sum_{i,j} a_{ij}$. Once a minimizing $\left[\widehat{A}, \widehat{b}\right]$ is found, then any x satisfying $\widehat{A}x = \widehat{b}$ is called a *TLS solution* of the initial system $Ax \approx b$.

The trouble is that the minimization problem may not be solvable, or its solution may not be unique. As an example one can set

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } b = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

It is known that the TLS solution (if exists) is always better than the ordinary LS in the sense that the correction $b - A\widehat{x}$ has smaller l_2-norm. The main tool in solving the TLS problems is the following Singular Value Decomposition:

For any matrix A of $n \times k$ with real entries there exist orthonormal matrices $P = \left[p_1, \ldots, p_n\right]$ and $Q = \left[q_1, \ldots, q_k\right]$ such that

$$P^T AQ = diag(\sigma_1, \ldots, \sigma_m), \text{ where } \sigma_1 \geq \cdots \geq \sigma_m \text{ and } m = \min\{n, k\}.$$

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Absolute Penalty Estimation
►Adaptive Linear Regression
►Analysis of Areal and Spatial Interaction Data
►Autocorrelation in Regression
►Best Linear Unbiased Estimation in Linear Models
►Estimation
►Gauss-Markov Theorem
►General Linear Models
►Least Absolute Residuals Procedure
►Linear Regression Models
►Nonlinear Models
►Optimum Experimental Design
►Partial Least Squares Regression Versus Other Methods
►Simple Linear Regression
►Statistics, History of
►Two-Stage Least Squares

References and Further Reading

Björk A (1996) Numerical methods for least squares problems. SIAM, Philadelphia
Kariya T (2004) Generalized least squares. Wiley, New York
Rao CR, Toutenberg H (1995) Linear models, least squares and alternatives. Springer, New York
Van Huffel S, Vandewalle J (1991) The total least squares problem: computational aspects and analysis. SIAM, Philadelphia
Wolberg J (2005) Data analysis using the method of least squares: extracting the most information from experiments. Springer, New York

Lévy Processes

Mohamed Abdel-Hameed
Professor
United Arab Emirates University, Al Ain, United Arab Emirates

Introduction

Lévy processes have become increasingly popular in engineering (reliability, dams, telecommunication) and mathematical finance. Their applications in reliability stems from the fact that they provide a realistic model for the degradation of devices, while their applications in the mathematical theory of dams as they provide a basis for describing the water input of dams. Their popularity in finance is because they describe the financial markets in a more accurate way than the celebrated Black–Scholes model. The latter model assumes that the rate of returns on assets are normally distributed, thus the process describing the asset price over time is continuous process. In reality, the asset prices have jumps or spikes, and the asset returns exhibit fat tails and ►skewness, which negates the normality assumption inherited in the Black–Scholes model.

Because of the deficiencies in the Black–Scholes model researchers in mathematical finance have been trying to find more suitable models for asset prices. Certain types of Lévy processes have been found to provide a good model for creep of concrete, fatigue crack growth, corroded steel gates, and chloride ingress into concrete. Furthermore, certain types of Lévy processes have been used to model the water input in dams.

In this entry, we will review Lévy processes and give important examples of such processes and state some references to their applications.

Lévy Processes

A stochastic process $X = \{X_t, t \geq 0\}$ that has right continuous sample paths with left limits is said to be a Lévy process if the following hold:

1. X has *stationary increments*, i.e., for every $s, t \geq 0$, the distribution of $X_{t+s} - X_t$ is independent of t.
2. X has independent increments, i.e., for every $t, s \geq 0$, $X_{t+s} - X_t$ is independent of $(X_u, u \leq t)$.
3. X is stochastically continuous, i.e., for every $t \geq 0$ and $\epsilon > 0$:

$$lim_{s \to t} P(|X_t - X_s| > \epsilon) = 0.$$

That is to say a Lévy process is a stochastically continuous process with stationary and independent increments whose sample paths are right continuous with left hand limits.

If $\Phi(z)$ is the characteristic function of a Lévy process, then its characteristic component $\varphi(z) \stackrel{def}{=} \frac{\ln \Phi(z)}{t}$ is of the form

$$\left\{ iza - \frac{z^2 b}{2} + \int_R [\exp(izx) - 1 - izx I_{\{|x|<1\}}] v(dx) \right\}$$

where $a \in R$, $b \in R_+$ and v is a measure on R satisfying $v(\{0\}) = 0, \int_R (1 \wedge x^2) v(dx) < \infty$.

The measure v characterizes the size and frequency of the jumps. If the measure is infinite, then the process has infinitely many jumps of very small sizes in any small interval. The constant a defined above is called the drift term of the process, and b is the variance (volatility) term.

The Lévy–Itô decomposition identify any Lévy process as the sum of three independent processes, it is stated as follows:

Given any $a \in R$, $b \in R_+$ and measure v on R satisfying $v(\{0\}) = 0, \int_R (1 \wedge x^2) v(dx) < \infty$, there exists a probability space (Ω, F, P) on which a *Lévy* process X is defined. The process X is the sum of three independent processes $\overset{(1)}{X}$, $\overset{(2)}{X}$, and $\overset{(3)}{X}$, where $\overset{(1)}{X}$ is a Brownian motion with drift a and volatility b (in the sense defined below), $\overset{(2)}{X}$ is a compound Poisson process, and $\overset{(3)}{X}$ is a square integrable martingale. The characteristic components of $\overset{(1)}{X}$, $\overset{(2)}{X}$, and $\overset{(3)}{X}$ (denoted by $\overset{(1)}{\varphi}(z)$, $\overset{(2)}{\varphi}(z)$ and $\overset{(3)}{\varphi}(z)$, respectively) are as follows:

$$\overset{(1)}{\varphi}(z) = iza - \tfrac{z^2 b}{2},$$
$$\overset{(2)}{\varphi}(z) = \int_{\{|x| \geq 1\}} (\exp(izx) - 1) v(dx),$$
$$\overset{(3)}{\varphi}(z) = \int_{\{|x|<1\}} (\exp(izx) - 1 - izx) v(dx).$$

Examples of the Lévy Processes

The Brownian Motion

A Lévy process is said to be a Brownian motion (see ►Brownian Motion and Diffusions) with drift μ, and volatility rate σ^2, if $\mu = a$, $b = \sigma^2$, and $v(R) = 0$. Brownian motion is the only nondeterministic Lévy processes with continuous sample paths.

The Inverse Brownian Process

Let X be a Brownian motion with $\mu > 0$ and volatility rate σ^2. For any $x > 0$, let $T_x = \inf\{t : X_t > x\}$. Then T_x is an increasing Lévy process (called inverse Brownian motion), its Lévy measure is given by

$$v(dx) = \frac{1}{\sqrt{2\pi\sigma^2 x^3}} \exp\left(\frac{-x\mu^2}{2\sigma^2}\right).$$

The Compound Poisson Process

The compound Poisson process (see ►Poisson Processes) is a Lévy process where $b = 0$ and v is a finite measure.

The Gamma Process

The gamma process is a nonnegative increasing Lévy process X, where $b = 0$, $a - \int_0^1 x v(dx) = 0$ and its Lévy measure is given by

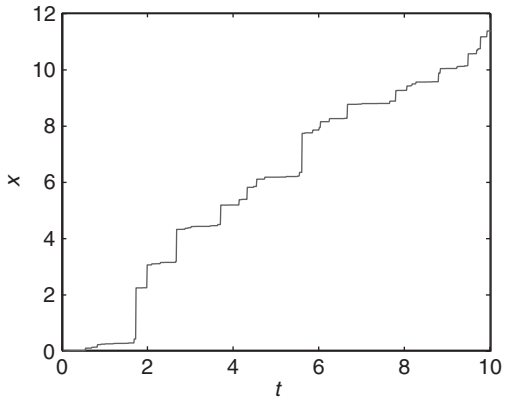
$$v(dx) = \frac{\alpha}{x} exp(-x/\beta) dx, \qquad x > 0$$

where $\alpha, \beta > 0$. It follows that the mean term $(E(X_1))$ and the variance term $(V(X_1))$ for the process are equal to $\alpha\beta$ and $\alpha\beta^2$, respectively.

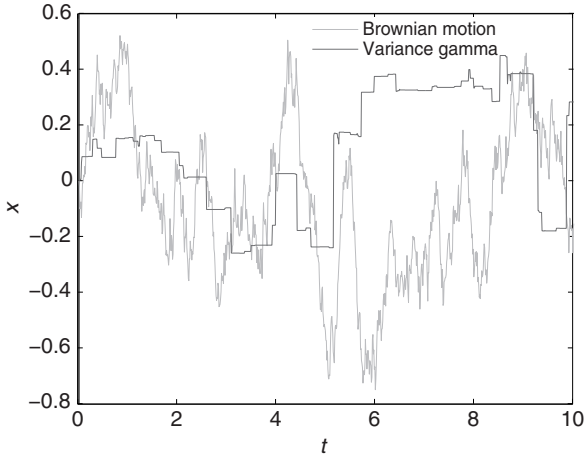
The following is a simulated sample path of a gamma process, where $\alpha = 2$ and $\beta = 0.5$ (Fig. 1).

The Variance Gamma Process

The variance gamma process is a Lévy process that can be represented as either the difference between two independent gamma processes or as a Brownian process subordinated by a gamma process. The latter is accomplished by a random time change, replacing the time of the Brownian process by a gamma process, with a mean term equal to 1. The variance gamma process has three parameters: μ – the



Lévy Processes. Fig. 1 Gamma process



Lévy Processes. Fig. 2 Brownian motion and variance gamma sample paths

Brownian process drift term, σ – the volatility of the Brownian process, and v – the variance term of the the gamma process.

The following are two simulated sample paths, one for a Brownian motion with a drift term $\mu = 0.2$ and volatility term $\sigma = 0.5$ and the other is for a variance gamma process with the same values for the drift term and the volatility terms and $v = 1$ (Fig. 2).

About the Author

Professor Mohamed Abdel-Hameed was the chairman of the Department of Statistics and Operations Research, Kuwait University (1988–1989). He was on the editorial board of *Applied Stochastic Models and Data Analysis* (1983–1990). He was the Editor (jointly with E. Cinlar and J. Quinn) of the text *Survival Models, Maintenance, Replacement Policies and Accelerated Life Testing* (Academic Press 1984). He is an elected member of the International Statistical Institute. He has published numerous papers in Statistics and Operations research.

Cross References

►Non-Uniform Random Variate Generations
►Poisson Processes
►Statistical Modeling of Financial Markets
►Stochastic Processes
►Stochastic Processes: Classification

References and Further Reading

Abdel-Hameed MS (1984) Life distribution of devices subject to a Lévy wear process. Math Oper Res 9:606–614

Abdel-Hameed M (2000) Optimal control of a dam using $P^M_{\lambda,\tau}$ policies and penalty cost when the input process is a compound Poisson process with a positive drift. J Appl Prob 37: 408–416

Black F, Scholes MS (1973) The pricing of options and corporate liabilities. J Pol Econ 81:637–654

Cont R, Tankov P (2003) Financial modeling with jump processes. Chapman & Hall/CRC Press, Boca Raton

Madan DB, Carr PP, Chang EC (1998) The variance gamma process and option pricing. Eur Finance Rev 2:79–105

Patel A, Kosko B (2008) Stochastic resonance in continuous and spiking Neutron models with Lévy noise. IEEE Trans Neural Netw 19:1993–2008

van Noortwijk JM (2009) A survey of the applications of gamma processes in maintenance. Reliab Eng Syst Safety 94:2–21

Life Expectancy

Maja Biljan-August
Full Professor
University of Rijeka, Rijeka, Croatia

Life expectancy is defined as the average number of years a person is expected to live from age x, as determined by statistics.

Statistics on life expectancy are derived from a mathematical model known as the ►life table. In order to calculate this indicator, the mortality rate at each age is assumed to be constant. Life expectancy (e_x) can be evaluated at any age and, in a hypothetical stationary population, can be written in discrete form as:

$$e_x = \frac{T_x}{l_x}$$

where x is age; T_x is the number of person-years lived aged x and over; and l_x is the number of survivors at age x according to the life table.

Life expectancy can be calculated for combined sexes or separately for males and females. There can be significant differences between sexes.

Life expectancy at birth (e_0) is the average number of years a newborn child can expect to live if current mortality trends remain constant:

$$e_0 = \frac{T_0}{l_0}$$

where T_0 is the total size of the population and l_0 is the number of births (the original number of persons in the birth cohort).

Life expectancy declines with age. Life expectancy at birth is highly influenced by infant mortality rate. *The paradox of the life table* refers to a situation where life expectancy at birth increases for several years after birth ($e_0 < e_1 < \ldots e_5$ and even beyond). The paradox reflects the higher rates of infant and child mortality in populations in pre-transition and middle stages of the demographic transition.

Life expectancy at birth is a summary measure of mortality in a population. It is a frequently used indicator of health standards and socio-economic living standards. Life expectancy is also one of the most commonly used indicators of social development. This indicator is easily comparable through time and between areas, including countries. Inequalities in life expectancy usually indicate inequalities in health and socio-economic development.

Life expectancy rose throughout human history. In ancient Greece and Rome, the average life expectancy was below 30 years; between the years 1800 and 2000, life expectancy at birth rose from about 30 years to a global average of 67 years, and to more than 75 years in the richest countries (Riley 2001). Furthermore, in most industrialized countries, in the early twenty-first century, life expectancy averaged at about 78 years (WHO). These changes, called the "health transition," are essentially the result of improvements in public health, medicine, and nutrition.

Life expectancy varies significantly across regions and continents: from life expectancies of about 40 years in some central African populations to life expectancies of 80 years and above in many European countries. The more developed regions have an average life expectancy of 76 years, while the population of less developed regions is at birth expected to live an average 12 years less. The two continents that display the most extreme differences in life expectancies are North America (77.6 years) and Africa (49.1 years) where, as of recently, the gap between life expectancies amounts to 29 years (UN, 2000–2005 data).

Countries with the highest life expectancies in the world (82 years) are Australia, Iceland, Italy, Switzerland, and Japan (83 years); Japanese men and women live an average of 79 and 86 years, respectively (WHO 2009).

In countries with a high rate of HIV infection, principally in Sub-Saharan Africa, the average life expectancy is 45 years and below. Some of the world's lowest life expectancies are in Sierra Leone (41 years), Afghanistan (42 years), Lesotho (45 years), and Zimbabwe (45 years).

In nearly all countries, women live longer than men. The world's average life expectancy at birth is 65 years for males and 70 years for females; the gap is about five years. The female-to-male gap is expected to narrow in the more developed regions and widen in the less developed regions. The Russian Federation has the greatest difference in life expectancies between the sexes (13 years less for men), whereas in Tonga, life expectancy for males exceeds that for females by 2 years (WHO 2009).

Life expectancy is assumed to rise continuously. According to estimation by the UN, global life expectancy at birth is likely to rise to an average 74 years by 2045–2050. By 2100, life expectancy is expected to vary across countries from 66 to 97 years. Long-range United Nations population projections predict that by 2300, on average, people can expect to live more than 95 years, from 87 (Liberia) up to 106 years (Japan).

For more details on the calculation of life expectancy, including continuous notation, see Keyfitz (1968, 2005) and Preston et al. (2003).

Cross References

►Biopharmaceutical Research, Statistics in
►Biostatistics
►Demography
►Life Table
►Mean Residual Life
►Measurement of Economic Progress

References and Further Reading

Keyfitz N (1968) Introduction to the Mathematics of Population. Addison-Wesly, Massachusetts

Keyfitz N (2005) Applied mathematical demography. Springer, New York

Preston SH, Heuveline P, Guillot M (2003) Demography: measuring and modelling potion processes. Blackwell, Oxford

Riley JC (2001) Rising life expectancy: a global history. Cambridge University Press, Cambridge

Rowland DT (2003) Demographic methods and concepts. Oxford University Press, New York

Siegel JS, Swanson DA (2004) The methods and materials of demography, 2nd edn. Elsevier Academic Press, Amsterdam

World Health Organization (2009) World health statistics 2009. Geneva

World Population Prospects: The 2004 revision. Highlights. United Nations, New York
World Population to 2300 (2004) United Nations, New York

Life Table

Jan M. Hoem
Professor, Director Emeritus
Max Planck Institute for Demographic Research, Rostock, Germany

The life table is a classical tabular representation of central features of the distribution function F of a positive variable, say X, which normally is taken to represent the lifetime of a newborn individual. The life table was introduced well before modern conventions concerning statistical distributions were developed, and it comes with some special terminology and notation, as follows. Suppose that F has a density $f(x) = \frac{d}{dx}F(x)$ and define *the force of mortality* or *death intensity* at age x as the function

$$\mu(x) = -\frac{d}{dx}\ln\{1 - F(x)\} = \frac{f(x)}{\{1 - F(x)\}}.$$

Heuristically, it is interpreted by the relation $\mu(x)\mathrm{d}x = P\{x < X < x + \mathrm{d}x | X > x\}$. Conversely $F(x) = 1 - \exp\left\{-\int_0^x \mu(s)ds\right\}$. The *survivor function* is defined as $\ell(x) = \ell(0)\{1 - F(x)\}$, normally with $\ell(0) = 100,000$. In mortality applications $\ell(x)$ is the expected number of survivors to exact age x out of an original cohort of 100,000 newborn babies. The *survival probability* is

$$\begin{aligned} {}_tp_x &= P\{X > x + t | X > x\} = \ell(x + t)/\ell(x) \\ &= \exp\left\{-\int_0^t \mu(x + s)ds\right\}, \end{aligned}$$

and the non-survival probability is (the converse) ${}_tq_x = 1 - {}_tp_x$. For $t = 1$ one writes $q_x = {}_1q_x$ and $p_x = {}_1p_x$. In particular, we get $\ell(x + 1) = \ell(x)p_x$. This is a practical recursion formula that permits us to compute all values of $\ell(x)$ once we know the values of p_x for all relevant x.

The life expectancy is $e_0^o = EX = \int_0^\infty \ell(x)dx/\ell(0)$ (The subscript 0 in e_0^o indicates that the expected value is computed at age 0 (i.e., for newborn individuals) and the superscript o indicates that the computation is made in the continuous mode.). The remaining life expectancy at age x is:

$$e_x^o = E(X - x | X > x) = \int_0^\infty \ell(x + t)dt/\ell(x),$$

i.e., it is the expected lifetime remaining to someone who has attained age x.

To turn to the statistical estimation of these various quantities, suppose that the function $\mu(x)$ is piecewise constant, which means that we take it to equal some constant, say μ_j, over each interval (x_j, x_{j+1}) for some partition $\{x_j\}$ of the age axis. For a collection $\{X_i\}$ of independent observations of X, let D_j be the number of X_i that fall in the interval (x_j, x_{j+1}). In mortality applications, this is the number of deaths observed in the given interval. For the cohort of the initially newborn, D_j is the number of individuals who die in the interval (called the *occurrences* in the interval). If individual i dies in the interval, he or she will of course have lived for $X_i - x_j$ time units *during* the interval. Individuals who survive the interval, will have lived for $x_{j+1} - x_j$ time units in the interval, and individuals who do not survive to age x_j, will not have lived during this interval at all. When we aggregate the time units lived in (x_j, x_{j+1}) over all individuals, we get a total R_j which is called the *exposures* for the interval, the idea being that individuals are *exposed to the risk* of death for as long as they live in the interval. In the simple case where there are no relations between the individual parameters μ_j, the collection $\{D_j, R_j\}$ constitutes a statistically sufficient set of observations with a likelihood Λ that satisfies the relation $\ln\Lambda = \sum_j\{-\mu_j R_j + D_j \ln\mu_j\}$ which is easily seen to be maximized by $\hat{\mu}_j = D_j/R_j$. The latter fraction is therefore the maximum-likelihood estimator for μ_j (In some connections an age schedule of mortality will be specified, such as the classical Gompertz–Makeham function $\mu_x = a + bc^x$, which does represent a relationship between the intensity values at the different ages x, normally for single-year age groups. Maximum likelihood estimators can then be found by plugging this functional specification of the intensities into the likelihood function, finding the values $\hat{a}, \hat{b}$, and $\hat{c}$ that maximize Λ, and using $\hat{a} + \hat{b}\hat{c}^x$ for the intensity in the rest of the life table computations. Methods that do not amount to maximum likelihood estimation will sometimes be used because they involve simpler computations. With some luck they provide starting values for the iterative process that must usually be applied to produce the maximum likelihood estimators. For an example, see Forsén (1979)). This whole schema can be extended trivially to cover censoring (*withdrawals*) provided the censoring mechanism is unrelated to the mortality process.

If the force of mortality is constant over a single-year age interval $(x, x+1)$, say, and is estimated by $\hat{\mu}_x$ in this interval, then $\hat{p}_x = e^{-\hat{\mu}_x}$ is an estimator of the single-year survival probability p_x. This allows us to estimate the survival function recursively for all corresponding ages, using $\hat{\ell}(x+1) = \hat{\ell}(x)\hat{p}_x$ for $x = 0, 1, \ldots$, and the rest of the life table computations follow suit. Life table construction consists in the estimation of the parameters and the tabulation of functions like those above from empirical data. The data can be for age at death for individuals, as in the example indicated above, but they can also be observations of duration until recovery from an illness, of intervals between births, of time until breakdown of some piece of machinery, or of any other positive duration variable.

So far we have argued as if the life table is computed for a group of mutually independent individuals who have all been observed in parallel, essentially a cohort that is followed from a significant common starting point (namely from birth in our mortality example) and which is diminished over time due to *decrements* (*attrition*) caused by the risk in question and also subject to reduction due to censoring (withdrawals). The corresponding table is then called a *cohort life table*. It is more common, however, to estimate a $\{p_x\}$ schedule from data collected for the members of a population during a limited time period and to use the mechanics of life-table construction to produce a *period life table* from the p_x values.

Life table techniques are described in detail in most introductory textbooks in actuarial statistics, ►biostatistics, ►demography, and epidemiology. See, e.g., Chiang (1984), Elandt-Johnson and Johnson (1980), Manton and Stallard (1984), Preston et al. (2001). For the history of the topic, consult Seal (1977), Smith and Keyfitz (1977), and Dupâquier (1996).

About the Author

For biography *see* the entry ►Demography.

Cross References

►Demographic Analysis: A Stochastic Approach
►Demography
►Event History Analysis
►Kaplan-Meier Estimator
►Population Projections
►Statistics: An Overview

References and Further Reading

Chiang CL (1984) The life table and its applications. Krieger, Malabar
Dupâquier J (1996) L'invention de la table de mortalité. Presses universitaires de France, Paris
Elandt-Johnson RC, Johnson NL (1980) Survival models and data analysis. Wiley, New York
Forsén L (1979) The efficiency of selected moment methods in Gompertz–Makeham graduation of mortality. Scand Actuarial J 167–178
Manton KG, Stallard E (1984) Recent trends in mortality analysis. Academic Press, Orlando
Preston SH, Heuveline P, Guillot M (2001) Demography. Measuring and modeling populations. Blackwell, Oxford
Seal H (1977) Studies in history of probability and statistics, 35: multiple decrements of competing risks. Biometrica 63(3):429–439
Smith D, Keyfitz N (1977) Mathematical demography: selected papers. Springer, Heidelberg

Likelihood

Nancy Reid
Professor
University of Toronto, Toronto, ON, Canada

Introduction

The likelihood function in a statistical model is proportional to the density function for the random variable to be observed in the model. Most often in applications of likelihood we have a parametric model $f(y;\theta)$, where the parameter θ is assumed to take values in a subset of $\mathbb{R}^k$, and the variable y is assumed to take values in a subset of $\mathbb{R}^n$: the likelihood function is defined by

$$L(\theta) = L(\theta; y) = cf(y;\theta), \tag{1}$$

where c can depend on y but not on θ. In more general settings where the model is semi-parametric or non-parametric the explicit definition is more difficult, because the density needs to be defined relative to a dominating measure, which may not exist: see Van der Vaart (1996) and Murphy and Van der Vaart (1997). This article will consider only finite-dimensional parametric models.

Within the context of the given parametric model, the likelihood function measures the relative plausibility of various values of θ, for a given observed data point y. Values of the likelihood function are only meaningful relative to each other, and for this reason are sometimes standardized by the maximum value of the likelihood function, although other reference points might be of interest depending on the context.

If our model is $f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, y = 0, 1, \ldots, n;$ $\theta \in [0,1]$, then the likelihood function is (any function proportional to)

$$L(\theta; y) = \theta^y(1-\theta)^{n-y}$$

and can be plotted as a function of θ for any fixed value of y. The likelihood function is maximized at $\theta = y/n$. This model might be appropriate for a sampling scheme which recorded the number of successes among n independent trials that result in success or failure, each trial having the same probability of success, θ. Another example is the likelihood function for the mean and variance parameters when sampling from a normal distribution with mean μ and variance σ^2:

$$L(\theta; y) = \exp\{-n \log \sigma - (1/2\sigma^2)\Sigma(y_i - \mu)^2\},$$

where $\theta = (\mu, \sigma^2)$. This could also be plotted as a function of μ and σ^2 for a given sample $y_1, \ldots, y_n$, and it is not difficult to show that this likelihood function only depends on the sample through the sample mean $\bar{y} = n^{-1}\Sigma y_i$ and sample variance $s^2 = (n-1)^{-1}\Sigma(y_i - \bar{y})^2$, or equivalently through Σy_i and Σy_i^2. It is a general property of likelihood functions that they depend on the data only through the minimal sufficient statistic.

Inference

The likelihood function was defined in a seminal paper of Fisher (1922), and has since become the basis for most methods of statistical inference. One version of likelihood inference, suggested by Fisher, is to use some rule such as $L(\hat{\theta})/L(\theta) > k$ to define a range of "likely" or "plausible" values of θ. Many authors, including Royall (1997) and Edwards (1960), have promoted the use of plots of the likelihood function, along with interval estimates of plausible values. This approach is somewhat limited, however, as it requires that θ have dimension 1 or possibly 2, or that a likelihood function can be constructed that only depends on a component of θ that is of interest; see section "►Nuisance Parameters" below.

In general, we would wish to calibrate our inference for θ by referring to the probabilistic properties of the inferential method. One way to do this is to introduce a probability measure on the unknown parameter θ, typically called a prior distribution, and use Bayes' rule for conditional probabilities to conclude

$$\pi(\theta \mid y) = L(\theta; y)\pi(\theta) / \int_\theta L(\theta; y)\pi(\theta)d\theta,$$

where $\pi(\theta)$ is the density for the prior measure, and $\pi(\theta \mid y)$ provides a probabilistic assessment of θ after observing $Y = y$ in the model $f(y; \theta)$. We could then make conclusions of the form, "having observed 5 successes in 20 trials, and assuming $\pi(\theta) = 1$, the posterior probability that $\theta > 0.5$ is 0.013," and so on.

This is a very brief description of Bayesian inference, in which probability statements refer to that generated from the prior through the likelihood to the posterior. A major difficulty with this approach is the choice of prior probability function. In some applications there may be an accumulation of previous data that can be incorporated into a probability distribution, but in general there is not, and some rather *ad hoc* choices are often made. Another difficulty is the meaning to be attached to probability statements about the parameter.

Inference based on the likelihood function can also be calibrated with reference to the probability model $f(y; \theta)$, by examining the distribution of $L(\theta; Y)$ as a random function, or more usually, by examining the distribution of various derived quantities. This is the basis for likelihood inference from a frequentist point of view. In particular, it can be shown that $2\log\{L(\hat{\theta}; Y)/L(\theta; Y)\}$, where $\hat{\theta} = \hat{\theta}(Y)$ is the value of θ at which $L(\theta; Y)$ is maximized, is approximately distributed as a χ^2_k random variable, where k is the dimension of θ. To make this precise requires an asymptotic theory for likelihood, which is based on a central limit theorem (see ►Central Limit Theorems) for the *score function*

$$U(\theta; Y) = \frac{\partial}{\partial\theta} \log L(\theta; Y).$$

If $Y = (Y_1, \ldots, Y_n)$ has independent components, then $U(\theta)$ is a sum of n independent components, which under mild regularity conditions will be asymptotically normal. To obtain the χ^2 result quoted above it is also necessary to investigate the convergence of $\hat{\theta}$ to the true value governing the model $f(y; \theta)$. Showing this convergence, usually in probability, but sometimes almost surely, can be difficult: see Scholz (2006) for a summary of some of the issues that arise.

Assuming that $\hat{\theta}$ is consistent for θ, and that $L(\theta; Y)$ has sufficient regularity, the follow asymptotic results can be established:

$$(\hat{\theta} - \theta)^T i(\theta)(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_k, \tag{2}$$

$$U(\theta)^T i^{-1}(\theta) U(\theta) \xrightarrow{d} \chi^2_k, \tag{3}$$

$$2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{d} \chi^2_k, \tag{4}$$

where $i(\theta) = E\{-\ell''(\theta; Y); \theta\}$ is the expected Fisher information function, $\ell(\theta) = \log L(\theta)$ is the log-likelihood function, and χ^2_k is the ►chi-square distribution with k degrees of freedom.

These results are all versions of a more general result that the log-likelihood function converges to the quadratic form corresponding to a multivariate normal distribution (see ►Multivariate Normal Distributions), under suitably stated limiting conditions. There is a similar asymptotic result showing that the posterior density is asymptotically

normal, and in fact asymptotically free of the prior distribution, although this result requires that the prior distribution be a proper probability density, i.e., has integral over the parameter space equal to 1.

Nuisance Parameters

In models where the dimension of θ is large, plotting the likelihood function is not possible, and inference based on the multivariate normal distribution for $\hat{\theta}$ or the χ^2_k distribution of the log-likelihood ratio doesn't lead easily to interval estimates for components of θ. However it is possible to use the likelihood function to construct inference for parameters of interest, using various methods that have been proposed to eliminate nuisance parameters.

Suppose in the model $f(y;\theta)$ that $\theta = (\psi, \lambda)$, where ψ is a k_1-dimensional parameter of interest (which will often be 1). The *profile log-likelihood* function of ψ is

$$\ell_P(\psi) = \ell(\psi, \hat{\lambda}_\psi),$$

where $\hat{\lambda}_\psi$ is the *constrained* maximum likelihood estimate: it maximizes the likelihood function $L(\psi, \lambda)$ when ψ is held fixed. The profile log-likelihood function is also called the concentrated log-likelihood function, especially in econometrics. If the parameter of interest is not expressed explicitly as a subvector of θ, then the constrained maximum likelihood estimate is found using Lagrange multipliers.

It can be verified under suitable smoothness conditions that results similar to those at (2 – 4) hold as well for the profile log-likelihood function: in particular

$$2\{\ell_P(\hat{\psi}) - \ell_P(\psi)\} = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\} \xrightarrow{d} \chi^2_{k_1},$$

This method of eliminating nuisance parameters is not completely satisfactory, especially when there are many nuisance parameters: in particular it doesn't allow for errors in estimation of λ. For example the profile likelihood approach to estimation of σ^2 in the linear regression model (see ▶Linear Regression Models) $y \sim N(X\beta, \sigma^2)$ will lead to the estimator $\hat{\sigma}^2 = \Sigma(y_i - \hat{y}_i)^2/n$, whereas the estimator usually preferred has divisor $n-p$, where p is the dimension of β.

Thus a large literature has developed on improvements to the profile log-likelihood. For Bayesian inference such improvements are "automatically" included in the formulation of the marginal posterior density for ψ:

$$\pi_M(\psi \mid y) \propto \int \pi(\psi, \lambda \mid y) d\lambda,$$

but it is typically quite difficult to specify priors for possibly high-dimensional nuisance parameters. For non-Bayesian inference most modifications to the profile log-likelihood are derived by considering conditional or marginal inference in models that admit factorizations, at least approximately, like the following:

$$f(y;\theta) = f_1(y_1;\psi) f_2(y_2 \mid y_1;\lambda), \quad \text{or}$$
$$f(y;\theta) = f_1(y_1 \mid y_2;\psi) f_2(y_2;\lambda).$$

A discussion of conditional inference and density factorisations is given in Reid (1995). This literature is closely tied to that on higher order asymptotic theory for likelihood. The latter theory builds on saddlepoint and Laplace expansions to derive more accurate versions of (2–4): see, for example, Severini (2000) and Brazzale et al. (2007). The direct likelihood approach of Royall (1997) and others does not generalize very well to the nuisance parameter setting, although Royall and Tsou (2003) present some results in this direction.

Extensions to Likelihood

The likelihood function is such an important aspect of inference based on models that it has been extended to "likelihood-like" functions for more complex data settings. Examples include nonparametric and semi-parametric likelihoods: the most famous semi-parametric likelihood is the proportional hazards model of Cox (1972). But many other extensions have been suggested: to empirical likelihood (Owen 1988), which is a type of nonparametric likelihood supported on the observed sample; to quasi-likelihood (McCullagh 1983) which starts from the score function $U(\theta)$ and works backwards to an inference function; to bootstrap likelihood (Davison et al. 1992); and many modifications of profile likelihood (Barndorff-Nielsen and Cox 1994; Fraser 2003). There is recent interest for multi-dimensional responses Y_i in composite likelihoods, which are products of lower dimensional conditional or marginal distributions (Varin 2008). Reid (2000) concluded a review article on likelihood by stating:

> From either a Bayesian or frequentist perspective, the likelihood function plays an essential role in inference. The maximum likelihood estimator, once regarded on an equal footing among competing point estimators, is now typically the estimator of choice, although some refinement is needed in problems with large numbers of nuisance parameters. The likelihood ratio statistic is the basis for most tests of hypotheses and interval estimates. The emergence of the centrality of the likelihood function for inference, partly due to the large increase in computing power, is one of the central developments in the theory of statistics during the latter half of the twentieth century.

Further Reading

The book by Cox and Hinkley (1974) gives a detailed account of likelihood inference and principles of statistical inference; see also Cox (2006). There are several book-length treatments of likelihood inference, including Edwards (1960), Azzalini (1998), Pawitan (2000), and Severini (2000): this last discusses higher order asymptotic theory in detail, as does Barndorff-Nielsen and Cox (1994), and Brazzale, Davison and Reid (2007). A short review paper is Reid (2000). An excellent overview of consistency results for maximum likelihood estimators is Scholz (2006); see also Lehmann and Casella (1998). Foundational issues surrounding likelihood inference are discussed in Berger and Wolpert (1980).

About the Author

Professor Reid is a Past President of the Statistical Society of Canada (2004–2005). During (1996–1997) she served as the President of the Institute of Mathematical Statistics. Among many awards, she received the Emanuel and Carol Parzen Prize for Statistical Innovation (2008) "for leadership in statistical science, for outstanding research in theoretical statistics and highly accurate inference from the likelihood function, and for influential contributions to statistical methods in biology, environmental science, high energy physics, and complex social surveys." She was awarded the Gold Medal, Statistical Society of Canada (2009) and Florence Nightingale David Award, Committee of Presidents of Statistical Societies (2009). She is Associate Editor of *Statistical Science*, (2008–), *Bernoulli* (2007–) and *Metrika* (2008–).

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Chi-Square Test: Analysis of Contingency Tables
►Empirical Likelihood Approach to Inference from Sample Survey Data
►Estimation
►Fiducial Inference
►General Linear Models
►Generalized Linear Models
►Generalized Quasi-Likelihood (GQL) Inferences
►Mixture Models
►Philosophical Foundations of Statistics
►Risk Analysis
►Statistical Evidence
►Statistical Inference
►Statistical Inference: An Overview
►Statistics: An Overview
►Statistics: Nelder's view
►Testing Variance Components in Mixed Linear Models
►Uniform Distribution in Statistics

References and Further Reading

Azzalini A (1998) Statistical inference. Chapman and Hall, London
Barndorff-Nielsen OE, Cox DR (1994) Inference and asymptotics. Chapman and Hall, London
Berger JO, Wolpert R (1980) The likelihood principle. Institute of Mathematical Statistics, Hayward
Birnbaum A (1962) On the foundations of statistical inference. Am Stat Assoc 57:269–306
Brazzale AR, Davison AC, Reid N (2007) Applied asymptotics. Cambridge University Press, Cambridge
Cox DR (1972) Regression models and life tables. J R Stat Soc B 34:187–220 (with discussion)
Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge
Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, London
Davison AC, Hinkley DV, Worton B (1992) Bootstrap likelihoods. Biometrika 79:133–130
Edwards AF (1960) Likelihood. Oxford University Press, Oxford
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Phil Trans R Soc A 222:309–368
Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer, New York
McCullagh P (1983) Quasi-likelihood functions. Ann Stat 11:59–67
Murphy SA, Van der Vaart A (1997) Semiparametric likelihood ratio inference. Ann Stat 25:1471–1509
Owen A (1988) Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75:237–249
Pawitan Y (2000) In all likelihood. Oxford University Press, Oxford
Reid N (1995) The roles of conditioning in inference. Stat Sci 10:138–157
Reid N (2000) Likelihood. J Am Stat Assoc 95:1335–1340
Royall RM (1997) Statistical evidence: a likelihood paradigm. Chapman and Hall, London
Royall RM, Tsou TS (2003) Interpreting statistical evidence using imperfect models: robust adjusted likelihoodfunctions. J R Stat Soc B 65:391404
Scholz F (2006) Maximum likelihood estimation. In: Encyclopedia of statistical sciences. Wiley, New York, doi: 10.1002/0471667196.ess1571.pub2. Accessed 23 Aug 2009
Severini TA (2000) Likelihood methods in statistics. Oxford University Press, Oxford
Van der Vaart AW (1996) Infinite-dimensional likelihood methods in statistics. http://www.stieltjes.org/archief/biennial9596/frame/node17.html. Accessed 18 Aug 2009
Varin C (2008) On composite marginal likelihood. Adv Stat Anal 92:1–28

Limit Theorems of Probability Theory

Alexandr Alekseevich Borovkov
Professor, Head of Probability and Statistics Chair at the Novosibirsk University
Novosibirsk University, Novosibirsk, Russia

Limit Theorems of Probability Theory is a broad name referring to the most essential and extensive research area in Probability Theory which, at the same time, has the greatest impact on the numerous applications of the latter.

By its very nature, Probability Theory is concerned with asymptotic (limiting) laws that emerge in a long series of observations on random events. Because of this, in the early twentieth century even the very definition of probability of an event was given by a group of specialists (R. von Mises and some others) as the limit of the relative frequency of the occurrence of this event in a long row of independent random experiments. The "stability" of this frequency (i.e., that such a limit always exists) was postulated. After the 1930s, Kolmogorov's axiomatic construction of probability theory has prevailed. One of the main assertions in this axiomatic theory is the *Law of Large Numbers* (LLN) on the convergence of the averages of large numbers of random variables to their expectation. This law implies the aforementioned stability of the relative frequencies and their convergence to the probability of the corresponding event.

The LLN is the simplest limit theorem (LT) of probability theory, elucidating the physical meaning of probability. The LLN is stated as follows: if $X, X_1, X_2, \ldots$ is a sequence of i.i.d. random variables,

$$S_n := \sum_{j=1}^{n} X_j,$$

and the expectation $a := \mathbf{E}X$ exists then $n^{-1}S_n \xrightarrow{a.s.} a$ (almost surely, i.e., with probability 1). Thus the value na can be called the *first order approximation* for the sums S_n. The *Central Limit Theorem* (CLT) gives one a more precise approximation for S_n. It says that, if $\sigma^2 := \mathbf{E}(X-a)^2 < \infty$, then the distribution of the standardized sum $\zeta_n := (S_n - na)/\sigma\sqrt{n}$ converges, as $n \to \infty$, to the standard normal (Gaussian) law. That is, for all x,

$$\mathbf{P}(\zeta_n < x) \to \Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt.$$

The quantity $n\mathbf{E}\xi + \zeta\sigma\sqrt{n}$, where ζ is a standard normal random variable (so that $\mathbf{P}(\zeta < x) = \Phi(x)$), can be called *the second order approximation* for S_n.

The first LLN (for the Bernoulli scheme) was proved by Jacob Bernoulli in the late 1690s (published posthumously in 1713). The first CLT (also for the Bernoulli scheme) was established by A. de Moivre (first published in 1738 and referred nowadays to as the de Moivre–Laplace theorem). In the beginning of the nineteenth century, P.S. Laplace and C.F. Gauss contributed to the generalization of these assertions and appreciation of their enormous applied importance (in particular, for the *theory of errors of observations*), while later in that century further breakthroughs in both methodology and applicability range of the CLT were achieved by P.L. Chebyshev (1887) and A.M. Lyapunov (1900).

The main directions in which the two aforementioned main LTs have been extended and refined since then are:

1. Relaxing the assumption $\mathbf{E}X^2 < \infty$. When the second moment is infinite, one needs to assume that the "tail" $P(x) := \mathbf{P}(X > x) + \mathbf{P}(X < -x)$ is a function regularly varying at infinity such that the limit $\lim_{x\to\infty} \mathbf{P}(X > x)/P(x)$ exists. Then the distribution of the normalized sum $S_n/\sigma(n)$, where $\sigma(n) := P^{-1}(n^{-1})$, P^{-1} being the generalized inverse of the function P, and we assume that $\mathbf{E}\xi = 0$ when the expectation is finite, converges to one of the so-called stable laws as $n \to \infty$. The ▶characteristic functions of these laws have simple closed-form representations.
2. Relaxing the assumption that the X_j's are identically distributed and proceeding to study the so-called *triangular array scheme*, where the distributions of the summands $X_j = X_{j,n}$ forming the sum S_n depend not only on j but on n as well. In this case, the class of all limit laws for the distribution of S_n (under suitable normalization) is substantially wider: it coincides with the class of the so-called infinitely divisible distributions. An important special case here is the Poisson limit theorem on convergence in distribution of the number of occurrences of rare events to a Poisson law.
3. Relaxing the assumption of independence of the X_j's. Several types of "weak dependence" assumptions on X_j under which the LLN and CLT still hold true have been suggested and investigated. One should also mention here the so-called ergodic theorems (see ▶Ergodic Theorem) for a wide class of random sequences and processes.
4. Refinement of the main LTs and derivation of asymptotic expansions. For instance, in the CLT, bounds of the rate of convergence $\mathbf{P}(\zeta_n < x) - \Phi(x) \to 0$ and

asymptotic expansions for this difference (in the powers of $n^{-1/2}$ in the case of i.i.d. summands) have been obtained under broad assumptions.

5. Studying large deviation probabilities for the sums S_n (theorems on rare events). If $x \to \infty$ together with n then the CLT can only assert that $\mathbf{P}(\zeta_n > x) \to 0$. Theorems on large deviation probabilities aim to find a function $P(x,n)$ such that

$$\frac{\mathbf{P}(\zeta_n > x)}{P(x,n)} \to 1 \quad \text{as} \quad n \to \infty, \quad x \to \infty.$$

The nature of the function $P(x,n)$ essentially depends on the rate of decay of $\mathbf{P}(X > x)$ as $x \to \infty$ and on the "deviation zone," i.e., on the asymptotic behavior of the ratio x/n as $n \to \infty$.

6. Considering observations $X_1, \ldots, X_n$ of a more complex nature – first of all, multivariate random vectors. If $X_j \in \mathbb{R}^d$ then the role of the limit law in the CLT will be played by a d-dimensional normal (Gaussian) distribution with the covariance matrix $\mathbf{E}(X - \mathbf{E}X)(X - \mathbf{E}X)^T$.

The variety of application areas of the LLN and CLT is enormous. Thus, *Mathematical Statistics* is based on these LTs. Let $X_n^* := (X_1, \ldots, X_n)$ be a sample from a distribution F and $F_n^*(u)$ the corresponding empirical distribution function. The fundamental Glivenko–Cantelli theorem (see ▶Glivenko-Cantelli Theorems) stating that $\sup_u |F_n^*(u) - F(u)| \xrightarrow{a.s.} 0$ as $n \to \infty$ is of the same nature as the LLN and basically means that the unknown distribution F can be estimated arbitrary well from the random sample X_n^* of a large enough size n.

The existence of consistent estimators for the unknown parameters $a = \mathbf{E}\xi$ and $\sigma^2 = \mathbf{E}(X - a)^2$ also follows from the LLN since, as $n \to \infty$,

$$a^* := \frac{1}{n}\sum_{j=1}^n X_j \xrightarrow{a.s.} a, \quad (\sigma^2)^* := \frac{1}{n}\sum_{j=1}^n (X_j - a^*)^2$$
$$= \frac{1}{n}\sum_{j=1}^n X_j^2 - (a^*)^2 \xrightarrow{a.s.} \sigma^2.$$

Under additional moment assumptions on the distribution F, one can also construct asymptotic confidence intervals for the parameters a and σ^2, as the distributions of the quantities $\sqrt{n}(a^* - a)$ and $\sqrt{n}\big((\sigma^2)^* - \sigma^2\big)$ converge, as $n \to \infty$, to the normal ones. The same can also be said about other parameters that are "smooth" enough functionals of the unknown distribution F.

The theorem on the ▶asymptotic normality and asymptotic efficiency of maximum likelihood estimators is another classical example of LTs' applications in mathematical statistics (see e.g., Borovkov 1998). Furthermore, in estimation theory and hypotheses testing, one also needs theorems on large deviation probabilities for the respective statistics, as it is statistical procedures with small error probabilities that are often required in applications.

It is worth noting that summation of random variables is by no means the only situation in which LTs appear in Probability Theory.

Generally speaking, the main objective of Probability Theory in applications is finding appropriate stochastic models for objects under study and then determining the distributions or parameters one is interested in. As a rule, the explicit form of these distributions and parameters is not known. LTs can be used to find suitable approximations to the characteristics in question.

At least two possible approaches to this problem should be noted here.

1. Suppose that the unknown distribution F_θ depends on a parameter θ such that, as θ approaches some "critical" value θ_0, the distributions F_θ become "degenerate" in one sense or another. Then, in a number of cases, one can find an approximation for F_θ which is valid for the values of θ that are close to θ_0. For instance, in actuarial studies, ▶queueing theory and some other applications one of the main problems is concerned with the distribution of $\overline{S} := \sup_{k \geq 1}(S_k - \theta k)$, under the assumption that $\mathbf{E}X = 0$. If $\theta > 0$ then $\overline{S}$ is a proper random variable. If, however, $\theta \to 0$ then $\overline{S} \xrightarrow{a.s.} \infty$. Here we deal with the so-called "transient phenomena." It turns out that if $\sigma^2 := \mathrm{Var}\,(X) < \infty$ then there exists the limit

$$\lim_{\theta \downarrow 0} \mathbf{P}(\theta \overline{S} > x) = e^{-2x/\sigma^2}, \quad x > 0.$$

This (Kingman–Prokhorov) LT enables one to find approximations for the distribution of $\overline{S}$ in situations where θ is small.

2. Sometimes one can estimate the "tails" of the unknown distributions, i.e., their asymptotic behavior at infinity. This is of importance in those applications where one needs to evaluate the probabilities of rare events. If the equation $\mathbf{E}e^{\mu(X-\theta)} = 1$ has a solution $\mu_0 > 0$ then, in the above example, one has

$$\mathbf{P}(\overline{S} > x) \sim ce^{-\mu_0 x}, \quad x \to \infty,$$

where c is a known constant. If the distribution F of X is *subexponential* (in this case, $\mathbf{E}e^{\mu(X-\theta)} = \infty$ for any $\mu > 0$) then

$$\mathbf{P}(\overline{S} > x) \sim \frac{1}{\theta}\int_x^\infty (1 - F(t))\,dt, \quad x \to \infty.$$

This LT enables one to find approximations for $\mathbf{P}(\overline{S} > x)$ for large x.

For both approaches, the obtained approximations can be refined.

An important part of Probability Theory is concerned with LTs for *random processes*. Their main objective is to find conditions under which random processes converge, in some sense, to some limit process. An extension of the CLT to that context is the so-called *Functional CLT* (a.k.a. the Donsker–Prokhorov invariance principle) which states that, as $n \to \infty$, the processes $\{\zeta_n(t) := (S_{\lfloor nt \rfloor} - ant)/\sigma\sqrt{n}\}_{t\in[0,1]}$ converge in distribution to the standard Wiener process $\{w(t)\}_{t\in[0,1]}$. The LTs (including large deviation theorems) for a broad class of functionals of the sequence (►random walk) $\{S_1, \ldots, S_n\}$ can also be classified as LTs for ►stochastic processes. The same can be said about Law of iterated logarithm which states that, for an arbitrary $\varepsilon > 0$, the random walk $\{S_k\}_{k=1}^{\infty}$ crosses the boundary $(1-\varepsilon)\sigma\sqrt{2k \ln \ln k}$ infinitely many times but crosses the boundary $(1+\varepsilon)\sigma\sqrt{2k \ln \ln k}$ finitely many times with probability 1. Similar results hold true for trajectories of Wiener processes $\{w(t)\}_{t\in[0,1]}$ and $\{w(t)\}_{t\in[1,\infty)}$.

In mathematical statistics a closely related to functional CLT result says that the so-called "empirical process" $\{\sqrt{n}(F_n^*(u) - F(u))\}_{u\in(-\infty,\infty)}$ converges in distribution to $\{w^0(F(u))\}_{u\in(-\infty,\infty)}$, where $w^0(t) := w(t) - tw(1)$ is the Brownian bridge process. This LT implies ►asymptotic normality of a great many estimators that can be represented as smooth functionals of the empirical distribution function $F_n^*(u)$.

There are many other areas in Probability Theory and its applications where various LTs appear and are extensively used. For instance, convergence theorems for ►martingales, asymptotics of extinction probability of a branching processes and conditional (under non-extinction condition) LTs on a number of particles etc.

About the Author

Professor Borovkov is Head of Probability and Statistics Department of the Sobolev Institute of Mathematics, Novosibirsk (Russian Academy of Sciences), since 1962. He is Head of Probability and Statistics Chair at the Novosibirsk University since 1966. He is Concelour of Russian Academy of Sciences and full member of Russian Academy of Sciences (Academician) (1990). He was awarded the State Prize of the USSR (1979), the Markov Prize of the Russian Academy of Sciences (2003) and Government Prize in Education (2003). Professor Borovkov is Editor-in-Chief of the journal "*Siberian Advances in Mathematics*" and Associated Editor of journals "*Theory of Probability and its Applications*," "*Siberian Mathematical Journal*," "*Mathematical Methods of Statistics*," "*Electronic Journal of Probability*."

Cross References

►Almost Sure Convergence of Random Variables
►Approximations to Distributions
►Asymptotic Normality
►Asymptotic Relative Efficiency in Estimation
►Asymptotic Relative Efficiency in Testing
►Central Limit Theorems
►Empirical Processes
►Ergodic Theorem
►Glivenko-Cantelli Theorems
►Large Deviations and Applications
►Laws of Large Numbers
►Martingale Central Limit Theorem
►Probability Theory: An Outline
►Random Matrix Theory
►Strong Approximations in Probability and Statistics
►Weak Convergence of Probability Measures

References and Further Reading

Billingsley P (1968) Convergence of probability measures. Wiley, New York
Borovkov AA (1998) Mathematical statistics. Gordon & Breach, Amsterdam
Gnedenko BV, Kolmogorov AN (1954) Limit distributions for sums of independent random variables. Addison-Wesley, Cambridge
Lévy P (1937) Théorie de l'Addition des Variables Aléatories. Gauthier-Villars, Paris
Loève M (1977–1978) Probability theory, vols I and II, 4th edn. Springer, New York
Petrov VV (1995) Limit theorems of probability theory: sequences of independent random variables. Clarendon/Oxford University Press, New York

Linear Mixed Models

Geert Molenberghs
Professor
Universiteit Hasselt & Katholieke Universiteit Leuven, Leuven, Belgium

In observational studies, repeated measurements may be taken at almost arbitrary time points, resulting in an extremely large number of time points at which only one

or only a few measurements have been taken. Many of the parametric covariance models described so far may then contain too many parameters to make them useful in practice, while other, more parsimonious, models may be based on assumptions which are too simplistic to be realistic. A general, and very flexible, class of parametric models for continuous longitudinal data is formulated as follows:

$$\boldsymbol{y_i}|\boldsymbol{b_i} \sim N(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i}, \Sigma_i), \quad (1)$$

$$\boldsymbol{b_i} \sim N(\mathbf{0}, D), \quad (2)$$

where X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ dimensional matrices of known covariates, $\boldsymbol{\beta}$ is a p-dimensional vector of regression parameters, called the fixed effects, D is a general $(q \times q)$ covariance matrix, and Σ_i is a $(n_i \times n_i)$ covariance matrix which depends on i only through its dimension n_i, i.e., the set of unknown parameters in Σ_i will not depend upon i. Finally, $\boldsymbol{b_i}$ is a vector of subject-specific or random effects.

The above model can be interpreted as a linear regression model (see ►Linear Regression Models) for the vector $\boldsymbol{y_i}$ of repeated measurements for each unit separately, where some of the regression parameters are specific (random effects, $\boldsymbol{b_i}$), while others are not (fixed effects, $\boldsymbol{\beta}$). The distributional assumptions in (2) with respect to the random effects can be motivated as follows. First, $\mathrm{E}(\boldsymbol{b_i}) = \mathbf{0}$ implies that the mean of $\boldsymbol{y_i}$ still equals $X_i\boldsymbol{\beta}$, such that the fixed effects in the random-effects model (1) can also be interpreted marginally. Not only do they reflect the effect of changing covariates within specific units, they also measure the marginal effect in the population of changing the same covariates. Second, the normality assumption immediately implies that, marginally, $\boldsymbol{y_i}$ also follows a normal distribution with mean vector $X_i\boldsymbol{\beta}$ and with covariance matrix $V_i = Z_iDZ_i' + \Sigma_i$.

Note that the random effects in (1) implicitly imply the marginal covariance matrix V_i of $\boldsymbol{y_i}$ to be of the very specific form $V_i = Z_iDZ_i' + \Sigma_i$. Let us consider two examples under the assumption of conditional independence, i.e., assuming $\Sigma_i = \sigma^2 I_{n_i}$. First, consider the case where the random effects are univariate and represent unit-specific intercepts. This corresponds to covariates Z_i which are n_i-dimensional vectors containing only ones.

The marginal model implied by expressions (1) and (2) is

$$\boldsymbol{y_i} \sim N(X_i\boldsymbol{\beta}, V_i), \quad V_i = Z_iDZ_i' + \Sigma_i$$

which can be viewed as a multivariate linear regression model, with a very particular parameterization of the covariance matrix V_i.

With respect to the estimation of unit-specific parameters $\boldsymbol{b_i}$, the posterior distribution of $\boldsymbol{b_i}$ given the observed data $\boldsymbol{y_i}$ can be shown to be (multivariate) normal with mean vector equal to $DZ_i'V_i^{-1}(\boldsymbol{\alpha})(\boldsymbol{y_i} - X_i\boldsymbol{\beta})$. Replacing $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by their maximum likelihood estimates, we obtain the so-called empirical Bayes estimates $\widehat{\boldsymbol{b_i}}$ for the $\boldsymbol{b_i}$. A key property of these EB estimates is shrinkage, which is best illustrated by considering the prediction $\widehat{\boldsymbol{y_i}} \equiv X_i\widehat{\boldsymbol{\beta}} + Z_i\widehat{\boldsymbol{b_i}}$ of the ith profile. It can easily be shown that

$$\widehat{\boldsymbol{y_i}} = \Sigma_i V_i^{-1} X_i\widehat{\boldsymbol{\beta}} + \left(I_{n_i} - \Sigma_i V_i^{-1}\right)\boldsymbol{y_i},$$

which can be interpreted as a weighted average of the population-averaged profile $X_i\widehat{\boldsymbol{\beta}}$ and the observed data $\boldsymbol{y_i}$, with weights $\Sigma_i V_i^{-1}$ and $I_{n_i} - \Sigma_i V_i^{-1}$, respectively. Note that the "numerator" of $\Sigma_i V_i^{-1}$ represents within-unit variability and the "denominator" is the overall covariance matrix V_i. Hence, much weight will be given to the overall average profile if the within-unit variability is large in comparison to the between-unit variability (modeled by the random effects), whereas much weight will be given to the observed data if the opposite is true. This phenomenon is referred to as shrinkage toward the average profile $X_i\widehat{\boldsymbol{\beta}}$. An immediate consequence of shrinkage is that the EB estimates show less variability than actually present in the random-effects distribution, i.e., for any linear combination $\boldsymbol{\lambda}$ of the random effects,

$$\mathrm{var}(\lambda'\widehat{\boldsymbol{b_i}}) \leq \mathrm{var}(\lambda'\boldsymbol{b_i}) = \lambda' D\lambda.$$

About the Author

Geert Molenberghs is Professor of Biostatistics at Universiteit Hasselt and Katholieke Universiteit Leuven in Belgium. He was Joint Editor of *Applied Statistics* (2001–2004), and Co-Editor of *Biometrics* (2007–2009). Currently, he is Co-Editor of Biostatistics (2010–2012). He was President of the International Biometric Society (2004–2005), received the Guy Medal in Bronze from the Royal Statistical Society and the Myrto Lefkopoulou award from the Harvard School of Public Health. Geert Molenberghs is founding director of the Center for Statistics. He is also the director of the Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat). Jointly with Geert Verbeke, Mike Kenward, Tomasz Burzykowski, Marc Buyse, and Marc Aerts, he authored books on longitudinal and incomplete data, and on surrogate marker evaluation. Geert Molenberghs received several Excellence in Continuing Education Awards of the American Statistical Association, for courses at Joint Statistical Meetings.

Cross References

►Best Linear Unbiased Estimation in Linear Models
►General Linear Models
►Testing Variance Components in Mixed Linear Models
►Trend Estimation

References and Further Reading

Brown H, Prescott R (1999) Applied mixed models in medicine. Wiley, New York
Crowder MJ, Hand DJ (1990) Analysis of repeated measures. Chapman & Hall, London
Davidian M, Giltinan DM (1995) Nonlinear models for repeated measurement data. Chapman & Hall, London
Davis CS (2002) Statistical methods for the analysis of repeated measurements. Springer, New York
Demidenko E (2004) Mixed models: theory and applications. Wiley, New York
Diggle PJ, Heagerty PJ, Liang KY, Zeger SL (2002) Analysis of longitudinal data, 2nd edn. Oxford University Press, Oxford
Fahrmeir L, Tutz G (2002) Multivariate statistical modelling based on generalized linear models, 2nd edn. Springer, New York
Fitzmaurice GM, Davidian M, Verbeke G, Molenberghs G (2009) Longitudinal data analysis. Handbook. Wiley, Hoboken
Goldstein H (1995) Multilevel statistical models. Edward Arnold, London
Hand DJ, Crowder MJ (1995) Practical longitudinal data analysis. Chapman & Hall, London
Hedeker D, Gibbons RD (2006) Longitudinal data analysis. Wiley, New York
Kshirsagar AM, Smith WB (1995) Growth curves. Marcel Dekker, New York
Leyland AH, Goldstein H (2001) Multilevel modelling of health statistics. Wiley, Chichester
Lindsey JK (1993) Models for repeated measurements. Oxford University Press, Oxford
Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2005) SAS for mixed models, 2nd edn. SAS Press, Cary
Longford NT (1993) Random coefficient models. Oxford University Press, Oxford
Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. Springer, New York
Pinheiro JC, Bates DM (2000) Mixed effects models in S and S-Plus. Springer, New York
Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York
Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer series in statistics. Springer, New York
Vonesh EF, Chinchilli VM (1997) Linear and non-linear models for the analysis of repeated measurements. Marcel Dekker, Basel
Weiss RE (2005) Modeling longitudinal data. Springer, New York
West BT, Welch KB, Gałecki AT (2007) Linear mixed models: a practical guide using statistical software. Chapman & Hall/CRC, Boca Raton
Wu H, Zhang J-T (2006) Nonparametric regression methods for longitudinal data analysis. Wiley, New York
Wu L (2010) Mixed effects models for complex data. Chapman & Hall/CRC Press, Boca Raton

Linear Regression Models

Raoul LePage
Professor
Michigan State University, East Lansing, MI, USA

► I did not want proof, because the theoretical exigencies of the problem would afford that. What I wanted was to be started in the right direction.

(F. Galton)

The *linear regression model* of statistics is any functional relationship $y = f(x, \beta, \varepsilon)$ involving a *dependent* real-valued variable y, *independent* variables x, *model parameters* β and *random variables* ε, such that a measure of central tendency for y in relation to x termed the *regression function* is linear in β. Possible regression functions include the conditional mean $E(y|x, \beta)$ (as when β is itself random as in Bayesian approaches), conditional medians, quantiles or other forms. Perhaps y is corn yield from a given plot of earth and variables x include levels of water, sunlight, fertilization, discrete variables identifying the genetic variety of seed, and combinations of these intended to model interactive effects they may have on y. The form of this linkage is specified by a function f known to the experimenter, one that depends upon parameters β whose values are not known, and also upon unseen random errors ε about which statistical assumptions are made. These models prove surprisingly flexible, as when localized linear regression models are knit together to estimate a regression function nonlinear in β. Draper and Smith (1981) is a plainly written elementary introduction to linear regression models, Rao (1965) is one of many established general references at the calculus level.

Aspects of Data, Model and Notation

Suppose a time varying sound signal is the superposition of sine waves of unknown amplitudes at two fixed known frequencies embedded in white noise background $y[t] = \beta_1 + \beta_2 \sin[.2t] + \beta_3 \sin[.35t] + \varepsilon$. We write $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ for $\beta = (\beta_1, \beta_2, \beta_3), x = (x_1, x_2, x_3), x_1 \equiv 1, x_2(t) = \sin[.2t], x_3(t) = \sin[.35t], t \geq 0$. A natural choice of regression function is $m(x, \beta) = E(y|x, \beta) = \beta_1 + \beta_2 x_2 + \beta_3 x_3$ provided $E\varepsilon \equiv 0$. In the *classical linear regression model* one assumes for different instances "i" of observation that random errors satisfy $E\varepsilon_i \equiv 0$, $E\varepsilon_i \varepsilon_k \equiv \sigma^2 > 0$, $i = k \leq n$, $E\varepsilon_i \varepsilon_k \equiv 0$ otherwise. Errors in linear regression models typically depend upon instances i at which we select observations and may in some formulations depend also on the values of x associated with instance i (perhaps the

errors are correlated and that correlation depends upon the x values). What we observe are y_i and associated values of the independent variables x. That is, we observe $(y_i, 1, \sin[.2t_i], \sin[.35t_i]), i \le n$. The linear model on data may be expressed $y = x\beta + \varepsilon$ with y = column vector $\{y_i,\ i \le n\}$, likewise for ε, and matrix x (the *design matrix*) whose $3n$ entries are $x_{ik} = x_k[t_i]$.

Terminology

Independent variables, as employed in this context, is misleading. It derives from language used in connection with mathematical equations and does not refer to statistically independent random variables. Independent variables may be of any dimension, in some applications functions or surfaces. If y is not scalar-valued the model is instead a *multivariate linear regression*. In some versions either x, β or both may also be random and subject to statistical models. Do not confuse multivariate linear regression with *multiple linear regression* which refers to a model having more than one non-constant independent variable.

General Remarks on Fitting Linear Regression Models to Data

Early work (the classical linear model) emphasized independent identically distributed (i.i.d.) additive *normal* errors in linear regression where ▶*least squares* has particular advantages (connections with ▶multivariate normal distributions are discussed below). In that setup least squares would arguably be a principle method of fitting linear regression models to data, perhaps with modifications such as Lasso or other constrained optimizations that achieve reduced sampling variations of coefficient estimators while introducing bias (Efron et al. 2004). Absent a breakthrough enlarging the applicability of the classical linear model other methods gain traction such as Bayesian methods (▶Markov Chain Monte Carlo having enabled their calculation); Non-parametric methods (good performance relative to more relaxed assumptions about errors); Iteratively Reweighted least squares (having under some conditions behavior like maximum likelihood estimators without knowing the precise form of the likelihood). The Dantzig selector is good news for dealing with far fewer observations than independent variables when a relatively small fraction of them matter (Candès and Tao 2007).

Background

C.F. Gauss may have used least squares as early as 1795. In 1801 he was able to predict the apparent position at which asteroid Ceres would re-appear from behind the sun after it had been lost to view following discovery only 40 days before. Gauss' prediction was well removed from all others and he soon followed up with numerous other high-caliber successes, each achieved by fitting relatively simple models motivated by Keppler's Laws, work at which he was very adept and quick. These were fits to imperfect, sometimes limited, yet fairly precise data. Legendre (1805) published a substantive account of least squares following which the method became widely adopted in astronomy and other fields. See Stigler (1986).

By contrast Galton (1877), working with what might today be described as "low correlation" data, discovered deep truths not already known by fitting a straight line. No theoretical model previously available had prepared Galton for these discoveries which were made in a study of his own data w = standard score of weight of parental sweet pea seed(s), y = standard score of seed weights(s) of their immediate issue. Each sweet pea seed has but one parent and the distributions of x and y the same. Working at a time when correlation and its role in regression were yet unknown, Galton found to his astonishment a nearly perfect straight line tracking points (parental seed weight w, median filial seed weight $m(w)$). Since for this data $s_y \sim s_w$ this was the least squares line (also the regression line since the data was bivariate normal) and its slope was $rs_y/s_w = r$ (the correlation). Medians $m(w)$ being essentially equal to means of y for each w greatly facilitated calculations owing to his choice to select equal numbers of parent seeds at weights $0, \pm1, \pm2, \pm3$ standard deviations from the mean of w. Galton gave the name co-relation (later correlation) to the slope $\sim$ 0.33 of this line and for a brief time thought it might be a universal constant. Although the correlation was small, this slope nonetheless gave measure to the previously vague principle of reversion (later regression, as when larger parental examples beget offspring typically not quite so large). Galton deduced the general principle that if $0 < r < 1$ then for a value $w > Ew$ the relation $Ew < m(w) < w$ follows. Having sensibly selected equal numbers of parental seeds at intervals may have helped him observe that points (w, y) departed on each vertical from the regression line by statistically independent $N(0, \theta^2)$ random residuals whose variance $\theta^2 > 0$ was the same for all w. Of course this likewise amazed him and by 1886 he had identified all these properties as a consequence of bivariate normal observations (w, y), (Galton 1886).

Echoes of those long ago events reverberate today in our many statistical models "driven," as we now proclaim, by random errors subject to ever broadening statistical modeling. In the beginning it was very close to truth.

Aspects of Linear Regression Models

Data of the real world seldom conform exactly to any deterministic mathematical model $y = f(x, \beta)$ and through the device of incorporating random errors we have now an established paradigm for fitting models to data (x, y) by statistically estimating model parameters. In consequence we obtain methods for such purposes as predicting what will be the average response y to particular given inputs x; providing *margins of error* (and *prediction error*) for various quantities being estimated or predicted, tests of hypotheses and the like. It is important to note in all this that more than one statistical model may apply to a given problem, the function f and the other model components differing among them. Two statistical models may disagree substantially in structure and yet neither, either or both may produce useful results. In this respect statistical modeling is more a matter of how much we gain from using a statistical model and whether we trust and can agree upon the assumptions placed on the model, at least as a practical expedient. In some cases the regression function conforms precisely to underlying mathematical relationships but that does not reflect the majority of statistical practice. It may be that a given statistical model, although far from being an underlying truth, confers advantage by capturing some portion of the variation of y vis-a-vis x. The method *principle components,* which seeks to find relatively small numbers of linear combinations of the independent variables that together account for most of the variation of y, illustrates this point well. In one application electromagnetic theory was used to generate by computer an elaborate database of theoretical responses of an induced electromagnetic field near a metal surface to various combinations of flaws in the metal. The role of principle components and linear modeling was to establish a simple model reflecting those findings so that a portable device could be devised to make detections in real time based on the model.

If there is any weakness to the statistical approach it lies in the fact that margins of error, statistical tests and the like can be seriously incorrect even if the predictions afforded by a model have apparent value. Refer to Hinkelmann and Kempthorne (2008), Berk (2004), Freedman (2005), Freedman (1991).

Classical Linear Regression Model and Least Squares

The classical linear regression model may be expressed $y = x\beta + \varepsilon$, an abbreviated matrix formulation of the system of equations in which random errors ε are assumed to satisfy $E\varepsilon_I \equiv 0, E\varepsilon_i^2 \equiv \sigma^2 > 0, i \le n$:

$$y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i,\ i \le n. \tag{1}$$

The interpretation is that response y_i is observed for the ith sample in conjunction with numerical values $(x_{i1}, \ldots, x_{ip})$ of the independent variables. If these errors $\{\varepsilon_I\}$ are jointly normally distributed (and therefore statistically independent having been assumed to be uncorrelated) and if the matrix $x^{\mathrm{tr}}x$ is non-singular then the maximum likelihood (ML) estimates of the model coefficients $\{\beta_k, k \le p\}$ are produced by ordinary least squares (LS) as follows:

$$\beta_{ML} = \beta_{LS} = (x^{\mathrm{tr}}x)^{-1}x^{\mathrm{tr}}y = \beta + M\varepsilon \tag{2}$$

for $M = (x^{\mathrm{tr}}x)^{-1}x^{\mathrm{tr}}$ with x^{tr} denoting matrix transpose of x and $(x^{\mathrm{tr}}x)^{-1}$ the matrix inverse. These coefficient estimates β_{LS} are linear functions in y and satisfy the Gauss–Markov properties (3)(4):

$$E(\beta_{LS})_k = \beta_k, k \le p. \tag{3}$$

and, among all unbiased estimators β_k^* (of β_k) that are linear in y,

$$E((\beta_{LS})_k - \beta_k)^2 \le E\left(\beta_k^* - \beta_k\right)^2, \text{ for every } k \le p. \tag{4}$$

Least squares estimator (2) is frequently employed without the assumption of normality owing to the fact that properties (3)(4) must hold in that case as well. Many statistical distributions F, t, *chi-square* have important roles in connection with model (1) either as exact distributions for quantities of interest (normality assumed) or more generally as limit distributions when data are suitably enriched.

Algebraic Properties of Least Squares

Setting all randomness assumptions aside we may examine the algebraic properties of least squares. If $y = x\beta + \varepsilon$ then $\beta_{LS} = My = M(x\beta + \varepsilon) = \beta + M\varepsilon$ as in (2). That is, the least squares estimate of model coefficients acts on $x\beta + \varepsilon$ returning β plus the result of applying least squares to ε. This has nothing to do with the model being correct or ε being error but is purely algebraic. If ε itself has the form $xb + e$ then $M\varepsilon = b + Me$. Another useful observation is that if x has first column identically one, as would typically be the case for a model with constant term, then each row $M_k, k > 1$, of M satisfies $1.M_k = 0$ (i.e., M_k is a *contrast*) so $(\beta_{LS})_k = \beta_k + M_k.\varepsilon$ and $M_k.(\varepsilon + c) = M_k.\varepsilon$ so ε may as well be assumed to be centered for $k > 1$. There are many of these interesting algebraic properties such as $s^2(y - x\beta_{LS}) = (1 - R^2)s^2(y)$ where $s(.)$ denotes the sample standard deviation and R is the *multiple correlation* defined as the correlation between y and the *fitted values* $x\beta_{LS}$. Yet another algebraic identity, this one involving

an interplay of permutations with projections, is exploited to help establish for *exchangeable* errors ε, and contrasts v, a permutation bootstrap of least squares residuals that consistently estimates the *conditional sampling distribution* of $v.(\beta_{LS} - \beta)$ conditional on the ▶order statistics of ε. (See LePage and Podgorski 1996). Freedman and Lane in 1983 advocated tests based on permutation bootstrap of residuals as a descriptive method.

Generalized Least Squares

If errors ε are $N(0, \Sigma)$ distributed for a covariance matrix Σ known up to a constant multiple then the maximum likelihood estimates of coefficients β are produced by a *generalized* least squares solution retaining properties (3)(4) (any positive multiple of Σ will produce the same result) given by:

$$\beta_{ML} = (x^{\mathrm{tr}}\Sigma^{-1}x)^{-1}x^{\mathrm{tr}}\Sigma^{-1}y = \beta + (x^{\mathrm{tr}}\Sigma^{-1}x)^{-1}x^{\mathrm{tr}}\Sigma^{-1}\varepsilon. \quad (5)$$

Generalized least squares solution (5) retains properties (3)(4) even if normality is not assumed. It must not be confused with ▶*generalized linear models* which refers to models equating moments of y to nonlinear functions of $x\beta$.

A very large body of work has been devoted to linear regression models and the closely related subject areas of experimental design, ▶analysis of variance, principle component analysis and their consequent distribution theory.

Reproducing Kernel Generalized Linear Regression Model

Parzen (1961, Sect. 6) developed the reproducing kernel framework extending generalized least squares to spaces of arbitrary finite or infinite dimension when the random error function $\varepsilon = \{\varepsilon(t), t \in T\}$ has zero means $E\varepsilon(t) \equiv 0$ and a covariance function $K(s,t) = E\varepsilon(s)\varepsilon(t)$ that is assumed known up to some positive constant multiple. In this formulation:

$$y(t) = m(t) + \varepsilon(t),\ t \in T,$$
$$m(t) = Ey(t) = \Sigma_i \beta_i w_i(t),\ t \in T,$$

where $w_i(.)$ are *known* linearly independent functions in *the reproducing kernel Hilbert* (*RKHS*) *space* $H(K)$ of the kernel K. For reasons having to do with singularity of Gaussian measures it is assumed that the series defining m is convergent in H(K). Parzen extends to that context and solves the problem of best linear unbiased estimation of the model coefficients β and more generally of *estimable* linear functions of them, developing confidence regions, prediction intervals, exact or approximate distributions, tests and other matters of interest, and establishing the Gauss–Markov properties (3)(4). The *RKHS* setup has been examined from an on-line learning perspective (Vovk 2008).

Joint Normal Distributed (*x*, *y*) as Motivation for the Linear Regression Model and Least Squares

For the moment, think of (x, y) as following a multivariate normal distribution, as might be the case under process control or in natural systems. The (regular) conditional expectation of y relative to x is then, for some β:

$$E(y|x) = Ey + E((y - Ey)|x) = Ey + x.\beta \text{ for every } x$$

and the discrepancies $y - E(y|x)$ are for each x distributed $N(0, \sigma^2)$ for a fixed σ^2, independent for different x.

Comparing Two Basic Linear Regression Models

Freedman (1981) compares the analysis of two superficially similar but differing models:

Errors model: Model (1) above.

Sampling model: Data $(x_{i1}, \ldots, x_{ip}, y_i), i \le n$ represent a *random sample* from a finite population (e.g., an actual physical population).

In the sampling model, $\{\varepsilon_i\}$ are simply the *residual discrepancies* y-$x\beta_{LS}$ of a least squares fit of linear model $x\beta$ *to the population*. Galton's seed study is an example of this if we regard his data (w, y) as resulting from equal probability without-replacement random sampling of a population of pairs (w, y) with w restricted to be at $0, \pm1, \pm2, \pm3$ standard deviations from the mean. Both with and without-replacement equal-probability sampling are considered by Freedman. Unlike the errors model there is no assumption in the sampling model that the population linear regression model is in any way correct, although least squares may not be recommended if the population residuals depart significantly from i.i.d. normal. Our only purpose is to estimate the coefficients of the population *LS* fit of the model using *LS* fit of the model to our sample, give estimates of the likely proximity of our sample least squares fit to the population fit and estimate the quality of the population fit (e.g., multiple correlation).

Freedman (1981) established the applicability of Efron's Bootstrap to each of the two models above but under different assumptions. His results for the sampling model are a textbook application of Bootstrap since a description of the sampling theory of least squares estimates for the sampling model has complexities largely, as had been said, out of the way when the Bootstrap approach is used. It would be an interesting exercise to examine data, such as Galton's

seed data, analyzing it by the two different models, obtaining confidence intervals for the estimated coefficients of a straight line fit in each case to see how closely they agree.

Balancing Good Fit Against Reproducibility

A balance in the linear regression model is necessary. Including too many independent variables in order to assure a close fit of the model to the data is called overfitting. Models over-fit in this manner tend not to work with fresh data, for example to predict y from a fresh choice of the values of the independent variables. Galton's regression line, although it did not afford very accurate predictions of y from w, owing to the modest correlation ~ 0.33, was arguably best for his bi-variate normal data (w, y). Tossing in another independent variable such as w^2 for a parabolic fit would have over-fit the data, possibly spoiling discovery of the principle of regression to mediocrity.

A model might well be used even when it is understood that incorporating additional independent variables will yield a better fit to data and a model closer to truth. How could this be? If the more parsimonious choice of x accounts for enough of the variation of y in relation to the variables of interest to be useful and if its fewer coefficients β are estimated more reliably perhaps. Intentional use of a simpler model might do a reasonably good job of giving us the estimates we need but at the same time violate assumptions about the errors thereby invalidating confidence intervals and tests. Gauss needed to come close to identifying the location at which Ceres would appear. Going for too much accuracy by complicating the model risked overfitting owing to the limited number of observations available.

One possible resolution to this tradeoff between reliable estimation of a few model coefficients, versus the risk that by doing so too much model related material is left in the error term, is to include all of several hierarchically ordered layers of independent variables, more than may be needed, then remove those that the data suggests are not required to explain the greater share of the variation of y (Raudenbush and Bryk 2001). New results on data compression (Candès and Tao 2007) may offer fresh ideas for reliably removing, in some cases, less relevant independent variables without first arranging them in a hierarchy.

Regression to Mediocrity Versus Reversion to Mediocrity or Beyond

Regression (when applicable) is often used to prove that a high performing group on some scoring, i.e., $X > c > EX$, will not average so highly on another scoring Y, as they do on X, i.e., $E(Y|X > c) < E(X|X > c)$. Termed reversion to mediocrity or beyond by Samuels (1991) this property is easily come by when X, Y have the same distribution. The following result and brief proof are Samuels' except for clarifications made here (***italics***). These comments are addressed only to the *formal mathematical proof* of the paper.

Proposition Let random variables X, Y be *identically distributed* with finite mean EX and fix any $c > \max(0, EX)$. If $P(X > c \text{ and } Y > c) < P(Y > c)$ then there is reversion to mediocrity or beyond for that c.

Proof For any given $c > \max(0, EX)$ define the difference J of indicator random variables $J = (X > c) - (Y > c)$. J is zero unless one indicator is 1 and the other 0. YJ is less *or equal* $cJ = c$ on $J = 1$ (i.e., on $X > c, Y \le c$) and YJ is *strictly less* than $cJ = -c$ on $J = -1$ (i.e., on $X \le c, Y > c$). *Since the event* $J = -1$ *has positive probability by assumption*, the previous implies $E\,YJ < c\,EJ$ and so

$$EY(X > c) = E(Y(Y > c) + YJ) = EX(X > c) + E\,YJ < EX(X > c) + cEJ = EX(X > c),$$

yielding $E(Y|X > c) < E(X|X > c)$. □

Cross References

►Adaptive Linear Regression
►Analysis of Areal and Spatial Interaction Data
►Business Forecasting Methods
►Gauss-Markov Theorem
►General Linear Models
►Heteroscedasticity
►Least Squares
►Optimum Experimental Design
►Partial Least Squares Regression Versus Other Methods
►Regression Diagnostics
►Regression Models with Increasing Numbers of Unknown Parameters
►Ridge and Surrogate Ridge Regressions
►Simple Linear Regression
►Two-Stage Least Squares

References and Further Reading

Berk RA (2004) Regression analysis: a constructive critique. Sage, Newbury park

Candès E, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat 35(6):2313–2351

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Stat 32(2):407–499

Freedman D (1981) Bootstrapping regression models. Ann Stat 9:1218–1228

Freedman D (1991) Statistical models and shoe leather. Sociol Methodol 21:291–313
Freedman D (2005) Statistical models: theory and practice. Cambridge University Press, New York
Freedman D, Lane D (1983) A non-stochastic interpretation of reported significance levels. J Bus Econ Stat 1:292–298
Galton F (1877) Typical laws of heredity. Nature 15:493–495, 512–514, 532–533
Galton F (1886) Regression towards mediocrity in hereditary stature. J Anthropolocal Inst Great Britain and Ireland 15:246–263
Hinkelmann K, Kempthorne O (2008) Design and analysis of experiments, vol 1, 2, 2nd edn. Wiley, Hoboken
Kendall M, Stuart A (1979) The advanced theory of statistics, volume 2: Inference and relationship. Charles Griffin, London
LePage R, Podgorski K (1996) Resampling permutations in regression without second moments. J Multivariate Anal 57(11):119–141, Elsevier
Parzen E (1961) An approach to time series analysis. Ann Math Stat 32(4):951–989
Rao CR (1965) Linear statistical inference and its applications. Wiley, New York
Raudenbush S, Bryk A (2001) Hierarchical linear models applications and data analysis methods, 2nd edn. Sage, Thousand Oaks
Samuels ML (1991) Statistical reversion toward the mean: more universal than regression toward the mean. Am Stat 45:344–346
Stigler S (1986) The history of statistics: the measurement of uncertainty before 1900. Belknap Press of Harvard University Press, Cambridge
Vovk V (2006) On-line regression competitive with reproducing kernel Hilbert spaces. Lecture notes in computer science, vol 3959. Springer, Berlin, pp 452–463

Local Asymptotic Mixed Normal Family

Ishwar V. Basawa
Professor
University of Georgia, Athens, GA, USA

Suppose $x(n) = (x_1, \ldots, x_n)$ is a sample from a stochastic process $x = \{x_1, x_2, \ldots\}$. Let $p_n(x(n);\theta)$ denote the joint density function of $x(n)$, where $\theta \epsilon \Omega \subset \mathcal{R}^k$ is a parameter. Define the log-likelihood ratio $\Lambda_n = \left[\frac{p_n(x(n);\theta_n)}{p_n(x(n);\theta)}\right]$, where $\theta_n = \theta + n^{-\frac{1}{2}}h$, and h is a $(k \times 1)$ vector. The joint density $p_n(x(n);\theta)$ belongs to a local asymptotic normal (LAN) family if Λ_n satisfies

$$\Lambda_n = n^{-\frac{1}{2}} h^t S_n(\theta) - n^{-1}\left(\frac{1}{2} h^t J_n(\theta) h\right) + o_p(1) \qquad (1)$$

where $S_n(\theta) = \frac{dlnp_n(x(n);\theta)}{d\theta}$, $J_n(\theta) = -\frac{d^2 lnp_n(x(n);\theta)}{d\theta d\theta^t}$, and

$$(i) n^{-\frac{1}{2}} S_n(\theta) \xrightarrow{d} N_k(0, F(\theta)), (ii) n^{-1} J_n(\theta) \xrightarrow{p} F(\theta), \qquad (2)$$

$F(\theta)$ being the limiting Fisher information matrix. Here, $F(\theta)$ ia assumed to be non-random. See LaCam and Yang (1990) for a review of the LAN family.

For the LAN family defined by (1) and (2), it is well known that, under some regularity conditions, the maximum likelihood (ML) estimator $\widehat{\theta}_n$ is consistent asymptotically normal and efficient estimator of θ with

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N_k(0, F^{-1}(\theta)). \qquad (3)$$

A large class of models involving the classical *i.i.d.* (independent and identically distributed) observations are covered by the LAN framework. Many time series models and ►Markov processes also are included in the LAN family.

If the limiting Fisher information matrix $F(\theta)$ is non-degenerate random, we obtain a generalization of the LAN family for which the limit distribution of the ML estimator in (3) will be a mixture of normals (and hence non-normal). If Λ_n satisfies (1) and (2) with $F(\theta)$ random, the density $p_n(x(n);\theta)$ belongs to a local asymptotic mixed normal (LAMN) family. See Basawa and Scott (1983) for a discussion of the LAMN family and related asymptotic inference questions for this family.

For the LAMN family, one can replace the norm $\sqrt{n}$ by a random norm $J_n^{\frac{1}{2}}(\theta)$ to get the limiting normal distribution, viz.,

$$J_n^{\frac{1}{2}}(\theta)(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, I), \qquad (4)$$

where I is the identity matrix.

Two examples belonging to the LAMN family are given below:

Example 1 Variance mixture of normals

Suppose, conditionally on $\mathcal{V} = v$, $(x_1, x_2, \ldots, x_n)$ are *i.i.d.* $N(\theta, v^{-1})$ random variables, and $\mathcal{V}$ is an exponential random variable with mean 1. The marginal joint density of $x(n)$ is then given by $p_n(x(n);\theta) \propto \left[1 + \frac{1}{2}\sum_1^n (x_i - \theta)^2\right]^{-\left(\frac{n}{2}+1\right)}$. It can be verified that $F(\theta)$ is an exponential random variable with mean 1. The ML estimator $\widehat{\theta}_n = \bar{x}$ and $\sqrt{n}(\bar{x} - \theta) \xrightarrow{d} t(2)$. It is interesting to note that the variance of the limit distribution of $\bar{x}$ is ∞!

Example 2 Autoregressive process

Consider a first-order autoregressive process $\{x_t\}$ defined by $x_t = \theta x_{t-1} + e_t, t = 1, 2, \ldots,$ with $x_0 = 0$, where $\{e_t\}$ are assumed to be *i.i.d.* $N(0,1)$ random variables.

We then have $p_n(x(n);\theta) \propto exp\left[-\frac{1}{2}\sum_1^n (x_t - \theta x_{t-1})^2\right]$. For the stationary case, $|\theta| < 1$, this model belongs to the LAN family. However, for $|\theta| > 1$, the model belongs to the LAMN family. For any θ, the ML estimator $\widehat{\theta}_n = \left(\sum_1^n x_i x_{i-1}\right)\left(\sum_1^n x_{i-1}^2\right)^{-1}$. One can verify that $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0,(1-\theta^2)^{-1})$, for $|\theta| < 1$, and $(\theta^2-1)^{-1}\theta^n(\widehat{\theta}_n - \theta) \xrightarrow{d}$ Cauchy, for $|\theta| > 1$.

About the Author

Dr. Ishwar Basawa is a Professor of Statistics at the University of Georgia, USA. He has served as interim head of the department (2000–2003), Executive Editor of the *Journal of Statistical Planning and Inference* (1995–1997), on the editorial board of *Communications in Statistics*, and currently the online *Journal of Probability and Statistics*. Professor Basawa is a Fellow of the Institute of Mathematical Statistics and he was an Elected member of the International Statistical Institute. He has co-authored two books and co-edited eight Proceedings/Monographs/Special Issues of journals. He has authored more than 125 publications. His areas of research include inference for stochastic processes, time series, and asymptotic statistics.

Cross References

►Asymptotic Normality
►Optimal Statistical Inference in Financial Engineering
►Sampling Problems for Stochastic Processes

References and Further Reading

Basawa IV, Scott DJ (1983) Asymptotic optimal inference for non-ergodic models. Springer, New York
LeCam L, Yang GL (1990) Asymptotics in statistics. Springer, New York

Location-Scale Distributions

Horst Rinne
Professor Emeritus for Statistics and Econometrics
Justus–Liebig–University Giessen, Giessen, Germany

A random variable X with realization x belongs to the location-scale family when its cumulative distribution is a function only of $(x-a)/b$:

$$F_X(x\,|\,a,b) = \Pr(X \le x|a,b) = F\left(\frac{x-a}{b}\right);\ a \in \mathbb{R},\ b > 0;$$

where $F(\cdot)$ is a distribution having no other parameters. Different $F(\cdot)$'s correspond to different members of the family. (a,b) is called the location–scale parameter, a being the location parameter and b being the scale parameter. For fixed $b = 1$ we have a subfamily which is a location family with parameter a, and for fixed $a = 0$ we have a scale family with parameter b. The variable

$$Y = \frac{X-a}{b}$$

is called the reduced or standardized variable. It has $a = 0$ and $b = 1$. If the distribution of X is absolutely continuous with density function

$$f_X(x\,|\,a,b) = \frac{\mathrm{d}F_X(x\,|\,a,b)}{\mathrm{d}\,x}$$

then (a,b) is a location scale-parameter for the distribution of X if (and only if)

$$f_X(x\,|\,a,b) = \frac{1}{b} f\left(\frac{x-a}{b}\right)$$

for some density $f(\cdot)$, called the reduced density. All distributions in a given family have the same shape, i.e., the same skewness and the same kurtosis. When Y has mean μ_Y and standard deviation σ_Y then, the mean of X is $\mathrm{E}(X) = a + b\,\mu_Y$ and the standard deviation of X is $\sqrt{\mathrm{Var}(X)} = b\,\sigma_Y$.

The location parameter a, $a \in \mathbb{R}$ is responsible for the distribution's position on the abscissa. An enlargement (reduction) of a causes a movement of the distribution to the right (left). The location parameter is either a measure of central tendency e.g., the mean, median and mode or it is an upper or lower threshold parameter. The scale parameter b, $b > 0$, is responsible for the dispersion or variation of the variate X. Increasing (decreasing) b results in an enlargement (reduction) of the spread and a corresponding reduction (enlargement) of the density. b may be the standard deviation, the full or half length of the support, or the length of a central $(1-\alpha)$–interval.

The location-scale family has a great number of members:

- Arc-sine distribution
- Special cases of the beta distribution like the rectangular, the asymmetric triangular, the U–shaped or the power–function distributions
- Cauchy and half–Cauchy distributions
- Special cases of the χ–distribution like the half–normal, the Rayleigh and the Maxwell–Boltzmann distributions
- Ordinary and raised cosine distributions
- Exponential and reflected exponential distributions
- Extreme value distribution of the maximum and the minimum, each of type I
- Hyperbolic secant distribution

- Laplace distribution
- Logistic and half–logistic distributions
- Normal and half–normal distributions
- Parabolic distributions
- Rectangular or uniform distribution
- Semi–elliptical distribution
- Symmetric triangular distribution
- Teissier distribution with reduced density $f(y) = \left[\exp(y) - 1\right] \exp\left[1 + y - \exp(y)\right], y \geq 0$
- V–shaped distribution

For each of the above mentioned distributions we can design a special probability paper. Conventionally, the abscissa is for the realization of the variate and the ordinate, called the probability axis, displays the values of the cumulative distribution function, but its underlying scaling is according to the percentile function. The ordinate value belonging to a given sample data on the abscissa is called plotting position; for its choice see Barnett (1975, 1976), Blom (1958), Kimball (1960). When the sample comes from the probability paper's distribution the plotted data will randomly scatter around a straight line, thus, we have a graphical goodness-fit-test. When we fit the straight line by eye we may read off estimates for a and b as the abscissa or difference on the abscissa for certain percentiles. A more objective method is to fit a least-squares line to the data, and the estimates of a and b will be the the parameters of this line.

The latter approach takes the order statistics $X_{i:n}$, $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ as regressand and the mean of the reduced order statistics $\alpha_{i:n} := \mathrm{E}\left(Y_{i:n}\right)$ as regressor, which under these circumstances acts as plotting position. The regression model reads:

$$X_{i:n} = a + b\,\alpha_{i:n} + \varepsilon_i,$$

where ε_i is a random variable expressing the difference between $X_{i:n}$ and its mean $\mathrm{E}\left(X_{i:n}\right) = a + b\,\alpha_{i:n}$. As the order statistics $X_{i:n}$ and – as a consequence – the disturbance terms ε_i are neither homoscedastic nor uncorrelated we have to use – according to Lloyd (1952) – the general-least-squares method to find best linear unbiased estimators of a and b. Introducing the following vectors and matrices:

$$\boldsymbol{x} := \begin{pmatrix} X_{1:n} \\ X_{2:n} \\ \vdots \\ X_{n:n} \end{pmatrix}, \mathbf{1} := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \boldsymbol{\alpha} := \begin{pmatrix} \alpha_{1:n} \\ \alpha_{2:n} \\ \vdots \\ \alpha_{n:n} \end{pmatrix}, \boldsymbol{\varepsilon} := \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \boldsymbol{\theta} := \begin{pmatrix} a \\ b \end{pmatrix},$$

$$\boldsymbol{A} := \begin{pmatrix} \mathbf{1} & \boldsymbol{\alpha} \end{pmatrix}$$

the regression model now reads

$$\boldsymbol{x} = \boldsymbol{A}\,\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

with variance–covariance matrix

$$\mathrm{Var}(\boldsymbol{x}) = b^2\,\boldsymbol{B}.$$

The GLS estimator of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \left(\boldsymbol{A}'\,\boldsymbol{\Omega}\,\boldsymbol{A}\right)^{-1} \boldsymbol{A}'\,\boldsymbol{\Omega}\,\boldsymbol{x}$$

and its variance–covariance matrix reads

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) = b^2 \left(\boldsymbol{A}'\,\boldsymbol{\Omega}\,\boldsymbol{A}\right)^{-1}.$$

The vector $\boldsymbol{\alpha}$ and the matrix $\boldsymbol{B}$ are not always easy to find. For only a few location–scale distributions like the exponential, the reflected exponential, the extreme value, the logistic and the rectangular distributions we have closed-form expressions, in all other cases we have to evaluate the integrals defining $\mathrm{E}\left(Y_{i:n}\right)$ and $\mathrm{E}\left(Y_{i:n}\,Y_{j:n}\right)$. For more details on linear estimation and probability plotting for location-scale distributions and for distributions which can be transformed to location–scale type see Rinne (2010). Maximum likelihood estimation for location–scale distributions is treated by Mi (2006).

About the Author

Dr. Horst Rinne is Professor Emeritus (since 2005) of statistics and econometrics. In 1971 he was awarded the Venia legendi in statistics and econometrics by the Faculty of economics of Berlin Technical University. From 1965 to 1972 he worked as a part-time lecturer of statistics at Berlin Polytechnic and at the Technical University of Hannover. In 1969 he joined Volkswagen AG to do operations research. In 1972 he was appointed full professor of statistics and econometrics at the Justus-Liebig University in Giessen, where later on he was Dean of the faculty of economics and management science for three years. He got further appointments to the universities of Tübingen and Cologne and to Berlin Polytechnic. He was a member of the board of the German Statistical Society and in charge of editing its journal *Allgemeines Statistisches Archiv*, now *AStA – Advances in Statistical Analysis*, from 1981 to 1997. He is Co-founder of the Econometric Board of the German Society of Economics. Since 1985 he is Elected member of the International Statistical Institute. He is Associate editor for *Quality Technology and Quantitative Management*. His scientific interests are wide-spread, ranging from financial mathematics, business and economics statistics, technical statistics to econometrics. He has written numerous papers in these fields and is author of several textbooks and monographs in statistics, econometrics, time series analysis, multivariate statistics and statistical quality control including process capability. He is the author of the text *The Weibull Distribution: A Handbook* (Chapman and Hall/CRC, 2008).

Cross References

►Statistical Distributions: An Overview

References and Further Reading

Barnett V (1975) Probability plotting methods and order statistics. Appl Stat 24:95–108

Barnett V (1976) Convenient probability plotting positions for the normal distribution. Appl Stat 25:47–50

Blom G (1958) Statistical estimates and transformed beta variables. Almqvist and Wiksell, Stockholm

Kimball BF (1960) On the choice of plotting positions on probability paper. J Am Stat Assoc 55:546–560

Lloyd EH (1952) Least-squares estimation of location and scale parameters using order statistics. Biometrika 39:88–95

Mi J (2006) MLE of parameters of location–scale distributions for complete and partially grouped data. J Stat Planning Inference 136:3565–3582

Rinne H (2010) Location-scale distributions – linear estimation and probability plotting; http:geb.uni-giessen/geb/volltexte/2010/7607/

Logistic Normal Distribution

John Hinde
Professor of Statistics
National University of Ireland, Galway, Ireland

The logistic-normal distribution arises by assuming that the *logit* (or logistic transformation) of a proportion has a normal distribution, with an obvious extension to a vector of proportions through taking a logistic transformation of a multivariate normal distribution, see Aitchison and Shen (1980). In the univariate case, this provides a family of distributions on $(0,1)$ that is distinct from the ►*beta distribution*, while the multivariate version is an alternative to the *Dirichlet distribution*. Note that in the multivariate case there is no unique way to define the set of logits for the multinomial proportions (just as in multinomial logit models, see Agresti 2002) and different formulations may be appropriate in particular applications (Aitchison 1982). The univariate distribution has been used, often implicitly, in random effects models for binary data and the multivariate version was pioneered by Aitchison for statistical diagnosis/discrimination (Aitchison and Begg 1976), the Bayesian analysis of contingency tables and the analysis of compositional data (Aitchison 1982, 1986).

The use of the logistic-normal distribution is most easily seen in the analysis of binary data where the logit model (based on a logistic tolerance distribution) is extended to the *logit-normal* model. For grouped binary data with responses r_i out of m_i trials ($i = 1,\ldots,n$), the response probabilities, P_i, are assumed to have a logistic-normal distribution with $\text{logit}(P_i) = \log(P_i/(1-P_i)) \sim N(\mu_i, \sigma^2)$, where μ_i is modelled as a linear function of explanatory variables, $x_1,\ldots,x_p$. The resulting model can be summarized as

$$\begin{aligned} R_i|P_i &\sim \text{Binomial}(m_i, P_i) \\ \text{logit}(P_i)|Z &= \eta_i + \sigma Z = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \sigma Z \\ Z &\sim N(0,1) \end{aligned}$$

This is a simple extension of the basic logit model with the inclusion of a single normally distributed random effect in the linear predictor, an example of a *generalized linear mixed model*, see McCulloch and Searle (2001). Maximum likelihood estimation for this model is complicated by the fact that the likelihood has no closed form and involves integration over the normal density, which requires numerical methods using Gaussian quadrature; routines now exist as part of generalized linear mixed model fitting in all major software packages, such as SAS, R, Stata and Genstat. Approximate moment-based estimation methods make use of the fact that if σ^2 is small then, as derived in Williams (1982),

$$\begin{aligned} E[R_i] &= m_i\pi_i \quad \text{and} \\ Var(R_i) &= m_i\pi_i(1-\pi_1)[1+\sigma^2(m_i-1)\pi_i(1-\pi_i)] \end{aligned}$$

where $\text{logit}(\pi_i) = \eta_i$. The form of the variance function shows that this model is *overdispersed* compared to the binomial, that is it exhibits greater variability; the random effect Z allows for unexplained variation across the grouped observations. However, note that for binary data ($m_i = 1$) it is not possible to have overdispersion arising in this way.

About the Author

For biography *see* the entry ►Logistic Distribution.

Cross References

►Logistic Distribution
►Mixed Membership Models
►Multivariate Normal Distributions
►Normal Distribution, Univariate

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York

Aitchison J (1982) The statistical analysis of compositional data (with discussion). J R Stat Soc Ser B 44:139–177

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London

Aitchison J, Begg CB (1976) Statistical diagnosis when basic cases are not classified with certainty. Biometrika 63:1–12

Aitchison J, Shen SM (1980) Logistic-normal distributions: some properties and uses. Biometrika 67:261–272
McCulloch CE, Searle SR (2001) Generalized, linear and mixed models. Wiley, New York
Williams D (1982) Extra-binomial variation in logistic linear models. Appl Stat 31:144–148

Logistic Regression

Joseph M. Hilbe
Emeritus Professor
University of Hawaii, Honolulu, HI, USA
Adjunct Professor of Statistics
Arizona State University, Tempe, AZ, USA
Solar System Ambassador
California Institute of Technology, Pasadena, CA, USA

Logistic regression is the most common method used to model binary response data. When the response is binary, it typically takes the form of 1/0, with 1 generally indicating a success and 0 a failure. However, the actual values that 1 and 0 can take vary widely, depending on the purpose of the study. For example, for a study of the odds of failure in a school setting, 1 may have the value of *fail*, and 0 of *not-fail*, or pass. The important point is that 1 indicates the foremost subject of interest for which a binary response study is designed. Modeling a binary response variable using normal linear regression introduces substantial bias into the parameter estimates. The standard linear model assumes that the response and error terms are normally or Gaussian distributed, that the variance, σ^2, is constant across observations, and that observations in the model are independent. When a binary variable is modeled using this method, the first two of the above assumptions are violated. Analogical to the normal regression model being based on the Gaussian probability distribution function (*pdf*), a binary response model is derived from a Bernoulli distribution, which is a subset of the binomial *pdf* with the binomial denominator taking the value of 1. The Bernoulli *pdf* may be expressed as:

$$f(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \quad (1)$$

Binary logistic regression derives from the canonical form of the Bernoulli distribution. The Bernoulli *pdf* is a member of the exponential family of probability distributions, which has properties allowing for a much easier estimation of its parameters than traditional Newton–Raphson-based maximum likelihood estimation (*MLE*) methods.

In 1972 Nelder and Wedderbrun discovered that it was possible to construct a single algorithm for estimating models based on the exponential family of distributions. The algorithm was termed ▶Generalized linear models (*GLM*), and became a standard method to estimate binary response models such as logistic, probit, and complimentary-loglog regression, count response models such as Poisson and negative binomial regression, and continuous response models such as gamma and inverse Gaussian regression. The standard normal model, or Gaussian regression, is also a generalized linear model, and may be estimated under its algorithm. The form of the exponential distribution appropriate for generalized linear models may be expressed as

$$f(y_i;\theta_i,\phi) = \exp\{(y_i\theta_i - b(\theta_i))/\alpha(\phi) + c(y_i;\phi)\}, \quad (2)$$

with θ representing the link function, $\alpha(\phi)$ the scale parameter, $b(\theta)$ the cumulant, and $c(y;\phi)$ the normalization term, which guarantees that the probability function sums to 1. The link, a monotonically increasing function, linearizes the relationship of the expected mean and explanatory predictors. The scale, for binary and count models, is constrained to a value of 1, and the cumulant is used to calculate the model mean and variance functions. The mean is given as the first derivative of the cumulant with respect to θ, $b'(\theta)$; the variance is given as the second derivative, $b''(\theta)$. Taken together, the above four terms define a specific *GLM* model.

We may structure the Bernoulli distribution (3) into exponential family form (2) as:

$$f(y_i;\pi_i) = \exp\{y_i \ln(\pi_i/(1-\pi_i)) + \ln(1-\pi_i)\}. \quad (3)$$

The link function is therefore $\ln(\pi/(1-\pi))$, and cumulant $-\ln(1-\pi)$ or $\ln(1/(1-\pi))$. For the Bernoulli, π is defined as the probability of success. The first derivative of the cumulant is π, the second derivative, $\pi(1-\pi)$. These two values are, respectively, the mean and variance functions of the Bernoulli *pdf*. Recalling that the logistic model is the canonical form of the distribution, meaning that it is the form that is directly derived from the *pdf*, the values expressed in (3), and the values we gave for the mean and variance, are the values for the logistic model.

Estimation of statistical models using the *GLM* algorithm, as well as *MLE*, are both based on the log-likelihood function. The likelihood is simply a re-parameterization of the *pdf* which seeks to estimate π, for example, rather than y. The log-likelihood is formed from the likelihood by taking the natural log of the function, allowing summation

across observations during the estimation process rather than multiplication.

The traditional *GLM* symbol for the mean, μ, is typically substituted for π, when *GLM* is used to estimate a logistic model. In that form, the log-likelihood function for the binary-logistic model is given as:

$$L(\mu_i; y_i) = \sum_{i=1}^{n} \{y_i \ln(\mu_i/(1-\mu_i)) + \ln(1-\mu_i)\}, \quad (4)$$

or

$$L(\mu_i; y_i) = \sum_{i=1}^{n} \{y_i \ln(\mu_i) + (1-y_i)\ln(1-\mu_i)\}. \quad (5)$$

The Bernoulli-logistic log-likelihood function is essential to logistic regression. When *GLM* is used to estimate logistic models, many software algorithms use the deviance rather than the log-likelihood function as the basis of convergence. The deviance, which can be used as a goodness-of-fit statistic, is defined as twice the difference of the saturated log-likelihood and model log-likelihood. For logistic model, the deviance is expressed as

$$D = 2\sum_{i=1}^{n} \{y_i \ln(y_i/\mu_i) + (1-y_i)\ln((1-y_i)/(1-\mu_i))\}. \quad (6)$$

Whether estimated using maximum likelihood techniques or as *GLM*, the value of μ for each observation in the model is calculated on the basis of the linear predictor, $x'\beta$. For the normal model, the predicted fit, $\hat{y}$, is identical to $x'\beta$, the right side of (7). However, for logistic models, the response is expressed in terms of the link function, $\ln(\mu_i/(1-\mu_i))$. We have, therefore,

$$x_i'\beta = \ln(\mu_i/(1-\mu_i)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n. \quad (7)$$

The value of μ_i, for each observation in the logistic model, is calculated as

$$\mu_i = 1/\left(1+\exp\left(-x_i'\beta\right)\right) = \exp\left(x_i'\beta\right)/\left(1+\exp\left(x_i'\beta\right)\right). \quad (8)$$

The functions to the right of μ are commonly used ways of expressing the logistic inverse link function, which converts the linear predictor to the fitted value. For the logistic model, μ is a probability.

When logistic regression is estimated using a Newton-Raphson type of *MLE* algorithm, the log-likelihood function as parameterized to $x'\beta$ rather than μ. The estimated fit is then determined by taking the first derivative of the log-likelihood function with respect to β, setting it to zero, and solving. The first derivative of the log-likelihood function is commonly referred to as the gradient, or score function. The second derivative of the log-likelihood with respect to β produces the Hessian matrix, from which the standard errors of the predictor parameter estimates are derived. The logistic gradient and hessian functions are given as

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \mu_i) x_i \quad (9)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} \left\{x_i x_i' \mu_i (1-\mu_i)\right\} \quad (10)$$

One of the primary values of using the logistic regression model is the ability to interpret the exponentiated parameter estimates as odds ratios. Note that the link function is the log of the odds of μ, $\ln(\mu/(1-\mu))$, where the odds are understood as the success of μ over its failure, $1-\mu$. The log-odds is commonly referred to as the *logit* function. An example will help clarify the relationship, as well as the interpretation of the odds ratio.

We use data from the 1912 Titanic accident, comparing the odds of survival for adult passengers to children. A tabulation of the data is given as:

Survived	Age (Child vs Adult)		Total
	child	adults	
no	52	765	817
yes	57	442	499
Total	109	1,207	1,316

The odds of survival for adult passengers is 442/765, or 0.578. The odds of survival for children is 57/52, or 1.096. The ratio of the odds of survival for adults to the odds of survival for children is (442/765)/(57/52), or 0.52709552. This value is referred to as the *odds ratio*, or ratio of the two component odds relationships. Using a logistic regression procedure to estimate the odds ratio of age produces the following results

survived	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	.5270955	.1058718	−3.19	0.001	.3555642	.7813771

With 1 = *adult* and 0 = *child*, the estimated odds ratio may be interpreted as:

- *The odds of an adult surviving were about half the odds of a child surviving.*

By inverting the estimated odds ratio above, we may conclude that children had [1/.527 ~ 1.9] some 90% – or

nearly two times – greater odds of surviving than did adults.

For continuous predictors, a one-unit increase in a predictor value indicates the change in odds expressed by the displayed odds ratio. For example, if age was recorded as a continuous predictor in the Titanic data, and the odds ratio was calculated as 1.015, we would interpret the relationship as:

- *The odds of surviving is one and a half percent greater for each increasing year of age.*

Non-exponentiated logistic regression parameter estimates are interpreted as log-odds relationships, which carry little meaning in ordinary discourse. Logistic models are typically interpreted in terms of odds ratios, unless a researcher is interested in estimating predicted probabilities for given patterns of model covariates; i.e., in estimating μ.

Logistic regression may also be used for grouped or proportional data. For these models the response consists of a numerator, indicating the number of successes (1s) for a specific covariate pattern, and the denominator (m), the number of observations having the specific covariate pattern. The response y/m is binomially distributed as:

$$f(y_i;\pi_i,m_i) = \binom{m_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{m_i-y_i}, \qquad (11)$$

with a corresponding log-likelihood function expressed as

$$L(\mu_i;y_i,m_i) = \sum_{i=1}^{n}\left\{y_i\ln(\mu_i/(1-\mu_i)) + m_i\ln(1-\mu_i) + \binom{m_i}{y_i}\right\}. \qquad (12)$$

Taking derivatives of the cumulant, $-m_i\ln(1-\mu_i)$, as we did for the binary response model, produces a mean of $\mu_i = m_i\pi_i$ and variance, $\mu_i(1-\mu_i/m_i)$.

Consider the data below:

y	cases	x_1	x_2	x_3
1	3	1	0	1
1	1	1	1	1
2	2	0	0	1
0	1	0	1	1
2	2	1	0	0
0	1	0	1	0

y indicates the number of times a specific pattern of covariates is successful. *Cases* is the number of observations having the specific covariate pattern. The first observation in the table informs us that there are three cases having predictor values of $x_1 = 1, x_2 = 0$, and $x_3 = 1$. Of those three cases, one has a value of y equal to 1, the other two have values of 0. All current commercial software applications estimate this type of logistic model using *GLM* methodology.

y	Odds ratio	OIM std. err.	z	$P > \|z\|$	[95% conf.	interval]
x_1	1.186947	1.769584	0.11	0.908	0.0638853	22.05271
x_2	0.2024631	0.3241584	−1.00	0.318	0.0087803	4.668551
x_3	0.5770337	0.9126937	−0.35	0.728	0.025993	12.8099

The data in the above table may be restructured so that it is in individual observation format, rather than grouped. The new table would have ten observations, having the same logic as described. Modeling would result in identical parameter estimates. It is not uncommon to find an individual-based data set of, for example, 10,000 observations, being grouped into 10–15 rows or observations as above described. Data in tables is nearly always expressed in grouped format.

Logistic models are subject to a variety of fit tests. Some of the more popular tests include the Hosmer-Lemeshow goodness-of-fit test, *ROC* analysis, various information criteria tests, link tests, and residual analysis. The Hosmer–Lemeshow test, once well used, is now only used with caution. The test is heavily influenced by the manner in which tied data is classified. Comparing observed with expected probabilities across levels, it is now preferred to construct tables of risk having different numbers of levels. If there is consistency in results across tables, then the statistic is more trustworthy.

Information criteria tests, e.g., Akaike information Criteria (see ►Akaike's Information Criterion and ►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements) (*AIC*) and Bayesian Information Criteria (*BIC*) are the most used of this type of test. Information tests are comparative, with lower values indicating the preferred model. Recent research indicates that *AIC* and *BIC* both are biased when data is correlated to any degree. Statisticians have attempted to develop enhancements of these two tests, but have not been entirely successful. The best advice is to use several different types of tests, aiming for consistency of results.

Several types of residual analyses are typically recommended for logistic models. The references below provide extensive discussion of these methods, together with

appropriate caveats. However, it appears well established that m-asymptotic residual analyses is most appropriate for logistic models having no continuous predictors. m-asymptotics is based on grouping observations with the same covariate pattern, in a similar manner to the grouped or binomial logistic regression discussed earlier. The Hilbe (2009) and Hosmer and Lemeshow (2000) references below provide guidance on how best to construct and interpret this type of residual.

Logistic models have been expanded to include categorical responses, e.g., proportional odds models and multinomial logistic regression. They have also been enhanced to include the modeling of panel and correlated data, e.g., generalized estimating equations, fixed and random effects, and mixed effects logistic models.

Finally, exact logistic regression models have recently been developed to allow the modeling of perfectly predicted data, as well as small and unbalanced datasets. In these cases, logistic models which are estimated using GLM or full maximum likelihood will not converge. Exact models employ entirely different methods of estimation, based on large numbers of permutations.

About the Author

Joseph M. Hilbe is an emeritus professor, University of Hawaii and adjunct professor of statistics, Arizona State University. He is also a Solar System Ambassador with NASA/Jet Propulsion Laboratory, at California Institute of Technology. Hilbe is a Fellow of the American Statistical Association and Elected Member of the International Statistical institute, for which he is founder and chair of the ISI astrostatistics committee and Network, the first global association of astrostatisticians. He is also chair of the ISI sports statistics committee, and was on the founding executive committee of the Health Policy Statistics Section of the American Statistical Association (1994–1996). Hilbe is author of *Negative Binomial Regression* (2007, Cambridge University Press), and *Logistic Regression Models* (2009, Chapman & Hall), two of the leading texts in their respective areas of statistics. He is also co-author (with James Hardin) of *Generalized Estimating Equations* (2002, Chapman & Hall/CRC) and two editions of *Generalized Linear Models and Extensions* (2001, 2007, Stata Press), and with Robert Muenchen is coauthor of the *R for Stata Users* (2010, Springer). Hilbe has also been influential in the production and review of statistical software, serving as Software Reviews Editor for *The American Statistician* for 12 years from 1997–2009. He was founding editor of the *Stata Technical Bulletin* (1991), and was the first to add the negative binomial family into commercial generalized linear models software. Professor Hilbe was presented the Distinguished Alumnus award at California State University, Chico in 2009, two years following his induction into the University's Athletic Hall of Fame (he was two-time US champion track & field athlete). He is the only graduate of the university to be recognized with both honors.

Cross References

►Case-Control Studies
►Categorical Data Analysis
►Generalized Linear Models
►Multivariate Data Analysis: An Overview
►Probit Analysis
►Recursive Partitioning
►Regression Models with Symmetrical Errors
►Robust Regression Estimation in Generalized Linear Models
►Statistics: An Overview
►Target Estimation: A New Approach to Parametric Estimation

References and Further Reading

Collett D (2003) Modeling binary regression, 2nd edn. Chapman & Hall/CRC Cox, London
Cox DR, Snell EJ (1989) Analysis of binary data, 2nd edn. Chapman & Hall, London
Hardin JW, Hilbe JM (2007) Generalized linear models and extensions, 2nd edn. Stata Press, College Station
Hilbe JM (2009) Logistic regression models. Chapman & Hall/CRC Press, Boca Raton
Hosmer D, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley, New York
Kleinbaum DG (1994) Logistic regression; a self-teaching guide. Springer, New York
Long JS (1997) Regression models for categorical and limited dependent variables. Sage, Thousand Oaks
McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman & Hall, London

Logistic Distribution

John Hinde
Professor of Statistics
National University of Ireland, Galway, Ireland

The logistic distribution is a location-scale family distribution with a very similar shape to the normal (Gaussian) distribution but with somewhat heavier tails. The distribution has applications in reliability and survival analysis. The cumulative distribution function has been used for

modelling growth functions and as a tolerance distribution in the analysis of binary data, leading to the widely used *logit* model. For a detailed discussion of the properties of the logistic and related distributions, see Johnson et al. (1995).

The probability density function is

$$f(x) = \frac{1}{\tau}\frac{\exp\{-(x-\mu)/\tau\}}{[1+\exp\{-(x-\mu)/\tau\}]^2}, \quad -\infty < x < \infty \quad (1)$$

and the cumulative distribution function is

$$F(x) = \frac{1}{[1+\exp\{-(x-\mu)/\tau\}]}, \quad -\infty < x < \infty .$$

The distribution is symmetric about the mean μ and has variance $\tau^2\pi^2/3$, so that when comparing the standard logistic distribution ($\mu = 0$, $\tau = 1$) with the standard normal distribution, $N(0,1)$, it is important to allow for the different variances. The suitably scaled logistic distribution has a very similar shape to the normal, although the kurtosis is 4.2 which is somewhat larger than the value of 3 for the normal, indicating the heavier tails of the logistic distribution.

In survival analysis, one advantage of the logistic distribution, over the normal, is that both *right-* and *left-censoring* can be easily handled. The *survivor* and *hazard* functions are given by

$$S(x) = \frac{1}{[1+\exp\{(x-\mu)/\tau\}]}, \quad -\infty < x < \infty$$

$$h(x) = \frac{1}{\tau}\frac{1}{[1+\exp\{-(x-\mu)/\tau\}]}.$$

The hazard function has the same logistic form and is monotonically increasing, so the model is only appropriate for ageing systems with an increasing failure rate over time. In modelling the dependence of failure times on explanatory variables, if we use a linear regression model for μ, then the fitted model has an *accelerated failure time* interpretation for the effect of the variables. Fitting of this model to right- and left-censored data is described in Aitkin et al. (2009).

One obvious extension for modelling failure times, T, is to assume a logistic model for $\log T$, giving a *log-logistic* model for T analagous to the *lognormal model*. The resulting hazard function based on the logistic distribution in (1) is

$$h(t) = \frac{\alpha}{\theta}\frac{(t/\theta)^{\alpha-1}}{1+(t/\theta)^{\alpha}}, \quad t, \theta, \alpha > 0$$

where $\theta = e^{\mu}$ and $\alpha = 1/\tau$. For $\alpha \leq 1$ the hazard is monotone decreasing, and for $\alpha > 1$ it has a single maximum as for the lognormal distribution; hazards of this form may be appropriate in the analysis of data such as heart transplant survival – there may be an initial period of increasing hazard associated with rejection, followed by decreasing hazard as the patient survives the procedure and the transplanted organ is accepted.

For the standard logistic distribution ($\mu = 0$, $\tau = 1$), the probability density and the cumulative distribution functions are related through the very simple identity

$$f(x) = F(x)\,[1 - F(x)]$$

which in turn, by elementary calculus, implies that

$$\text{logit}(F(x)) := \log_e\left[\frac{F(x)}{1-F(x)}\right] = x \quad (2)$$

and uniquely characterizes the standard logistic distribution. Equation (2) provides a very simple way for simulating from the standard logistic distribution by setting $X = \log_e[U/(1-U)]$ where $U \sim U(0,1)$; for the general logistic distribution in (1) we take $\tau X + \mu$.

The logit transformation is now very familiar in modelling probabilities for binary responses. Its use goes back to Berkson (1944), who suggested the use of the logistic distribution to replace the normal distribution as the underlying tolerance distribution in quantal bio-assays, where various dose levels are given to groups of subjects (animals) and a simple binary response (e.g., cure, death, etc.) is recorded for each individual (giving r-out-of-n type response data for the groups). The use of the normal distribution in this context had been pioneered by Finney through his work on ▶probit analysis and the same methods mapped across to the logit analysis, see Finney (1978) for a historical treatment of this area. The probability of response, $P(d)$, at a particular dose level d is modelled by a linear logit model

$$\text{logit}(P(d)) = \log_e\left[\frac{P(d)}{1-P(d)}\right] = \beta_0 + \beta_1 d$$

which, by the identity (2), implies a logistic tolerance distribution with parameters $\mu = -\beta_0/\beta_1$ and $\tau = 1/|\beta_1|$. The logit transformation is computationally convenient and has the nice interpretation of modelling the *log-odds* of a response. This goes across to general logistic regression models for binary data where parameter effects are on the log-odds scale and for a two-level factor the fitted effect corresponds to a log-odds-ratio. Approximate methods for parameter estimation involve using the empirical logits of the observed proportions. However, maximum likelihood estimates are easily obtained with standard generalized linear model fitting software, using a binomial response distribution with a logit link function for the response probability; this uses the iteratively reweighted least-squares Fisher-scoring algorithm of

Nelder and Wedderburn (1972), although Newton-based algorithms for maximum likelihood estimation of the logit model appeared well before the unifying treatment of ►generalized linear models. A comprehensive treatment of ►logistic regression including models and applications is given in Agresti (2002) and Hilbe (2009).

About the Author

John Hinde is Professor of Statistics, School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland. He is past President of the Irish Statistical Association (2006–2008), the Statistical Modelling Society (2004–2006) and the European Regional Section of the International Association of Statistical Computing (2000–2002). He has been an active member of the International Biometric Society and is the incoming President of the British and Irish Region (2011). He is an Elected member of the International Statistical Institute (2001). He has authored or coauthored over 50 papers mainly in the area of statistical modelling and statistical computing and is coauthor of several books, including most recently *Statistical Modelling in R* (with M. Aitkin, B. Francis, and R. Darnell, Oxford University Press, 2009). He is currently an Associate Editor of *Statistics and Computing* and was one of the joint Founding Editors of *Statistical Modelling*.

Cross References

►Asymptotic Relative Efficiency in Estimation
►Asymptotic Relative Efficiency in Testing
►Bivariate Distributions
►Location-Scale Distributions
►Logistic Regression
►Multivariate Statistical Distributions
►Statistical Distributions: An Overview

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York
Aitkin M, Francis B, Hinde J, Darnell R (2009) Statistical modelling in R. Oxford University Press, Oxford
Berkson J (1944) Application of the logistic function to bio-assay. J Am Stat Assoc 39:357–365
Finney DJ (1978) Statistical method in biological assay, 3rd edn. Griffin, London
Hilbe JM (2009) Logistic regression models. Chapman & Hall/CRC Press, Boca Raton
Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distributions, vol 2, 2nd edn. Wiley, New York
Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc A 135:370–384

Lorenz Curve

Johan Fellman
Professor Emeritus
Folkhälsan Institute of Genetics, Helsinki, Finland

Definition and Properties

It is a general rule that income distributions are skewed. Although various distribution models, such as the Lognormal and the Pareto have been proposed, they are usually applied in specific situations. For general studies, more wide-ranging tools have to be applied, the first and most common tool of which is the Lorenz curve. Lorenz (1905) developed it in order to analyze the distribution of income and wealth within populations, describing it in the following way:

► *Plot along one axis accumulated percentages of the population from poorest to richest, and along the other, wealth held by these percentages of the population.*

The Lorenz curve $L(p)$ is defined as a function of the proportion p of the population. $L(p)$ is a curve starting from the origin and ending at point $(1,1)$ with the following additional properties (I) $L(p)$ is monotone increasing, (II) $L(p) \leq p$, (III) $L(p)$ convex, (IV) $L(0) = 0$ and $L(1) = 1$. The Lorenz curve is convex because the income share of the poor is less than their proportion of the population (Fig. 1).

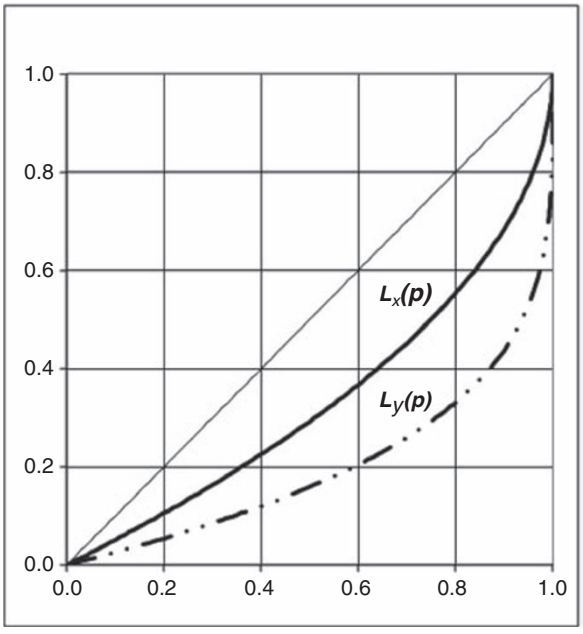
The Lorenz curve satisfies the general rules:

► *A unique Lorenz curve corresponds to every distribution. The contrary does not hold, but every Lorenz $L(p)$ is a common curve for a whole class of distributions $F(\theta x)$ where θ is an arbitrary constant.*

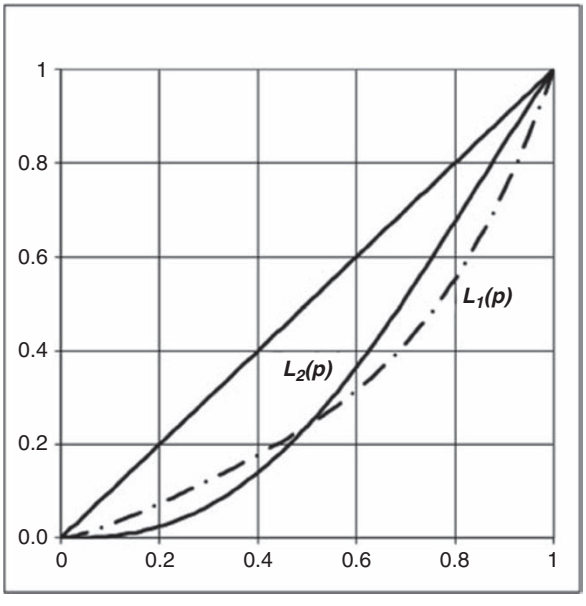
The higher the curve, the less inequality in the income distribution. If all individuals receive the same income, then the Lorenz curve coincides with the diagonal from $(0,0)$ to $(1,1)$. Increasing inequality lowers the Lorenz curve, which can converge towards the lower right corner of the square.

Consider two Lorenz curves $L_X(p)$ and $L_Y(p)$. If $L_X(p) \geq L_Y(p)$ for all p, then the distribution corresponding to $L_X(p)$ has lower inequality than the distribution corresponding to $L_Y(p)$ and is said to Lorenz dominate the other. Figure 1 shows an example of Lorenz curves.

The inequality can be of a different type, the corresponding Lorenz curves may intersect, and for these no Lorenz ordering holds. This case is seen in Fig. 2. Under such circumstances, alternative inequality measures have to be defined, the most frequently used being the Gini index, G, introduced by Gini (1912). This index is the ratio



Lorenz Curve. Fig. 1 Lorenz curves with Lorenz ordering; that is, $L_X(p) \geq L_Y(p)$



Lorenz Curve. Fig. 2 Two intersecting Lorenz curves. Using the Gini index $L_1(p)$ has greater inequality ($G_1 = 0.37$) than $L_2(p)$ ($G_2 = 0.33$)

between the area between the diagonal and the Lorenz curve and the whole area under the diagonal. This definition yields Gini indices satisfying the inequality $0 \leq G \leq 1$. The higher the G value, the greater the inequality in the income distribution.

Income Redistributions

It is a well-known fact that progressive taxation reduces inequality. Similar effects can be obtained by appropriate transfer policies, findings based on the following general theorem (Fellman 1976; Jakobsson 1976; Kakwani 1977):

Theorem *Let $u(x)$ be a continuous monotone increasing function and assume that $\mu_Y = E(u(X))$ exists. Then the Lorenz curve $L_Y(p)$ for $Y = u(X)$ exists and*

(I) $L_Y(p) \geq L_X(p)$ *if* $\frac{u(x)}{x}$ *is monotone decreasing*

(II) $L_Y(p) = L_X(p)$ *if* $\frac{u(x)}{x}$ *is constant*

(III) $L_Y(p) \leq L_X(p)$ *if* $\frac{u(x)}{x}$ *is monotone increasing.*

For progressive taxation rules, $\frac{u(x)}{x}$ measures the proportion of post-tax income to initial income and is a monotone-decreasing function satisfying condition (I), and the Gini index is reduced. Hemming and Keen (1993) gave an alternative condition for the Lorenz dominance, which is that $\frac{u(x)}{x}$ crosses the $\frac{\mu_Y}{\mu_X}$ level once from above. If the taxation rule is a flat tax, then (II) holds and the Lorenz curve and the Gini index remain. The third case in Theorem 1 indicates that the ratio $\frac{u(x)}{x}$ is increasing and the Gini index increases, but this case has only minor practical importance.

A crucial study concerning income distributions and redistributions is the monograph by Lambert (2001).

About the Author

Dr Johan Fellman is Professor Emeritus in statistics. He obtained Ph.D. in mathematics (University of Helsinki) in 1974 with the thesis On the Allocation of Linear Observations and in addition, he has published about 180 printed articles. He served as professor in statistics at Hanken School of Economics, (1977–1994) and has been a scientist at Folkhälsan Institute of Genetics since 1963. He is member of the Editorial Board of *Twin Research and Human Genetics* and of the Advisory Board of *Mathematica Slovaca* and Editor of *InterStat*. He has been awarded the Knight, First Class, of the Order of the White Rose of Finland and the Medal in Silver of Hanken School of Economics. He is Elected member of the Finnish Society of Sciences and Letters and of the International Statistical Institute.

Cross References

►Econometrics
►Economic Statistics
►Testing Exponentiality of Distribution

References and Further Reading

Fellman J (1976) The effect of transformations on Lorenz curves. Econometrica 44:823–824
Gini C (1912) Variabilità e mutabilità. Bologna, Italy
Hemming R, Keen MJ (1993) Single crossing conditions in comparisons of tax progressivity. J Publ Econ 20:373–390
Jakobsson U (1976) On the measurement of the degree of progression. J Publ Econ 5:161–168
Kakwani NC (1977) Applications of Lorenz curves in economic analysis. Econometrica 45:719–727
Lambert PJ (2001) The distribution and redistribution of income: a mathematical analysis, 3rd edn. Manchester university press, Manchester
Lorenz MO (1905) Methods for measuring concentration of wealth. J Am Stat Assoc New Ser 70:209–219

Loss Function

Wolfgang Bischoff
Professor and Dean of the Faculty of Mathematics and Geography
Catholic University Eichstätt–Ingolstadt, Eichstätt, Germany

Loss functions occur at several places in statistics. Here we attach importance to decision theory (see ►Decision Theory: An Introduction, and ►Decision Theory: An Overview) and regression. For both fields the same loss functions can be used. But the interpretation is different.

Decision theory gives a general framework to define and understand statistics as a mathematical discipline. The loss function is the essential component in decision theory. The loss function judges a decision with respect to the truth by a real value greater or equal to zero. In case the decision coincides with the truth then there is no loss. Therefore the value of the loss function is zero then, otherwise the value gives the loss which is suffered by the decision unequal the truth. The larger the value the larger the loss which is suffered.

To describe this more exactly let Θ be the known set of all outcomes for the problem under consideration on which we have information by data. We assume that one of the values $\theta \in \Theta$ is the true value. Each $d \in \Theta$ is a possible decision. The decision d is chosen according to a rule, more exactly according to a function with values in Θ and defined on the set of all possible data. Since the true value θ is unknown the loss function L has to be defined on $\Theta \times \Theta$, i.e.,

$$L : \Theta \times \Theta \to [0, \infty).$$

The first variable describes the true value, say, and the second one the decision. Thus $L(\theta, a)$ is the loss which is suffered if θ is the true value and a is the decision. Therefore, each (up to technical conditions) function $L : \Theta \times \Theta \to [0, \infty)$ with the property

$$L(\theta, \theta) = 0 \text{ for all } \theta \in \Theta$$

is a possible loss function. The loss function has to be chosen by the statistician according to the problem under consideration.

Next, we describe examples for loss functions. First let us consider a test problem. Then Θ is divided in two disjoint subsets Θ_0 and Θ_1 describing the null hypothesis and the alternative set, $\Theta = \Theta_0 + \Theta_1$. Then the usual loss function is given by

$$L(\theta, \vartheta) = \begin{cases} 0 & \text{if} \quad \theta, \vartheta \in \Theta_0 \qquad \text{or} \quad \theta, \vartheta \in \Theta_1 \\ 1 & \text{if} \quad \theta \in \Theta_0, \vartheta \in \Theta_1 \quad \text{or} \quad \theta \in \Theta_1, \vartheta \in \Theta_0 \end{cases}.$$

For point estimation problems we assume that Θ is a normed linear space and let $|\cdot|$ be its norm. Such a space is typical for estimating a location parameter. Then the loss $L(\theta, \vartheta) = |\theta - \vartheta|$, $\theta, \vartheta \in \Theta$, can be used. Next, let us consider the specific case $\Theta = \mathrm{R}$. Then $L(\theta, \vartheta) = \ell(\theta - \vartheta)$ is a typical form for loss functions, where $\ell : IR \to [0, \infty)$ is nonincreasing on $(-\infty, 0]$ and nondecreasing on $[0, \infty)$ with $\ell(0) = 0$. ℓ is also called loss function. An important class of such functions is given by choosing $\ell(t) = |t|^p$, where $p > 0$ is a fixed constant. There are two prominent cases, for $p = 2$ we get the classical square loss and for $p = 1$ the robust L_1-loss. Another class of robust losses are the famous Huber losses

$$\ell(t) = t^2/2, \text{ if } |t| \le \gamma, \text{ and } \ell(t) = \gamma|t| - \gamma^2/2, \text{ if } |t| > \gamma,$$

where $\gamma > 0$ is a fixed constant. Up to now we have shown symmetrical losses, i.e., $L(\theta, \vartheta) = L(\vartheta, \theta)$. There are many problems in which underestimating of the true value θ has to be differently judged than overestimating. For such problems Varian (1975) introduced LinEx losses

$$\ell(t) = b(exp(at) - at - 1),$$

where $a, b > 0$ can be chosen suitably. Here underestimating is judged exponentially and overestimating linearly.

For other estimation problems corresponding losses are used. For instance, let us consider the estimation of a

scale parameter and let $\Theta = (0, \infty)$. Then it is usual to consider losses of the form $L(\theta, \vartheta) = \ell(\vartheta/\theta)$, where ℓ must be chosen suitably. It is, however, more convenient to write $\ell(\ln \vartheta - \ln \theta)$. Then ℓ can be chosen as above.

In theoretical works the assumed properties for loss functions can be quite different. Classically it was assumed that the loss is convex (see ►Rao–Blackwell Theorem). If the space Θ is not bounded, then it seems to be more convenient in practice to assume that the loss is bounded which is also assumed in some branches of statistics. In case the loss is not continuous then it must be carefully defined to get no counter intuitive results in practice, see Bischoff (1999).

In case a density of the underlying distribution of the data is known up to an unknown parameter the class of divergence losses can be defined. Specific cases of these losses are the Hellinger and the Kulback-Leibler loss.

In regression, however, the loss is used in a different way. Here it is assumed that the unknown location parameter is an element of a known class $\mathcal{F}$ of real valued functions. Given n observations (data) $y_1, \ldots, y_n$ observed at design points $t_1, \ldots, t_n$ of the experimental region a loss function is used to determine an estimation for the unknown regression function by the 'best approximation', i.e., the function in $\mathcal{F}$ that minimizes $\sum_{i=1}^{n} \ell\left(r_i^f\right), f \in \mathcal{F}$, where $r_i^f = y_i - f(t_i)$ is the residual in the ith design point. Here ℓ is also called loss function and can be chosen as described above. For instance, the least squares estimation is obtained if $\ell(t) = t^2$.

Cross References

►Advantages of Bayesian Structuring: Estimating Ranks and Histograms
►Bayesian Statistics
►Decision Theory: An Introduction
►Decision Theory: An Overview
►Entropy and Cross Entropy as Diversity and Distance Measures
►Sequential Sampling
►Statistical Inference for Quantum Systems

References and Further Reading

Bischoff W (1999) Best ϕ-approximants for bounded weak loss functions. Stat Decis 17:49–61

Varian HR (1975) A Bayesian approach to real estate assessment. In: Fienberg SE, Zellner A (eds) Studies in Bayesian econometrics and statistics in honor of Leonard J. Savage, North Holland, Amsterdam, pp 195–208

M

Margin of Error

Judith M. Tanur
Distinguished Teaching Professor Emerita
Stony Brook University, Stony Brook, NY, USA

Margin of error is a term that probably originated in the popular reporting of results of ▶public opinion polls but has made its way into more professional usage. It usually represents half of the length of a confidence interval (most usually a 95% confidence interval, though it could in theory be any confidence interval) for a proportion or percentage, calculated under the assumption of simple random sampling. The sample value of the proportion, $\hat{p}$, is used as an estimate of the population proportion π, and the standard error (se) is estimated as $\sqrt{\hat{p}(1-\hat{p})/n}$. Then a 95% confidence interval is given as $\hat{p} \pm 1.96\times$ se and the margin of error is $1.96\times$ se. For example, if an opinion poll gives a result of 40% of 900 respondents in favor of a proposition (a proportion of .40), then the estimated se of the proportion is $\sqrt{(0.4 \times 0.6)/900} = .016$ and that is expressed as 1.6 percentage points. Then the margin of error would be presented as $1.96 \times 1.6 = 3.2$ percentage points.

The fact that the margin of error is often reported in the popular press represents progress from a time when sample results were not qualified at all by notions of sample-to-sample variability. Such reporting, however, is frequently subject to misinterpretation, though reporters often caution against such misinterpretation. First, like the confidence interval, the margin of error does not represent anything about the probability that the results are close to truth. A 95% confidence interval merely says that, with the procedure as carried out repeatedly by drawing a sample from this population, 95% of the time the stated interval would cover the true population parameter. There is no information whether this current interval does or does not cover the population parameter and similarly the margin of error gives no information whether it covers the true population percentage. Second, the procedure assumes simple random sampling, but frequently the sampling for a survey is more complicated than that and hence the standard error calculated under the assumption of simple random sampling is an underestimate. Third, the margin of error is frequently calculated for the sample as a whole, but when interest centers on a subgroup of respondents (e.g., the percentage of females who prefer a particular candidate) the sample size is smaller and a fresh margin of error should be calculated for the subgroup, though it frequently is not. And finally, and perhaps most importantly, there is a tendency to assume that the margin of error takes into account all possible "errors" when in fact it deals only with sampling error. Nonsampling errors, such as noncoverage, nonresponse, or inaccurate responses are not taken into account via a confidence interval or the margin of error and may indeed be of much larger magnitude than the sampling error measured by the standard error.

About the Author

For biography *see* the entry ▶Nonsampling Errors in Surveys.

Cross References

▶Confidence Interval
▶Estimation
▶Estimation: An Overview
▶Public Opinion Polls

Marginal Probability: Its Use in Bayesian Statistics as Model Evidence

Luis Raúl Pericchi
Professor
University of Puerto Rico, San Juan, Puerto Rico

Definition

Suppose that we have vectors of random variables $[\mathbf{v},\mathbf{w}] = [v_1, v_2, \ldots, v_I, w_1, \ldots, w_J]$ in $\Re^{(I+J)}$. Denote as the **joint** density function: $f_{\mathbf{v},\mathbf{w}}$, which obeys: $f_{\mathbf{v},\mathbf{w}}(v, w) \geq 0$ and

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{v},\mathbf{w}}(v,w)dv_1 \dots dv_I dw_1 \dots dw_I = 1$. Then the probability of the set $[A_v, B_w]$ is given by

$$P(A_v, B_w) = \int \dots \int_{A_v,B_w} f_{\mathbf{v},\mathbf{w}}(v,w)\mathbf{dvdw}.$$

The *marginal* density $f_{\mathbf{v}}$ is obtained as

$$f_{\mathbf{v}}(v) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\mathbf{v},\mathbf{w}}(v,w)dw_1 \dots dw_I.$$

The *marginal probability* of the set A_v is then obtained as,

$$P(A_v) = \int \dots \int_{A_v} f_{\mathbf{v}}(v)dv.$$

We have assumed that the random variables are continuous. When they are discrete, integrals are substituted by sums. We proceed to present an important application of marginal probabilities for measuring the probability of a model.

Measuring the Evidence in Favor of a Model

In Statistics, a parametric model, is denoted as $f(x_1, \dots, x_n|\theta_1, \dots, \theta_k)$, where $\mathbf{x} = (x_1, \dots, x_n)$ is the vector of n observations and $\theta = (\theta_1, \dots, \theta_k)$ is the vector of k parameters. For instance we may have $n = 15$ observations normally distributed and the vector of parameters is (θ_1, θ_2) the location and scale respectively, denoted by $f_{Normal}(\mathbf{x}|\theta) = \prod_{\mathbf{i=1}}^{\mathbf{n}} \frac{1}{\sqrt{2\pi}\theta_2} \exp\left(-\frac{1}{2\theta_2^2}(\mathbf{x}_i - \theta_1)^2\right)$.

Assume now that there is reason to suspect that the location is zero. As a second example, it may be suspected that the sampling model which usually has been assumed Normally distributed, is instead a Cauchy, $f_{Cauchy}(X|\theta) = \prod_1^n \frac{1}{\pi\theta_2} \frac{1}{\left(1+\left(\frac{x_i-\theta_1}{\theta_2}\right)^2\right)}$. The first problem is a *hypothesis test* denoted by

$$H_0 : \theta_1 = 0 \text{ VS } H_1 : \theta_1 \neq 0,$$

and the second problem is a *model selection* problem:

$$M_0 : f_{Normal} \text{ VS } M_1 : f_{Cauchy}.$$

How to measure the evidence in favor of H_0 or M_0? Instead of maximized likelihoods as it is done in traditional statistics, in ►Bayesian statistics the central concept is *the evidence* or *marginal probability density*

$$m_j(\mathbf{x}) = \int f_j(\mathbf{x}|\theta_j)\pi(\theta_j)\mathbf{d}\theta_j,$$

where j denotes either model or hypothesis j and $\pi(\theta)$ denotes the prior for the parameters under model or hypothesis j.

Marginal probabilities embodies the likelihood of a model or hypothesis in great generality and can be claimed it is the natural probabilistic quantity to compare models.

Marginal Probability of a Model

Once the marginal densities of the model j, for $j = 1, \dots, J$ models have been calculated and assuming the prior model probabilities $P(M_j), j = 1, \dots, J$ with $\sum_{j=1}^{J} P(M_j) = 1$ then, using Bayes Theorem, *the marginal probability of a model* $P(M_j|\mathbf{x})$ can be calculated as,

$$P(M_j|\mathbf{x}) = \frac{\mathbf{m_j}(\mathbf{x}) \cdot \mathbf{P}(\mathbf{M_j})}{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{m_i}(\mathbf{x}) \cdot \mathbf{P}(\mathbf{M_i})}.$$

We have then the following formula for any two models or hypotheses:

$$\frac{P(M_j|\mathbf{x})}{P(M_i|\mathbf{x})} = \frac{P(M_j)}{P(M_i)} \times \frac{m_j(\mathbf{x})}{m_i(\mathbf{x})},$$

or in words: Posterior Odds equals Prior Odds times Bayes Factor, where the Bayes Factor of M_j over M_i is

$$B_{j,i} = \frac{m_j(\mathbf{x})}{m_i(\mathbf{x})},$$

Jeffreys (1961).

In contrast to ►*p-values*, which have interpretations heavily dependent on the sample size n, and its definition is not the same as the scientific question, the posterior probabilities and Bayes Factors address the scientific question: "how probable is model or hypothesis j as compared with model or hypothesis i?," and the interpretation is the same for any sample size, Berger and Pericchi (2001). Bayes Factors and Marginal Posterior Model Probabilities have several advantages, like for example large sample consistency, that is as the sample size grows the Posterior Model Probability of the sampling model tends to one. Furthermore, if the goal is to predict future observations y_f it is **not** necessary to select one model as *the* predicting model since we may predict by the so called Bayesian Model Averaging, which if quadratic loss is assumed, the optimal predictor takes the form,

$$E[Y_f|\mathbf{x}] = \sum_{\mathbf{j=1}}^{\mathbf{J}} \mathbf{E}[\mathbf{Y_f}|\mathbf{x}, \mathbf{M_j}] \times \mathbf{P}(\mathbf{M_j}|\mathbf{x}),$$

where $E[Y_f|\mathbf{x}, \mathbf{M_j}]$ is the expected value of a future observation under the model or hypothesis M_j.

Intrinsic Priors for Model Selection and Hypothesis Testing

Having said some of the advantages of the marginal probabilities of models, the question arises: how to assign the conditional priors $\pi(\theta_j)$? In the two examples above which priors are sensible to use? The problem is *not* a simple one since it is not possible to use the usual Uniform priors since then the Bayes Factors are undetermined. To solve this problem with some generality, Berger and Pericchi (1996)

introduced the concepts of Intrinsic Bayes Factors and Intrinsic Priors. Start by splitting the sample in two subsamples $\mathbf{x} = [\mathbf{x}(\mathbf{l}), \mathbf{x}(-\mathbf{l})]$ where the training sample $\mathbf{x}(\mathbf{l})$ is as small as possible such that for $j = 1, \ldots, J : 0 < m_j(\mathbf{x}(\mathbf{l})) < \infty$. Thus starting with an improper prior $\pi^N(\theta_j)$, which does not integrate to one (for example the Uniform), by using the minimal training sample $\mathbf{x}(\mathbf{l})$, all the conditional prior densities $\pi(\theta_j|\mathbf{x}(\mathbf{l}))$ *become* proper. So we may form the Bayes Factor using the training sample $\mathbf{x}(\mathbf{l})$ as

$$B_{ji}(\mathbf{x}(\mathbf{l})) = \frac{\mathbf{m_j}(\mathbf{x}(-\mathbf{l})|\mathbf{x}(\mathbf{l}))}{\mathbf{m_i}(\mathbf{x}(-\mathbf{l})|\mathbf{x}(\mathbf{l}))}.$$

This however depends on the particular training sample $\mathbf{x}(\mathbf{l})$. So some sort of average of Bayes Factor is necessary. In Berger and Pericchi (1996) it is shown that the average should be the arithmetic average. It is also found a theoretical prior that is an approximation to the procedure just described as the sample size grows. This is called an *Intrinsic Prior*. In the examples above: (i) in the normal case, assuming for simplicity that the variance is known and $\theta_2^2 = 1$ then it turns out that the Intrinsic Prior is Normal centered at the null hypothesis $\theta_1 = 0$ and with variance 2. On the other hand in the Normal versus Cauchy example, it turns out that the improper prior $\pi(\theta_1, \theta_2) = 1/\theta_2$ is the appropriate prior for comparing the models. For other examples of Intrinsic Priors see for instance, Berger and Pericchi (1996a,b, 2001), and Moreno et al. (1998).

About the Author

Luis Raúl Pericchi is Full Professor Department of Mathematics, College of Natural Sciences, University of Puerto Rico, Rio Piedras Campus, San Juan, and Director of the Biostatistics and Bioinformatics Core of the Comprehensive Cancer Center of the University of Puerto Rico. He received his Ph.D. in 1981, Imperial College, London (his supervisor was Professor A.C. Atkinson). He was Founder Coordinator of the Graduate Studies in Statistics (1997–2000) and Director of the Department of Mathematics (2001–2006). Professor Pericchi is Elected Member of the International Statistical Institute (1989) and Past President of the Latin American Chapter of the Bernoulli Society for Probability and Mathematical Statistics (1997–2000). Dr Pericchi was Associate Editor, *International Statistical Review* (1988–1991), Associate Editor of *Bayesian Analysis* (2006–2009). He is currently Associate Editor of the Brazilean Journal of Bayesian Analysis. He has (co)-authored more than 70 scientific articles.

Cross References

►Bayes' Theorem
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Inversion of Bayes' Formula for Events
►Model Selection
►Statistical Evidence

References and Further Reading

Berger JO, Pericchi LR (1996a) The intrinsic Bayes factor for model selection and Prediction. J Am Stat Assoc 91:109–122
Berger JO, Pericchi LR (1996b) The intrinsic Bayes factors for linear models. In: Bernardo JM et al (eds) Bayesian statistics 5. Oxford University Press, London, pp 23–42
Berger JO, Pericchi LR (2001) Objective Bayesian methods for model selection: introduction and comparison. IMS LectureNotes-Monograph Series 38:135–207
Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, London
Moreno E, Bertolino F, Racugno W (1998) An intrinsic limiting procedure for model selection and hypothesis testing. J Am Stat Assoc 93(444):1451–1460

Marine Research, Statistics in

Gunnar Stefansson
Professor, Director of the Statistical Center
University of Iceland, Reykjavik, Iceland

Marine science is a wide field of research, including hydrography, chemistry, biological oceanography and fishery science. One may consider that the longer-term aspects of global warming and issues with pollution monitoring are the most critical statistical modeling issues. Somewhat subjectively, the next in line are probably issues which relate to the sustainable use of marine resources, commonly called fishery science. Statistics enters all of the above subfields but the most elaborate models have been developed for fishery science and aspects of these will mainly be described here. Within marine research it was quite common up through about 1980 to use models of the biological processes set up using differential equations, but had no error component and basically transformed observed data through an arbitrary computational mechanism into desired measures of population size, growth, yield potential and so forth (Baranov 1918; Beverton and Holt 1957; Gulland 1965).

Data in fishery science are quite noisy for several reasons. One source of variation is measurement error and one should expect considerable variability in data which

are almost always collected indirectly. Thus one cannot observe the marine community through simple population measurements but only with surveys (bottom trawl, divers etc) or sampling of catch, both of which will provide measures which only relate indirectly to the corresponding stock parameters, are often biased and always quite variable. The second source of variation is due to the biological processes themselves, all of which have natural variation. A typical such process is the recruitment process, i.e., the production of a new yearclass by the mature component of the stock in question. Even for biology, this process is incredibly variable and it is quite hard to extract meaningful signals out of the noise. Unfortunately this process is the single most important process with regard to sustainable utilization (Beverton and Holt 1957, 1993).

As is to be expected, noisy input data will lead to variation in estimates of stock sizes, productivity and predictions (Patterson et al. 2001). As is well-known to statisticians, it is therefore important not only to obtain point estimates but also estimates of variability. In addition to the general noise issue, fisheries data are almost never i.i.d. and examples show how ignoring this can easily lead to incorrect estimates of stock size, state of utilization and predictions (Myers and Cadigan 1995).

Bayesian approaches have been used to estimate stock sizes (Patterson 1999). A particular virtue of Bayesian analysis in this context is the potential to treat natural mortality more sensibly than in other models. The natural mortality rate, M, is traditionally treated as a constant in parametric models and it turns out that this is very hard to estimate unless data are quite exceptional. Thus, M is commonly assumed to be a known constant and different values are tested to evaluate the effect of different assumptions. The Bayesian approach simply sets a prior on the natural mortality like all other parameters and the resulting computations extend all the way into predictions. Other methods typically encounter problems in the prediction phase where it is difficult to encompass the uncertainty in M in the estimate of prediction uncertainty.

One approach to extracting general information on difficult biological parameters is to consider several stocks and even several species. For the stock-recruit question it is clear when many stocks are considered that the typical behavior is such that the stock tend to produce less at low stock sizes, but this signal can rarely be seen for individual stocks. Formalizing such analyses needs to include parameters (as random effects) for each stock and combining them reduces the noise enough to provide patterns which otherwise could not be seen (see e.g., Myers et al. 1999).

In addition to the overall view of sustainable use of resources, many smaller statistical models are commonly considered. For example, one can model growth alone, typically using a nonlinear model, sometimes incorporating environmental effects and/or random effects (Miller 1992; Taylor and Stefansson 1999; Brandão et al. 2004; Gudmundsson 2005).

Special efforts have been undertaken to make the use of nonlinear and/or random effects models easier for the user (Skaug 2002; Skaug and Fournier 2006). Although developed for fishery science, these are generic C++-based model-building languages which undertake automatic differentiation transparently to the user (Fournier 1996).

Most of the above models have been developed for "data-rich" scenarios but models designed for less informative data sets abound. Traditionally these include simple models which were non-statistical and were simply a static model of equilibrium catch but a more time-series orientated approach was set up by Collie and Sissenwine (1983). In some cases these simple population models have been extended to formal random effects models (Conser 1991; Trenkel 2008).

At the other extreme of the complexity scale, several multispecies models have been developed, some of which are formal statistical models (Taylor et al. 2007), though most are somewhat ad-hoc and do not take a statistical approach (Helgason and Gislason 1979; Fulton et al. 2005; Pauly et al. 2000). Simple mathematical descriptions of species interactions are not sufficient here since it is almost always essential to take into account spatial variation in species overlap, different nursery and spawning areas and so forth. For these reasons a useful multispecies model needs to take into account multiple areas, migration and maturation along with several other processes (Stefansson and Palsson 1998). To become statistical models, these need to be set up in the usual statistical manner with likelihood functions, parameters to be formally estimated, methods to estimate uncertainty and take into account the large number of different data sources available through appropriate weighting or comparisons (Richards 1991; Stefansson 1998, 2003).

In the year 2010, the single most promising venue of further research concerns the use of random effects in nonlinear fisheries models. Several of these have been described by Venables and Dichmont (2004) and some examples go a few decades back in time as seen above, often in debated implementations (de Valpine and Hilborn 2005). How this can be implemented in the context of complex multispecies models remains to be seen.

Cross References

►Adaptive Sampling
►Bayesian Statistics

►Mathematical and Statistical Modeling of Global Warming
►Statistical Inference in Ecology

References and Further Reading

Baranov FI (1918) On the question of the biological basis of fisheries. Proc Inst Icht Invest 1(1):81–128
Beverton RJH, Holt SJ (1957) On the dynamics of exploited fish populations, vol 19. Marine Fisheries, Great Britain Ministry of Agriculture, Fisheries and Food
Beverton RJH, Holt SJ (1993) On the dynamics of exploited fish populations, vol 11. Chapman and Hall, London
Brandão A, Butterworth DS, Johnston SJ, Glazer JP (2004) Using a GLMM to estimate the somatic growth rate trend for male South African west coast rock lobster, Jasus lalandii. Fish Res 70(2–3):339–349, 2004
Collie JS, Sissenwine MP (1983) Estimating population size from relative abundance data measured with error. Can J Fish Aquat Sci 40:1871–1879
Conser RJ (1991) A delury model for scallops incorporating length-based selectivity of the recruiting year-class to the survey gear and partial recruitment to the commercial fishery. Northeast Regional Stock Assessment Workshop Report, Woods Hole, MA, Res. Doc. SAW12/2, Appendix to CRD-91-03, 18pp
de Valpine P, Hilborn R (2005) State-space likelihoods for nonlinear fisheries timeseries. Can J Fish Aquat Sci 62(9):1937–1952
Fournier DA (1996) AUTODIF. A C++ array language extension with automatic differentiation for use in nonlinear modeling and statistic. Otter Research, Nanaimo, BC, 1996
Fulton EA, Smith ADM, Punt AE (2005) Which ecological indicators can robustly detect effects of fishing? ICES J Marine Sci 62(3):540
Gudmundsson G (2005) Stochastic growth. Can J Fish Aquat Sci 62(8):1746–1755
Gulland JA (1965) Estimation of mortality rates. Annex to Arctic Fisheries Working Group Report. ICES (Int. Counc. Explor. Sea) Document C.M. D:3 (mimeo), 1965
Helgason T, Gislason H (1979) VPA-analysis with species interaction due to predation. ICES C.M. 1979/G:52
Millar RB (1992) Modelling environmental effects on growth of cod: fitting to growth increment data versus fitting to size-at-age data. ICES J Marine Sci 49(3):289
Myers RA, Cadigan NG (1995) Statistical analysis of catch-at-age data with correlated errors. Can J Fish Aquat Sci (Print) 52(6):1265–1273
Myers RA, Bowen KG, Barrowman NJ (1999) Maximum reproductive rate of fish at low population sizes. Can J Fish Aquat Sci 56(12):2404–2419
Patterson KR (1999) Evaluating uncertainty in harvest control law catches using Bayesian Markov chain Monte Carlo virtual population analysis with adaptive rejection sampling and including structural uncertainty. Can J Fish Aquat Sci 56(2):208–221
Patterson K, Cook R, Darby C, Gavaris S, Kell L, Lewy P, Mesnil B, Punt A, Restrepo V, Skagen DW, Stefansson G (2001) Estimating uncertainty in fish stock assessment and forecasting. Fish Fish 2(2):125–157
Pauly D, Christensen V, Walters C (2000) Ecopath, Ecosim, and Ecospace as tools for evaluating ecosystem impact of fisheries. ICES J Marine Sci 57(3):697
Richards LJ (1991) Use of contradictory data sources in stock assessments. Fish Res 11(3–4):225–238
Skaug HJ (2002) Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. J Comput Gr Stat pp 458–470
Skaug HJ, Fournier DA (2006) Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Comput Stat Data Anal 51(2):699–709
Stefansson G (1998) Comparing different information sources in a multispecies context. In Funk F, Quinn II TJ, Heifetz J, Ianelli JN, Powers JE, Schweigert JF, Sullivan PJ, Zhang CI (eds.), Fishery Stock Assessment Models: Proceedings of the international symposium; Anchorage 1997, 15th Lowell Wakefield Fisheries Symposium, pp 741–758
Stefansson G (2003) Issues in multispecies models. Natural Res Model 16(4):415–437
Stefansson G, Palsson OK (1998) A framework for multispecies modelling of boreal systems. Rev Fish Biol Fish 8:101–104
Taylor L, Stefansson G (1999) Growth and maturation of haddock (Melanogrammus aeglefinus) in icelandic waters. J Northwest Atlantic Fish Sci 25:101–114
Taylor L, Begley J, Kupca V, Stefansson G (2007) A simple implementation of the statistical modelling framework Gadget for cod in Icelandic waters. African J Marine Sci 29(2):223–245, AUG 2007. ISSN 1814-232X. doi: 10.2989/AJMS.2007.29.2.7.190
Trenkel VM (2008) A two-stage biomass random effects model for stock assessment without catches: what can be estimated using only biomass survey indices? Can J Fish Aquat Sci 65(6):1024–1035
Venables WN, Dichmont CM (2004) GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. Fish Res 70(2–3):319–337

Markov Chain Monte Carlo

Siddhartha Chib
Harry C. Hartkopf Professor of Econometrics and Statistics
Washington University in St. Louis, St. Louis, MO, USA

Introduction

Suppose that π is a probability measure on the probability space $(S, \mathcal{A})$, h is a measurable function from $S \to \mathrm{R}$, and one is interested in the calculation of the expectation

$$\bar{h} = \int h d\pi$$

assuming that the integral exists. In many problems, especially when the sample space S is multivariate or when the normalizing constant of π is not easily calculable, finding the value of this integral is not feasible either by numerial methods of integration (such as the method of quadrature) or by classical Monte Carlo methods (such as the method of rejection sampling). In such instances, it is usually possible to find $\bar{h}$ by Markov chain Monte Carlo, or MCMC for short, a method that stems from Metropolis et al. (1953)

in connection with work related to the hydrogen bomb project. It found early and wide use in computational statistical mechanics and quantum field theory where it was used to sample the coordinates of a point in phase space. Applications and developments of this method in statistics, in particular for problems arising in ►Bayesian statistics, can be traced to Hastings (1970), Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990).

The idea behind MCMC is to generate a sequence of draws $\{\boldsymbol{\psi}^{(g)}, g \geq 0\}$ that follow a Markov chain (see ►Markov Chains) with the property that the unique invariant distribution of this Markov chain is the target distribution π. Then, after ignoring the first n_0 draws to remove the effect of the initial value $\boldsymbol{\psi}^{(0)}$, the sample

$$\{\boldsymbol{\psi}^{(n_0+1)}, ..., \boldsymbol{\psi}^{(n_0+M)}\}$$

for M large, is taken as an approximate sample from π and $\bar{h}$ estimated by the sample average

$$M^{-1}\sum_{g=1}^{M} h(\boldsymbol{\psi}^{(n_0+g)})$$

Laws of large numbers for Markov chains show that

$$M^{-1}\sum_{g=1}^{M} h(\boldsymbol{\psi}^{(n_0+g)}) \rightarrow \int h d\pi$$

as the simulation sample size M goes to infinity (Tierney 1994; Chib and Greenberg 1995; Chen et al. 2000; Liu 2001; Robert and Casella 2004).

A key reason for the interest in MCMC methods is that, somewhat surprisingly, it is straightforward to construct one or more Markov chains whose limiting invariant distribution is the desired target distribution. A leading method is the Metropolis–Hasting (M-H) method.

Metropolis–Hastings method

In the Metropolis–Hastings method, as the Hastings (1970) extension of the Metropolis et al. (1953) method is called, the Markov chain simulation is constructed by a recursive two step process.

Let $\pi(\boldsymbol{\psi})$ be a probabiliy measure that is dominated by a sigma-finite measure μ. Let the density of π with respect to μ be denoted by $p(\cdot)$. Let $q(\boldsymbol{\psi}, \boldsymbol{\psi}^{\dagger})$ denote a conditional density for $\boldsymbol{\psi}^{\dagger}$ given $\boldsymbol{\psi}$ with respect to μ. This density $q(\boldsymbol{\psi}, \cdot)$ is referred to as the proposal or candidate generating density. Then, the Markov chain in the M-H algorithm is constructed in two steps as follows.

Step 1 Sample a proposal value $\boldsymbol{\psi}^{\dagger}$ from $q(\boldsymbol{\psi}^{(g)}, \boldsymbol{\psi})$ and calculate the quantity (the *acceptance probability* or the *probability of move*)

$$\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}^{\dagger}) = \begin{cases} \min\left[\frac{p(\boldsymbol{\psi}^{\dagger})q(\boldsymbol{\psi}^{\dagger}, \boldsymbol{\psi})}{p(\boldsymbol{\psi})q(\boldsymbol{\psi}, \boldsymbol{\psi}^{\dagger})}, 1\right] & \text{if } p(\boldsymbol{\psi})q(\boldsymbol{\psi}, \boldsymbol{\psi}^{\dagger}) > 0; \\ 1 & \text{otherwise}. \end{cases}$$

Step 2 Set

$$\boldsymbol{\psi}^{(g+1)} = \begin{cases} \boldsymbol{\psi}^{\dagger} & \text{with prob } \alpha(\boldsymbol{\psi}^{(g)}, \boldsymbol{\psi}^{\dagger}) \\ \boldsymbol{\psi}^{(g)} & \text{with prob } 1 - \alpha(\boldsymbol{\psi}^{(g)}, \boldsymbol{\psi}^{\dagger}) \end{cases}$$

If the proposal value is rejected then the next sampled value is taken to be the current value which means that when a rejection occurs the current value is repeated and the chain stays at the current value. Given the new value, the same two step process is repeated and the whole process iterated a large number of times.

Given the form of the acceptance probability $\alpha(\boldsymbol{\psi}, \boldsymbol{\psi}')$ it is clear that the M-H algorithm does not require knowledge of the normalizing constant of $p(\cdot)$. Furthermore, if the proposal density satisfies the symmetry condition $q(\boldsymbol{\psi}, \boldsymbol{\psi}') = q(\boldsymbol{\psi}', \boldsymbol{\psi})$, the acceptance probability reduces to $p(\boldsymbol{\psi}')/p(\boldsymbol{\psi})$; hence, if $p(\boldsymbol{\psi}') \geq p(\boldsymbol{\psi})$, the chain moves to $\boldsymbol{\psi}'$, otherwise it moves to $\boldsymbol{\psi}'$ with probability given by $p(\boldsymbol{\psi}')/p(\boldsymbol{\psi})$. The latter is the algorithm originally proposed by Metropolis et al. (1953).

A full expository discussion of this algorithm, along with a derivation of the method from the logic of reversibility, is provided by Chib and Greenberg (1995).

The M-H method delivers variates from π under quite general conditions. A weak requirement for a law of large numbers for sample averages based on the M-H output involve positivity and continuity of $q(\boldsymbol{\psi}, \boldsymbol{\psi}')$ for $(\boldsymbol{\psi}, \boldsymbol{\psi}')$ and connectedness of the support of the target distribution. In addition, if π is bounded then conditions for ergodicity, required to establish the central limit theorem (see ►Central Limit Theorems), are satisfied (Tierney 1994).

It is important that the proposal density be chosen to ensure that the chain makes large moves through the support of the invariant distribution without staying at one place for many iterations. Generally, the empirical behavior of the M-H output is monitored by the autocorrelation time of each component of $\boldsymbol{\psi}$ defined as

$$\left\{1 + 2\sum_{s=1}^{M} \rho_{ks}\right\},$$

where ρ_{ks} is the sample autocorrelation at lag s for the kth component of $\boldsymbol{\psi}$, and by the acceptance rate which is the proportion of times a move is made as the sampling proceeds. Because independence sampling produces an autocorrelation time that is theoretically equal to one, one tries to tune the M-H algorithm to get values close to one, if possible.

Different proposal densities give rise to specific versions of the M-H algorithm, each with the correct invariant distribution π. One family of candidate-generating densities is given by $q(\psi, \psi') = q(\psi' - \psi)$. The candidate ψ' is thus drawn according to the process $\psi' = \psi + z$, where z follows the distribution q, and is referred to as the random walk M-H chain. The random walk M-H chain is perhaps the simplest version of the M-H algorithm and is quite popular in applications. One has to be careful, however, in setting the variance of z because if it is too large it is possible that the chain may remain stuck at a particular value for many iterations while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution. Hastings (1970) considers a second family of candidate-generating densities that are given by the form $q(\psi, \psi') = q(\psi')$. Proposal values are thus drawn independently of the current location ψ.

Multiple-Block M-H

In applications when the dimension of ψ is large it is usually necessary to construct the Markov chain simulation by first grouping the variables ψ into smaller blocks. Suppose that two blocks are adequate and that ψ is written as (ψ_1, ψ_2), with $\psi_k \in \Omega_k \subseteq \Re^{d_k}$. In that case the M-H algorithm requires the specification of two proposal densities,

$$q_1\left(\psi_1, \psi_1^{\dagger} | \psi_2\right) \; ; \; q_2\left(\psi_2, \psi_2^{\dagger} | \psi_1\right),$$

one for each block ψ_k, where the proposal density q_k may depend on the current value of the remaining block. Also, define

$$\alpha\left(\psi_1, \psi_1^{\dagger} | \psi_2\right) = \min\left\{\frac{p\left(\psi_1^{\dagger}, \psi_2\right) q_1\left(\psi_1^{\dagger}, \psi_1 | \psi_2\right)}{p(\psi_1, \psi_2) q_1\left(\psi_1, \psi_1^{\dagger} | \psi_2\right)}, 1\right\}$$

and

$$\alpha\left(\psi_2, \psi_2^{\dagger} | \psi_1\right) = \min\left\{\frac{p\left(\psi_1, \psi_2^{\dagger}\right) q_2\left(\psi_2^{\dagger}, \psi_2 | \psi_1\right)}{p(\psi_1, \psi_2) q_2\left(\psi_2, \psi_2^{\dagger} | \psi_1\right)}, 1\right\},$$

as the probability of move for block ψ_k conditioned on the other block. Then, one cycle of the algorithm is completed by updating each block using a M-H step with the above probability of move, given the most current value of the other block.

Gibbs Sampling

A special case of the multiple-block M-H method is the Gibbs sampling method which was introduced by Geman and Geman (1984) in the context of image-processing and broadened for use in Bayesian problems by Gelfand and Smith (1990). To describe this algorithm, suppose that the parameters are grouped into two blocks (ψ_1, ψ_2) and each block is sampled according to the full conditional distribution of block ψ_k,

$$p(\psi_1 | \psi_2) \; ; \; p(\psi_2 | \psi_1)$$

defined as the conditional distribution under π of ψ_k given the other block. In parallel with the multiple-block M-H algorithm, the most current value of the other block is used in sampling the full conditional distribution. Derivation of these full conditional distributions is usually quite simple since, by ▶Bayes' theorem, each full conditional is proportional to $p(\psi_1, \psi_2)$, the joint distribution of the two blocks. In addition, the introduction of latent or auxiliary variables can sometimes simplify the calculation and sampling of the full conditional distributions. Albert and Chib (1993) develop such an approach for the Bayesian analysis of categorical response data.

Concluding Remarks

Some of the recent theoretical work on MCMC methods is related to the question of the rates of convergence (Cai 2000; Fort et al. 2003; Jarner and Tweedie 2003; Douc et al. 2007) and in the development of adaptive MCMC methods (Atchade and Rosenthal; Andrieu and Moulines 2005; 2006).

The importance of MCMC methods in statistics and in particular Bayesian statistics cannot be overstated. The remarkable growth of Bayesian thinking over the last 20 years was made possible largely by the innovative use of MCMC methods. Software programs such as WINBUGS and the various MCMC packages in R have contributed to the use of MCMC methods in applications across the sciences and social sciences (Congdon 2006) and these applications are likely to continue unabated.

About the Author

Siddhartha Chib is the Harry Hartkopf Professor of Econometrics and Statistics at the Olin Business School, Washington University in St. Louis. He is a Fellow of the American Statistical Association and the Director of the NBER-NSF Seminar in Bayesian Inference in Econometrics and Statistics. Professor Chib has made several contributions in the areas of binary, categorical and censored response models, the Metropolis-Hastings algorithm and MCMC methods, the estimation of the marginal likelihood and Bayes factors, and in the treatment of hidden Markov and change-point models, and stochastic volatility and diffusion models. He has served as an Associate Editor

of the *Journal of the American Statistical Association* (Theory and Methods), *Journal of Econometrics*, the *Journal of Business and Economics Statistics*, and others. Currently he is an Associate Editor of the *Journal of Computational and Graphical Statistics*, and *Statistics and Computing*.

Cross References

►Bayesian Reliability Modeling
►Bayesian Statistics
►Bootstrap Methods
►Markov Chains
►Model Selection
►Model-Based Geostatistics
►Monte Carlo Methods in Statistics
►Non-Uniform Random Variate Generations
►Rubin Causal Model
►Small Area Estimation
►Social Network Analysis
►Statistics: An Overview

References and Further Reading

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc 88:669–679
Andrieu C, Moulines E (2006) On the ergodicity properties of some adaptive MCMC algorithms. Ann Appl Probab 16:1462–1505
Atchade YF, Rosenthal JS (2005) On adaptive Markov Chain Monte Carlo algorithms. Bernoulli 11:815–828
Cai HY (2000) Exact bound for the convergence of Metropolis chains. Stoch Anal Appl 18:63–71
Chen MH, Shao QM, Ibrahim JG (2000) Monte Carlo methods in Bayesian computation. Springer, New York
Chib S, Greenberg E (1995) Understanding the Metropolis-Hastings algorithm. Am Stat 49(4):327–335
Congdon P (2006) Bayesian statistical modelling, 2nd edn. Wiley, Chichester
Douc R, Moulines E, Soulier P (2007) Computable convergence rates for subgeometric ergodic Markov chains. Bernoulli 13:831–848
Fort G, Moulines E, Roberts GO, Rosenthal JS (2003) On the geometric ergodicity of hybrid samplers. J Appl Probab 40:123–146
Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409
Geman S, Geman D (1984) Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. IEEE Trans PAMI 6: 721–741
Hastings WK (1970) Monte-Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109
Jarner SF, Tweedie RL (2003) Necessary conditions for geometric and polynomial ergodicity of random-walk-type markov chains. Bernoulli 9:559–578
Liu JS (2001) Monte Carlo strategies in scientific computing. Springer, New York
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1092
Robert CP, Casella G (2004) Monte Carlo statistical methods, 2nd edn. Springer, New York
Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82:528–550 (with discussion)
Tierney L (1994) Markov-chains for exploring posterior distributions. Ann Stat 22:1701–1728

Markov Chains

Arnoldo Frigessi[1,2], Bernd Heidergott[3]
[1]Director
Norwegian Centre for Research-Based Innovation "Statistics for Innovation," Oslo, Norway
[2]Professor
University of Oslo & Norwegian Computing Centre, Oslo, Norway
[3]Associate Professor
Vrije Universiteit, Amsterdam, The Netherlands

Introduction

Markov chains, which comprise Markov chains and ►Markov processes, have been successfully applied in areas as divers as biology, finance, manufacturing, telecommunications, physics and transport planning, and even for experts it is impossible to have an overview on the full richness of Markovian theory. Roughly speaking, Markov chains are used for modeling how a system moves from one state to another at each time point. Transitions are random and governed by a conditional probability distribution which assigns a probability to the move into a new state, given the current state of the system. This dependence represents the memory of the system. A basic example of a Markov chain is the so-called random walk defined as follows. Let $X_t \in N$, for $t \in \mathbb{N}$, be a sequence of random variables with initial value $X_0 = 0$. Furthermore assume that $P(X_{t+1} = X_t + 1 | X_t \geq 1) = p = 1 - P(X_{t+1} = X_t - 1 | X_t \geq 1)$. The sequence $X = \{X_t : t \in \mathbb{N}\}$ is an example of a Markov chain (for a detailed definition see below) and the aspects of X one is usually interested in in Markov chain theory is (i) whether X returns to 0 in a finite number of steps (this holds for $0 \leq p \leq 1/2$), (ii) the expected number of steps until the chain returns to 0 (which is finite for $0 \leq p < 1/2$), and (iii) the limiting behavior of X_t.

In the following we present some realistic examples. A useful model in modeling infectious diseases assumes that there are four possible states: Susceptible (S), Infected (I), Immune (A), Dead (R). Possible transitions are from S to I, S or R; from I to A or R; from A to A or R; from R to R only. The transitions probabilities, from S to I, S to R

and the loop S to S, must sum to one and can depend on characteristics of the individuals modeled, like age, gender, life style, etc. All individuals start in S, and move at each time unit (say a day). Given observations of the sequence of visited states (called trajectory) for a sample of individuals, with their personal characteristics, one can estimate the transition probabilities, by ►logistic regression, for example. This model assumes that the transition probability at time t from one state A to state B, only depends on the state A, and not on the trajectory that lead to A. This might not be realistic, as for example a perdurance in the diseased state I over many days, could increase the probability of transition to R. It is possible to model a system with longer memory, and thus leave the simplest setting of a Markov Chain (though one can formulate such a model still as a Markov Chain over a more complex state space which includes the length of stay in the current state). A second example refers to finance. Here we follow the daily value in Euro of a stock. The state space is continuous, and one can model the transitions from state x Euro to y Euro with an appropriate Normal density with mean $x - y$. The time series of the value of the stock might well show a longer memory, which one would typically model with some autoregressive terms, leading to more complex process again. As a further example, consider the set of all web pages on the Internet as the state space of a giant Markov chain, where the user clicks from one page to the next, according to a transition probability. A Markov Chain has been used to model such a process. The transitions from the current web page to the next web page can be modeled as a mixture of two terms: with probability λ the user follows one of the links present in the current web page and among these uniformly; with probability $1 - \lambda$ the user chooses another web page at random among all other ones. Typically $\lambda = 0.85$. Again, one could discuss how correct the assumption is, that only the current web page determines the transition probability to the next one. The modeler has to critically validate such hypothesis before trusting results based on the Markov Chain model, or chains with higher order of memory. In general a stochastic process has the Markov property if the probability to enter a state in the future is independent of the states visited in the past given the current state. Finally, Markov Chain Monte Carlo (MCMC) algorithms (see ►Markov Chain Monte Carlo) are Markov chains, where at each iteration, a new state is visited according to a transition probability that depends on the current state. These stochastic algorithm are used to sample from a distribution on the state space, which is the marginal distribution of the chain in the limit, when enough iterations have been performed.

In the literature the term Markov processes is used for Markov chains for both discrete- and continuous time cases, which is the setting of this paper. Standard textbooks on Markov chains are Kijima (1997), Meyn and Tweedie (1993), Nummelin (1984), Revuz (1984). In this paper we follow (Iosifescu 1980) and use the term 'Markov chain' for the discrete time case and the term 'Markov process' for the continuous time case. General references on Markov chains are Feller (1968), Gilks et al. (1995), Haeggstroem (2002), Kemeny and Snell (1960), Seneta (1973).

Discrete Time Markov Chains

Consider a sequence of random variables $X = \{X_t : t \in \mathbb{N}\}$ defined on a common underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with state discrete space $(S, \mathcal{S})$, i.e., X_t is $\mathcal{F} - \mathcal{S}$-measurable for $t \in \mathbb{N}$. The defining property of a Markov chain is that the distribution of X_{t+1} depends on the past only through the immediate predecessor X_t, i.e., given $X_0, X_1, \ldots, X_t$ it holds that

$$\mathbb{P}(X_{t+1} = x | X_0 = x_0, X_1 = x_1, \ldots X_{t-1} = x_{t-1}, X_t = y) = \mathbb{P}(X_{t+1} = x | X_t = y),$$

where x, y and all other x_i are element of the given state space S. If $\mathbb{P}(X_{t+1} = x | X_t = y)$ does not depend on t, the chain is called *homogenous* and it is called inhomogeneous otherwise. Provided that S is at most countable, the transition probabilities of a homogeneous Markov Chain are given by $P = (p_{x,y})_{S \times S}$, where $p_{x,y} = \mathbb{P}(X_{t+1} = y | X_t = x)$ is the probability of a transition from x to y. The matrix P is called the *one-step transition probability matrix* of the Markov chain. For the introductory ►random walk example the transition matrix is given by $p_{i,i+1} = p$, $p_{i,i-1} = p - 1$, for $i \geq 1, p_{o,1} = 1$ and otherwise zero, for $i \in \mathbb{Z}$. The row sums are one and the k-th power of the transition matrix represent the probability to move between states in k time units.

In order to fully define a Markov Chain it is necessary to assign an initial distribution $\mu = (\mathbb{P}(X_0 = s) : s \in S)$. The marginal distribution at time t can then be computed, for example, as

$$\mathbb{P}(X_t = x) = \sum_{s \in S} p_{s,x}^{(t)} \mathbb{P}(X_0 = s),$$

where $p_{s,x}^{(t)}$ denotes the s, x element of the t-th power of the transition matrix. Note that given an initial distribution μ and a transition matrix P, the distribution of the Markov chain X is uniquely defined.

A Markov chain is said to be *aperiodic* if for each pair of states i, j the greatest common divisor of the set of all t such that $p_{ij}^{(t)} > 0$ is one. Note that the random walk in

our introductory example fails to be aperiodic as any path from starting in 0 and returning there has a length that is a multiple of 2.

A distribution $(\pi_i : i \in S)$ is called a *stationary distribution* of P if

$$\pi P = \pi.$$

A key topic in Markov chain theory is the study of the limiting behavior of X. Again, with initial distribution μ, X has limiting distribution ν for initial distribution μ if

$$\lim_{t\to\infty} \mu P^t = \nu. \tag{1}$$

Note that any limiting distribution is a stationary distribution. A case of particular interest is that when X has a unique stationary distribution, which is then also the unique limiting distribution and thus describes the limit behavior of the Markov chain. If P fails to be aperiodic, then the limit in (1) may not exists and should be replaced by the Cesaro limit

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} \mu P^k = \nu,$$

which always exists for finite Markov chains.

A Markov chain is called *ergodic* if the limit in (1) is independent of the initial distribution. Consequently, an ergodic Markov chain has a unique limiting distribution and this limiting distribution is also a stationary distribution, and since any stationary distribution is a limiting distribution it is also unique.

A Markov chain is called *irreducible* if for any pair of states $i, j \in S$, there exists a path from i to j that X will follow with positive probability. In words, any state can be reached from any other state with positive probability. An irreducible Markov chain is called *recurrent* if the number of steps from a state i to the first visit of a state j, denoted by $\tau_{i,j}$, is almost surely finite for all $i, j \in S$, and it is called *positive recurrent* if $\mathbb{E}[\tau_{i,i}] < \infty$ for at least one $i \in S$. Note that for $p = 1/2$ the random walk is recurrent and for $p < 1/2$ it is positive recurrent.

The terminology developed so far allows to present the main result of Markov chain theory: Any aperiodic, irreducible and positive recurrent Markov chain P possesses a unique stationary distribution π which is the unique probability vector solving $\pi P = \pi$ (and which is also the unique limiting distribution). This ▶ergodic theorem is one of the central results and it has been established in many variations and extensions, see the references. Also, efficient algorithms for computing π have been a focus of research as for Markov chains on large state-spaces computing π is a non-trivial task.

An important topic of the statistics of Markov chains is to estimate the (one-step) transition probabilities. Consider a discrete time, homogeneous Markov chain with finite state space $S = \{1, 2, \ldots, m\}$, observed at time points $0, 1, 2, \ldots, T$ on the trajectory $s_0, s_2, s_2, \ldots, s_T$. We wish to estimate the transition probabilities $p_{i,j}$ by maximum likelihood. The likelihood is

$$\begin{aligned}\mathbb{P}(X_0 = s_0) \prod_{t=0}^{'T-1} \mathbb{P}(X_t = s_{t+1} | X_y = s_t) \\ = \mathbb{P}(X_0 = s_0) \prod_{i=1}^{m}\prod_{j=1}^{m} p_{i,j}^{k(i,j)}\end{aligned}$$

where $k(i, j)$ is the number of transitions from i to j in the observed trajectory. Ignoring the initial factor, the maximum likelihood estimator of $p_{i,j}$ is found to be equal to $\hat{p}_{i,j} = \frac{k(i,j)}{k(i,\cdot)}$, where $k(i, \cdot)$ is the number of transitions out from state i. Standard likelihood asymptotic applies, despite the data are dependent, as $k(i, \cdot) \to \infty$, which will happen if the chain is ergodic. The asymptotic variance of the maximum likelihood estimates can be approximated as $\mathrm{var}(\hat{p}_{i,j}) \sim \hat{p}_{i,j}(1 - \hat{p}_{i,j})/k(i, \cdot)$. The covariances are zero, except $\mathrm{cov}(\hat{p}_{i,j}, \hat{p}_{i,j'}) \sim -\hat{p}_{i,j}\hat{p}_{i,j'}/k(i, \cdot)$ for $j \neq j'$. If the trajectory is short, the initial distribution should be considered. A possible model is to use the stationary distribution $\pi(s_0)$, which depend on the unknown transition probabilities. Hence numerical maximization is needed to obtain the maximum likelihood estimates. In certain medical applications, an alternative asymptotic regime can be of interest, when many (k) short trajectories are observed, and $k \to \infty$. In this case the initial distribution cannot be neglected.

Markov Chains and Markov Processes

Let $\{X_t : t \geq 0\}$ denote the (continuous time) Markov process on state space $(S, \mathcal{S})$ with transition matrix $P(t)$, i.e.,

$$(P(t))_{ij} = \mathbb{P}(X_{t+s} = j | X_s = i), \quad s \geq 0, \quad i, j \in S.$$

Under some mild regularity conditions is holds that the *generator matrix* Q, defined as

$$\left.\frac{d}{dt}\right|_{t=0} P(t) = Q,$$

exists for $P(t)$. The stationary distribution of a Markov process can be found as the unique probability π that solves $\pi Q = 0$, see Anderson (1991). A generator matrix Q is called uniformizable with rate μ if $\mu = \sup_j |q_{jj}| < \infty$. While any finite dimensional generator matrix is uniformizable a classical example of a Markov process on denumerable state space that fails to have this property is the M/M/∞

queue. Note that if Q is uniformizable with rate μ, then Q is uniformizable with rate η for any $\eta > \mu$. Let Q be uniformizable with rate μ and introduce the Markov chain P_μ as follows

$$[P_\mu]_{ij} = \begin{cases} q_{ij}/\mu & i \neq j \\ 1 + q_{ii}/\mu & i = j, \end{cases} \tag{2}$$

for $i, j \in S$, or, in shorthand notation,

$$P_\mu = I + \frac{1}{\mu} Q,$$

then it holds that

$$P(t) = e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\mu t)^n}{n!} (P_\mu)^n, \quad t \geq 0. \tag{3}$$

Moreover, the stationary distribution of P_μ and $P(t)$ coincide. The Markov chain $\mathcal{X}_\mu = \{X_n^\mu : n \geq 0\}$ with transition probability matrix P_μ is called the *sampled chain*. The relationship between $\mathcal{X}$ and $\mathcal{X}_\mu$ can be expressed as follows. Let $N_\mu(t)$ denote a Poisson process (see ►Poisson Processes) with rate μ, then $X^\mu_{N_\mu(t)}$ and X_t are equal in distribution for all $t \geq 0$. From the above it becomes clear that the analysis of the stationary behavior of a (uniformizable) continuous time Markov chain reduces to that of a discrete time Markov chain.

About the Authors

Arnoldo Frigessi is Professor in statistics, University of Oslo. He is director of the center for research based innovation Statistics for Innovation (sfi)2 and holds a position at the Norwegian Computing Center. Previously he hold positions at the University of Roma Tre and University of Venice. He is an Elected member of the Royal Norwegian Academy of Science and Letters. He is past scientific secretary of the Bernoulli Society for Mathematical Statistics and Probability. His research is mainly in the area of Bayesian statistics and MCMC, both methodological and applied.

Dr Bernd Heidergott is Associate Professor at the Department of Econometrics, Vrije Universiteit Amsterdam, the Netherlands. He is also research fellow at the Tinbergen Institute and at EURANDOM, both situated in the Netherlands. He has authored and co-authored more than 30 papers and two books, *Max-Plus linear Systems and Perturbation Analysis* (Springer, 2007), and *Max Plus at Work* (with Jacob van der Woude and Geert Jan Olsder, Princeton, 2006.)

Cross References

►Box–Jenkins Time Series Models
►Ergodic Theorem
►Graphical Markov Models
►Markov Processes
►Nonlinear Time Series Analysis
►Optimal Stopping Rules
►Record Statistics
►Statistical Inference for Stochastic Processes
►Stochastic Global Optimization
►Stochastic Modeling Analysis and Applications
►Stochastic Processes: Classification

References and Further Reading

Anderson W (1991) Continuous-time Markov chains: an applications oriented approach. Springer, New York
Feller W (1968) An Introduction to Probability Theory and its Applications, vol 1, 3rd edn. Wiley, New York
Gilks W, Richardson S, Spiegeihalter D (eds) (1995) Markov Chain Monte Carlo in practice. Chapman & Hall, London
Haeggstroem O (2002) Finite Markov chains and algorithmic applications, London Mathematical Society Student Texts (No. 52)
Iosifescu M (1980) Finite Markov processes and their applictions. Wiley, New York
Kemeny J, Snell J (1960) Finite Markov chains, (originally published by Van Nostrand Publishing Company Springer Verlag, 3rd printing, 1983)
Kijima M (1997) Markov processes for stochastic modelling. Chapman & Hall, London
Meyn S, Tweedie R (1993) Markov chains and stochastic stability. Springer, London
ummelin E (1984) General irreducible Markov chains and non-negative operators. Cambridge Univesity Press, Cambridge
Revuz D (1984) Markov chains, 2nd edn. Noth-Holland, Amsterdam
Seneta E (1973) Non-negative matrices and Markov chains (originally published by Allen & Unwin Ltd., London, Springer Series in Statistics, 2nd revised edition, 2006)

Markov Processes

Zoran R. Pop-Stojanović
Professor Emeritus
University of Florida, Gainesville, FL, USA

The class of Markov Processes is characterized by a special stochastic dependence known as the *Markov Dependence* that was introduced in 1907 by A.A. Markov while extending in a natural way the concept of stochastic independence that will preserve, for example, the asymptotic properties of sums of random variables such as the law of large numbers. One of his first applications of this dependence was in investigation of the way the vowels and consonants alternate in literary works in the Russian literature. This dependence that Markov introduced, dealt with what we

call today *a discrete-parameter Markov Chain with a finite number of states,* and it can be stated as follows: a sequence $\{X_n;\ n = 1, 2, \dots\}$ of real-valued random variables given on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, each taking on a finite number of values, satisfies

$$\mathrm{P}[X_{n+1} = x_{n+1} | X_1, X_2, \dots, X_n] = \mathrm{P}[X_{n+1} = x_{n+1} | X_n]. \quad (1)$$

Roughly speaking, (1) states that *any prediction of* X_{n+1} *knowing*

$$X_1, X_2, \dots, X_n,$$

can be achieved by using X_n *alone.*

This concept was further extended (as shown in what follows), for the *continuous-parameter Markov processes* by A.N. Kolmogorov in 1931. Further essential developments in the theory of continuous-parameter Markov Processes were due to W. Feller, J.L. Doob, G.A. Hunt, and E.B. Dynkin.

In order to introduce a continuous-parameter Markov Process, one needs the following setting. Let $\mathbf{T} \equiv [0, +\infty) \subset \mathbb{R}$ be the parameter set of the process, referred to in the sequel as *time*, where $\mathbb{R}$ denotes the one-dimensional Euclidean space; let $X = \{X_t,\ \mathcal{F}_t,\ t \in \mathbf{T}\}$ be the process given on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ that takes values in a topological space $(\mathcal{S}, \mathcal{E})$, where $\mathcal{E}$ is a Borel field of $\mathcal{S}$, that is, a σ-field generated by open sets in $\mathcal{S}$. The process X is adapted to the increasing family $\{\mathcal{F}_t,\ t \in \mathbf{T}\}$ of σ-fields of $\mathcal{F}$, where $\mathcal{F}_0$ contains all P-null sets. All X_t's are $\mathcal{E}$-measurable. Here, X_t *is adapted to* $\mathcal{F}_t$ *means that all random events related to* X_t *are contained in* $\mathcal{F}_t$ *for every value* t *of the parameter of the process, that is,* X_t *is* $\mathcal{F}_t$*-measurable in addition of being* $\mathcal{E}$*-measurable.* In order to describe the Markov dependence for the process X, the following two σ-fields are needed: $\forall t,\ t \in \mathbf{T},\ \mathcal{F}_t^{past} = \sigma(\{X_s,\ s \in [0, t]\})$ and $\mathcal{F}_t^{future} = \sigma(\{X_s,\ s \in [t, +\infty)\})$. Here, the *past* and the *future* are relative to the instant t that is considered as the *present.* Now the process $X = \{X_t, \mathcal{F}_t,\ t \in \mathbf{T}\}$ is called a *Markov Process* if and only if one of the following equivalent conditions is satisfied:

$$\begin{aligned}
&(i)\ \forall t,\ t \in \mathbf{T},\ A \in \mathcal{F}_t,\ B \in \mathcal{F}_t^{future}:\\
&\quad \mathrm{P}(A \cap B | X_t) = \mathrm{P}(A | X_t)\mathrm{P}(B | X_t).\\
&(ii)\ \forall t,\ t \in \mathbf{T},\ B \in \mathcal{F}_t^{future}:\\
&\quad \mathrm{P}(B | \mathcal{F}_t) = \mathrm{P}(B | X_t).\\
&(iii)\ \forall t,\ t \in \mathbf{T},\ A \in \mathcal{F}_t:\\
&\quad \mathrm{P}\left(A | \mathcal{F}_t^{future}\right) = \mathrm{P}(A | X_t).
\end{aligned} \quad (2)$$

Observe that (ii) in (2) is the analog of (1) stating that *the probability of an event in the future of the Markov process* X *depends only on the probability of the present state of the process and it is independent of the past history of the process.* There are numerous phenomena occurring in physical sciences, social sciences, econometrics, the world of finance, to name just a few, that can all be modelled by Markov processes. Among Markov processes there is a very important subclass of the so-called *strong Markov processes.* This proper subclass of Markov processes is obtained by *randomizing* the parameter of the process. This randomization of the parameter leads to the so-called *optional times of the process* and the Markov property (2) is replaced by the *strong Markov property*, where in (2) deterministic time t is replaced by an *optional time* of the process. The most important example of a strong Markov process is the *Brownian Motion Process* (see ►Brownian Motion and Diffusions) that models the physical phenomenon known as the *Brownian Movement of particles.* Another important class of processes – *Diffusion processes,* are *strong Markov Processes with continuous paths.*

One of the most important properties of Markov processes is *that times between transitions from one state to another, are random variables that are conditionally independent of each other given the successive states being visited, and each such sojourn time has an exponential distribution with the parameter dependent on the state being visited.* This property coupled with the property that successive states visited by the process form a Markov chain (see ►Markov Chains), clearly describe the structure of a Markov process. Other important examples of Markov processes are ►Poisson processes, Compound Poisson processes, The ►Random Walk, Birth and Death processes, to mention just a few. The last mentioned class of Markov processes has many applications in biology, ►demography, and ►queueing theory.

For further details and proofs of all facts mentioned here, a reader may consult the enclosed list of references.

Cross References

►Brownian Motion and Diffusions
►Markov Chains
►Martingale Central Limit Theorem
►Optimal Stopping Rules
►Poisson Processes
►Random Permutations and Partition Models
►Random Walk
►Statistical Inference for Stochastic Processes
►Stochastic Processes
►Stochastic Processes: Classification
►Structural Time Series Models

References and Further Reading

Blumenthal RM, Getoor RK (1968) Markov processes and potential theory. Academic Press, New York
Chung KL (1982) Lectures from Markov processes to Brownian motion. Springer, New York
Çinlar E (1975) Introduction to stochastic processes. Prentice Hall, New Jersey
Doob JL (1953) Stochastic processes. Wiley, New York
Dynkin EB (1965) Markov process, 2 Volumes. Springer, New York
Feller W (1971) An introduction to probability theory and its applications, vol 2. Wiley, New York

Martingale Central Limit Theorem

Petra Posedel
Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

The martingale central limit theorem (MCLT) links the notions of martingales and the Lindeberg–Feller classical central limit theorem (CLT, see ▶Central Limit Theorems) for independent summands.

Perhaps the greatest achievement of modern probability is the unified theory of limit results for sums of independent random variables, such as the law of large numbers, the central limit theorem, and the law of the iterated logarithm. In comparison to the classical strong law of large numbers, the classical CLT says something also about the rate of this convergence. We recall the CLT for the case of independent, but not necessarily identically distributed random variables. Suppose that $\{X_i,\ i \geq 1\}$ is a sequence of zero-mean independent random variables such that $Var[X_n] = \sigma_n^2 < \infty,\ n \geq 1$. Let $S_n = \sum_{i=1}^n X_i,\ n \geq 1$ and set $Var[S_n] = s_n^2$. If the Lindeberg condition holds, i.e., $\frac{\sum_{i=1}^n E\left[X_i \mathbb{1}_{\{|X_i| \geq \epsilon s_n\}}\right]}{s_n^2} \longrightarrow 0$ as $n \to \infty$, for all $\epsilon > 0$, and $\mathbb{1}_{\{.\}}$ denoting the indicator function, then $\frac{S_n}{s_n} \xrightarrow{\mathcal{D}} N(0,1)$, where $N(0,1)$ denotes the standard normal random variable.

Limit theorems have applicability far beyond the corresponding results for sums of independent random variables. Namely, since sums of independent random variables centered at their expectations have a specific dependence structure (i.e., are martingales), there is interest in extending the results to sums of dependent random variables.

In order to define martingales and state the MCLT attributed to Brown (1971), one needs the following setting. Let $(\Omega,\ \mathcal{F},\ \mathbb{P})$ be a probability space and let $\{\mathcal{F}_n,\ n \geq 0\}$ be an increasing sequence of σ-fields of $\mathcal{F}$ sets.

Definition 1 *A sequence $\{Y_n,\ n \geq 0\}$ of random variables on Ω is said to be a martingale with respect to $\{\mathcal{F}_n,\ n \geq 0\}$ if (1) Y_n is measurable with respect to $\mathcal{F}_n$, (2) $E|Y_n| < \infty$, and (3) $E[Y_n|\mathcal{F}_m] = Y_m$ a.s. for all $m < n,\ m,\ n \geq 0$.*

In order to highlight the dependence structure of the underlying random variables, one should note that condition (3) is weaker than independence since it cannot be deduced which structure conditional higher-order moments may have given the past. The mathematical theory of martingales may be regarded as an extension of the independence theory, and it too has its origins in limit results, beginning with Bernstein (1927) and Lévy's (1935) early central limit theorems. These authors introduced the martingale in the form of consecutive sums with a view to generalizing limit results for sums of independent random variables. However, it was the subsequent work of Doob, including the proof of the celebrated martingale convergence theorem, that completely changed the direction of the subject, and his book (Doob 1953), popularly called in academia the *Holy Bible for stochastic processes*, has remained a major influence for nearly three decades.

The main result that follows applies the CLT to sequences of random variables that are martingales. If $\{S_n,\ \mathcal{F}_n\}$ is a martingale, it seems natural to replace $Var[S_n]$ in the CLT by the sum of conditional variances. Secondly, the norming by $1/n$ is very restrictive. For a sequence of independent, but not identically distributed random variables, it seems appropriate to norm by a different constant, and for a sequence of dependent random variables norming by another random variable should be considered. The limit theory for martingales essentially covers that for the categories of processes with independent increments and ▶Markov processes. Using stochastic processes that are martingales for analyzing limit results, one has at their disposal all the machinery from martingale theory. This reason makes martingales considerably attractive for inference purposes. A standard reference on martingales is Williams (1991).

Theorem 1 *Let $\{S_n,\ \mathcal{F}_n,\ n \geq 1\}$ be a zero-mean martingale with $S_0 = 0$, whose increments have finite variance. Write*

$$S_n = \sum_{i=1}^n X_i, \quad V_n^2 = \sum_{i=1}^n E\left[X_i^2 | \mathcal{F}_{i-1}\right], \quad \text{and}$$
$$s_n^2 = E\left[V_n^2\right] = E\left[S_n^2\right]. \tag{1}$$

If

$$\frac{V_n^2}{s_n^2} \xrightarrow{\mathbb{P}} 1 \quad \textit{and} \quad \frac{\sum_{i=1}^{n} E\left[X_i^2 \mathbb{1}_{\{|X_i| \geq \epsilon s_n\}}\right]}{s_n^2} \xrightarrow{\mathbb{P}} 0 \qquad (2)$$

as $n \to \infty$, *for all* $\epsilon > 0$, *and* $\mathbb{1}_{\{.\}}$ *denoting the indicator function, then*

$$\frac{S_n}{s_n} \xrightarrow{\mathcal{D}} N(0,1), \qquad (3)$$

where $N(0,1)$ *denotes the standard normal random variable.*

Roughly speaking, (3) says that the sum of martingale differences, when scaled appropriately, is approximately normally distributed provided the conditional variances are sufficiently well behaved. The theorem seems relevant in any context in which conditional expectations, given the past, have a simple and possibly explicit form. Various results on sums of independent random variables in fact require only orthogonality of the increments, i.e., $E[X_i X_j] = 0$, $i \neq j$, and this property holds for martingales whose increments have finite variance. The MCLT reduces to the sufficiency part of the standard Lindeberg–Feller result in the case of independent random variables.

The interpretation of V_n^2 is highlighted and particularly interesting for inference purposes. Let $X_1, X_2, \ldots$ be a sequence of observations of a stochastic process whose distribution depends on a (single) parameter θ, and let $L_n(\theta)$ be the likelihood function associated with $X_1, X_2, \ldots$. Under very mild conditions, score functions $S_n = \partial \log L_n(\theta)/\partial\theta$ form a martingale whose conditional variance $V_n^2 = I_n(\theta)$ is a generalized form of the standard Fisher information, as shown in Hall and Heyde (1980). Namely, suppose that the likelihood function $L(\theta)$ is differentiable with respect to θ and that $E_\theta[\partial \log L(\theta)/\partial\theta]^2 < \infty$.
Let θ be a true parameter vector. We have

$$S_n = \frac{\partial \log L_n(\theta)}{\partial\theta} = \sum_{i=1}^{n} x_i(\theta),$$

$$x_i(\theta) = \frac{\partial}{\partial\theta}[\log L_i(\theta) - \log L_{i-1}(\theta)],$$

and thus $E_\theta[x_i(\theta)|\mathcal{F}_{i-1}] = 0$ a.s., so that $\{S_n, \mathcal{F}_n, n \geq 1\}$ is a square-integrable martingale. Set $V_n^2 = \sum_{i=1}^{n} E_\theta\left[x_i^2(\theta)\middle|\mathcal{F}_{i-1}\right]$. The quantity V_n^2 reduces to the standard Fisher information $I_n(\theta)$ in the case where the observations $\{X_i, i \geq 1\}$ are independent random variables. If the behavior of V_n^2 is very erratic, then so is that of S_n, and it may not be possible to obtain a CLT.

So, if we have a reasonably large sample, we can assume that estimators obtained from estimating functions that are martingales, have an approximately normal distribution, which can be used for testing and constructing confidence intervals. A standard reference for the more general theory of martingale estimating functions is Sørensen (1999).

Billingsley (1961), and independently Ibragimov (1963), proved the central limit theorem for martingales with stationary and ergodic differences. For such martingales the conditional variance V_n^2 is asymptotically constant, i.e., $\frac{V_n^2}{s_n^2} \xrightarrow{P} 1$. Brown (1971) showed that the first part of condition (2) and not stationarity or ergodicity is crucial for such a result to hold. Further extensions in view of other central limit theorems for double arrays are based on Dvoretzky (1970) and McLeish (1974), where limit results employ a double sequence schema $\{X_{n,j}, 1 \leq j \leq k_n < \infty, n \geq 1\}$ and furnish conditions for the row sums $S_n = \sum_{j=1}^{k_n} X_{n,j}$ to converge in distributions to a mixture of normal distributions with means zero. A large variety of negligibility assumptions have been made about differences $X_{n,j}$ during the formulation of martingale central limit theorems. The classic condition of negligibility in the theory of sums of independent random variables asks the $X_{n,j}$ to be uniformly asymptotically negligible.

A comprehensive review on mainly one-dimensional martingales can be found in Helland (1982). Multivariate versions of the central limit theorem for martingales satisfying different conditions or applicable to different frameworks, can be found in Hutton and Nelson (1984), Sørensen (1991), Küchler and Sørensen (1999), Crimaldi and Pratelli (2005), and Hubalek and Posedel (2007).

Cross References

►Central Limit Theorems
►Markov Processes
►Martingales
►Statistical Inference for Stochastic Processes

References and Further Reading

Bernstein S (1927) Sur l'extension du théorèmelimite du calcul des probabilitiés aux sommes de quantités dépendantes. Math Ann 85:1–59
Billingsley P (1961) The Lindeberg-Lévy theorem for martingales. Proc Am Math Soc 12:788–792
Brown BM (1971) Martingale central limit theorems. Ann Math Stat 42:59–66
Chow YS, Teicher H (1997) Probability theory, 3rd edn. Springer, New York
Crimaldi I, Pratelli L (2005) Convergence results for multivariate martingales. Stoch Proc Appl 115(4):571–577
Doob JL (1953) Stochastic processes. Wiley, New York

Dvoretzky A (1970) Asymptotic normality for sums of dependent random variables. Proceedings of the Sixth Berkeley Symposium on Statistics and Probability. pp 513–535
Hall P, Heyde CC (1980) Martingale limit theory and its application. Academic, New York
Helland IS (1982) Central limit theorems for martingales with discrete or continuous time. Scand J Stat 9:79–94
Hubalek F, Posedel P (2007) Asymptotic analysis for a simple explicit estimator in Barndorff-Nielsen and Shephard stochastic volatility models. Thiele Research Report 2007–2005
Hutton JE, Nelson PI (1984) A mixing and stable central limit theorem for continuous time martingales. Technical Report No. 42, Kansas State University, Kansas
Ibragimov IA (1963) A central limit theorem for a class of dependent random variables. Theor Probab Appl 8:83–89
Küchler U, Sørensen M (1999) A note on limit theorems for multivariate martingales. Bernoulli 5(3):483–493
Lévy P (1935) Propriétés asymptotiques des sommes de variables aléatoires enchainées. Bull Sci Math 59(series 2):84–96, 109–128
McLeish DL (1974) Dependent Central Limit Theorems and invariance principles. Ann Probab 2:620–628
Sørensen M (1991) Likelihood methods for diffusions with jumps. In: Prabhu NU, Basawa IV (eds) Statistical inference in stochastic processes. Marcel Dekker, New York, pp 67–105
Sørensen M (1999) On asymptotics of estimating functions. Brazilian J Probab Stat 13:111–136
Williams D (1991) Probability with martingales. Cambridge University Press, Cambridge

Martingales

Rüdiger Kiesel
Professor, Chair for energy trading and financial services
Universität Duisburg-Essen, Duisburg, Germany

The fundamental theorem of asset pricing (The term *fundamental theorem of asset pricing* was introduced in Dybvig and Ross [1987]. It is used for theorems establishing the equivalence of an economic modeling condition such as no-arbitrage to the existence of the mathematical modeling condition existence of equivalent martingale measures.) links the martingale property of (discounted) asset price processes under a particular class of probability measures to the 'fairness' (in this context no arbitrage condition) of financial markets. In elementary models one such result is *In an arbitrage-free complete financial market model, there exists a unique equivalent martingale measure*, see e.g., Bingham and Kiesel (2004).

So despite martingales have been around for more than three and a half centuries they are still at the forefront of applied mathematics and have not lost their original motivation of describing the notion of fairness in games of chance. The *Oxford English Dictionary* lists under the word *martingale* (we refer to Mansuy [2009] for a interesting account of the etymology of the word): A system of gambling which consists in doubling the stake when losing in order to recoup oneself (1815).

Indeed, the archetype of a martingale is the capital of a player during a fair gambling game, where the capital stays "constant on average"; a supermartingale is "decreasing on average," and models an unfavourable game; a submartingale is "increasing on average," and models a favorable game.

Gambling games have been studied since time immemorial – indeed, the Pascal–Fermat correspondence of 1654 which started the subject was on a problem (de Méré's problem) related to gambling. The doubling strategy above has been known at least since 1815. The term "martingale" in our sense is due to J. Ville (1910–1989) in his thesis in 1939. Martingales were studied by Paul Lévy (1886–1971) from 1934 on (see obituary Loève (1973)) and by J.L. Doob (1910–2004) from 1940 on. The first systematic exposition was Doob (1953). Nowadays many very readable accounts exist, see Neveu (1975), Williams (1991) and Williams (2001).

Martingales are of central importance in any modelling framework which uses ▶stochastic processes, be it in discrete or continuous time. The concept has been central to the theory of stochastic processes, stochastic analysis, in mathematical statistics, information theory, and in parts of mathematical physics, see Kallenberg (1997) and Meyer (2009) for further details. The Martingale gambling insight 'You can't beat the system' establishes properties of martingale transforms and lays the foundation of stochastic integrals, Øksendal (1998). Martingale stopping results establish optimality criteria which help develop optimal strategies for decision problems (and exercising financial options), see Chow (1971) and Shiryaev (2007).

We can here only give a few fundamental definitions and results and point to the vast literature for many more exiting results.

For the definition, let I be a suitable (discrete or continuous) index set and assume that an index t is always taken from I. Given a stochastic basis $(\Omega, \mathcal{F}, I\!P, I\!F = \{\mathcal{F}_t\})$ (where the filtration $I\!F$ models the flow of information) we call a process $X = (X_t)$ a *martingale* relative to $(\{\mathcal{F}_t\}, I\!P)$ if

(i) X is adapted (to $\{\mathcal{F}_t\}$).
(ii) $I\!E|X_t| < \infty$ for all t.
(iii) For $s \le t$ we have $I\!E[X_t|\mathcal{F}_s] = X_s \quad I\!P - a.s.$.

X is a *supermartingale* if in place of (ii)

$$\mathbb{E}[X_t|\mathcal{F}_s] \le X_s \quad \mathbb{P}-a.s.;$$

X is a *submartingale* if in place of (iii)

$$\mathbb{E}[X_t|\mathcal{F}_s] \ge X_s \quad \mathbb{P}-a.s..$$

Basic examples are the mean-zero ►random walk: $S_n = \sum X_i$, with X_i independent, where for $\mathbb{E}(X_i) = 0$ S_n is a martingale (submartingales: positive mean; supermartingale: negative mean) and stock prices: $S_n = S_0\zeta_1\cdots\zeta_n$ with ζ_i independent positive r.vs with existing first moment. (See Williams (1991) and Williams (2001) for many more examples). In continuous time the central example is that of Brownian motion, see Revuz and Yor (1991), Karatzas and Shreve (1991), which of course is a central process for many branches of probability (see also ►Brownian Motion and Diffusions).

Now think of a gambling game, or series of speculative investments, in discrete time. There is no play at time 0; there are plays at times $n = 1, 2, \ldots$, and

$$\Delta X_n := X_n - X_{n-1}$$

represents our net winnings per unit stake at play n. Thus if X_n is a martingale, the game is "fair on average."

Call a process $C = (C_n)_{n=1}^{\infty}$ *predictable* if C_n is $\mathcal{F}_{n-1}$-measurable for all $n \ge 1$. Think of C_n as your stake on play n (C_0 is not defined, as there is no play at time 0). Predictability says that you have to decide how much to stake on play n based on the history *before* time n (i.e., up to and including play $n-1$). Your winnings on game n are $C_n\Delta X_n = C_n(X_n - X_{n-1})$. Your total (net) winnings up to time n are

$$Y_n = \sum_{k=1}^{n} C_k\Delta X_k = \sum_{k=1}^{n} C_k(X_k - X_{k-1}).$$

This constitutes the *Martingale transform* of X by C.

The central theorem for betting and applications in finance says that "You can't beat the system!," i.e., if X is a martingale then the martingale transform is a martingale (under some mild regularity conditions on C). So in the martingale case, predictability of C means we can't foresee the future (which is realistic and fair). So we expect to gain nothing – as we should, see e.g., Neveu (1975). Likewise one can analyze different strategies to stop the game, then Doob's stopping time principle reassures that it is not possible to beat the system, see e.g., Williams (2001).

Martingale transforms were introduced and studied by Burkholder (1966). They are the discrete analogs of stochastic integrals and dominate the mathematical theory of finance in discrete time, see Shreve (2004), just as stochastic integrals dominate the theory in continuous time, see Harrison and Pliska (1981). The various links between mathematical finance and martingale theory are discussed in Musiela and Rutkowski (2004) and Karatzas and Shreve (1998).

Martingale-convergence results are among the most important results in probability (arguably in mathematics). Hall and Heyde (1980) and Chow (1988) are excellent sources, but Doob (1953) lays the foundations. Martingale techniques play a central role in many parts of probability, consult Rogers (1994), Revuz and Yor (1991), Karatzas and Shreve (1991) or Kallenberg (1997) for excellent accounts. Martingales appear in time series theory and sequential analysis, see Lai (2009) and Hamilton (1994).

About the Author

Rüdiger Kiesel holds the chair of energy trading and financial services (sponsored by the Stifterverband für die Deutsche Wissenschaft and RWE Supply & Trading; the first such chair in Europe). Previously, he was Professor and Head of the Institute of Financial Mathematics at Ulm University. Kiesel also holds guest professorships at the London School of Economics and the Centre of Mathematical Applications at the University of Oslo. His main research areas are currently risk management for power utility companies, design and analysis of credit risk models, valuation and hedging of derivatives (interest-rate, credit- and energy-related), methods of risk transfer and structuring of risk (securitization), and the stochastic modelling of financial markets using Lévy-type processes. He is on the editorial board of the *Journal of Energy Markets* and co-author (with Nicholas H. Bingham) of the Springer Finance monograph *Risk-Neutral Valuation: Pricing and Hedging of Financial Derivatives* (2nd edition, 2004).

Cross References

►Brownian Motion and Diffusions
►Central Limit Theorems
►Khmaladze Transformation
►Martingale Central Limit Theorem
►Point Processes
►Radon–Nikodým Theorem
►Statistical Inference for Stochastic Processes
►Statistics and Gambling
►Stochastic Processes
►Stochastic Processes: Applications in Finance and Insurance
►Stochastic Processes: Classification

References and Further Reading

Bingham N, Kiesel R (2004) Risk-Neutral valuation: pricing and hedging of financial derivatives, 2nd edn. Springer, London
Burkholder DL (1966) Martingale transforms. Ann Math Stat 37:1494–1504

Chow YS, Teicher H (1988) Probability theory: independence, interchangeability, martingales, 2nd edn. Springer, New York
Chow YS, Robbins H, Siegmaund D (1971) Great expectations: the theory of optimal stopping. Houghton Mifflin, Boston
Doob JL (1953) Stochastic processes. Wiley, New York
Dybvig PH, Ross SA (1987) Arbitrage. In: Milgate M, Eatwell J, Newman P (eds) The new palgrave: dictionary of economics. Macmillan, London
Hall P, Heyde CC (1980) Martingale limit theory and applications. Academic, New York
Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
Harrison JM, Pliska SR (1981) Martingales and stochastic integrals in the theory of continuous trading. Stoch Proc Appl 11: 215–260
Kallenberg O (1997) Foundations of probability. Springer, New York
Karatzas I, Shreve S (1991) Brownian motion and stochastic calculus, 2nd edn, 1st edn 1988. Springer, Berlin
Karatzas I, Shreve S (1998) Methods of mathematical finance. Springer, New York
Lai TL (2009) Martingales in sequential analysis and time series, 1945–1985. Electron J Hist Probab Stat 5
Loève M (1973) Paul Lévy (1886–1971), obituary. Ann Probab 1:1–18
Mansuy R (2009) The origins of the word 'martingale'. Electron J Hist Probab Stat 5
Meyer P-A (2009) Stochastic processes from 1950 to the present. Electron J Hist Probab Stat 5
Musiela M, Rutkowski M (2004) Martingale methods in financial modelling, 2nd edn. Springer, Heidelberg
Neveu J (1975) Discrete-parameter martingales. North-Holland, Amsterdam
Øksendal B (1998) Stochastic differential equations: an introduction with applications, 5th edn. Springer, Berlin
Revuz D, Yor M (1991) Continuous martingales and Brownian motion. Springer, New York
Rogers L, Williams D (1994) Diffusions, Markov processes and martingales. Volume 1: foundations, 2nd edn. Wiley, Chichester
Shiryaev AN (2007) Optimal stopping rules, 3rd edn. Springer, Berlin
Shreve S (2004) Stochastic calculus for finance I: the binomial asset pricing model. Springer, New York
Williams D (1991) Probability with martingales. Cambridge University Press, Cambridge
Williams D (2001) Weighing the odds. Cambridge University Press, Cambridge

Mathematical and Statistical Modeling of Global Warming

Chris P. Tsokos
Distinguished University Professor
University of South Florida, Tampa, FL, USA

Introduction

Do we scientifically understand the concept of "Global Warming"? A very basic definition of "Global Warming" is an increase in temperature at the surface of the earth supposedly caused by the greenhouse effect, carbon dioxide, CO_2 (greenhouse gas). The online encyclopedia, Wikipedia, defines the phenomenon of "GLOBAL WARMING" as the increase in the average temperature of the earth's near surface air and oceans in the recent decades and its projected continuation.

For the past 3 years this has been a media chaos: pro and concerned skeptics. The Intergovernmental Panel of the United States on Climate Change (IPCC) – "Climate Change 2007" claimed that the following are some of the causes of Global Warming:

- Increase in temperature – Increase in sea level
- Unpredictable pattern in rainfall
- Increase in extreme weather events
- Increase in river flows
- Etc.

Furthermore, the award winning documentary narrated by Vice President Gore strongly supports the IPCC findings. However, the ABC news program 20/20 "Give Me a Break," raises several questions and disputes the process by which IPCC stated their findings. A number of professional organizations, the American Meteorological Society, American Geographical Union, AAAS, supported the subject matter. The U.S. National Academics blame global warming on human activities.

The concerned skeptics raise several points of interest concerning Global Warming. Great Britain's Channel 4 Documentary entitled "*The Great Global Warming Swindle*" disputes several of the aspects of Vice President former documentary. NASA scientists reveal through their scientific experiments and studies that the increase in atmospheric temperature is due to the fact that sea spots are hotter than previously thought. Their findings are also reported by the *Danish National Space Center*, DNSC, on similar investigations conducted by NASA. DNSC stated that there is absolutely nothing we can do to correct this situation. *Times Washington Bureau Chief*, Bill Adair, states that "Global Warming has been called the most dire issue facing the planet and yet, if you are not a scientist, it can be difficult to sort out the truth." The Wall Street Journal in a leading article "Global Warming is 300-year-old News," stated that "the various kind of evidence examined by the *National Research Council*, NRC, led it to conclude that the observed disparity between the surface and atmospheric temperature trends during the 20-year period is probably at least partially real." It further stated that "uncertainties in all aspects exist- cannot draw any conclusion concerning *Global Warming*." However, the NRC study concluded with an important statement that "major advances in scientific

methods will be necessary before these questions on *Global Warming* can be resolved."

Furthermore, the temperature increase that we are experiencing are infinitesimal, during the past 100 years – the mean global surface air temperature increased by approximately $1.3^{\circ}F$ $(0.32^{\circ}F)$. Dr. Thomas G. Moore, Senior Fellow at the Hoover Institute at Stanford University, in his article entitled "Climate of Fear: Why We Shouldn't Worry About Global Warming" is not concerned with such small changes in temperatures. Furthermore, in his interview with *Newsweek*, he said more people die from cold than from warmth and an increase of a few degrees could prevent thousands of deaths.

It is well known that carbon dioxide, CO_2, and surface/atmospheric temperatures are the primary cause of "GLOBAL WARMING." Jim Verhult, Perspective Editor, *St. Petersburg Times*, writes, "carbon dioxide is invisible – no color, no odor, no taste. It puts out fires, puts the fizz in seltzer and it is to plants what oxygen is to us. It's hard to think of it as a poison." The U.S.A. is emitting approximately 5.91221 billion metric tons of CO_2 in the atmosphere, which makes us the world leader; however, by the end of 2007, the Republic of China became the new leader. Temperatures and CO_2 are related in that as CO_2 emissions increase, the gasses start to absorb too much sunlight and this interaction warms up the globe. Thus, the rise in temperature and the debate of "GLOBAL WARMING."

While working on the subject matter, an article appeared on the front page of the *St. Petersburg Times* on January 23, 2007. This article, entitled "Global Warming: Meet your New Adversary," was written by David Adams. The highlight of this article was a section called "By the Numbers," which stated some information concerning the continental United States: 2006 hottest year; U.S. top global warming polluter; 20% increase of CO_2 since 1990; 15% of CO_2 emissions by 2020; 78 number of days U.S. fire season has increased; and 200 million people that will be displaced due to global warming. Our data for the continental U.S. does not support the first four statistics, we have no data for the fifth, and the sixth is quite hypothetical. The final assertion, with "0" representing the number of federal bills passed by the Congress to cap America's global warming pollution. Thus, it is very important that we perform sophisticated statistical analysis and modeling to fully understand the subject matter. Also, very recently, the Supreme Court of the U.S., in one of its most important environmental decisions, ruled that the Environmental Protection Agency (EPA) has the authority to regulate the greenhouse gases that contribute to global climate changes unless it can provide a scientific basis for its refusal.

We believe that a contributing factor in creating these controversies among scientists (and this is passed onto the policymakers and the media) is a lack of precise and accurate statistical analysis and modeling of historical data with an appropriate degree of confidence. The problem of "GLOBAL WARMING" is very complex with a very large number of contributing entities with significant interactions. The complexity of the subject matter can be seen in the attached diagram "A Schematic View" (Fig. 1). We believe that statisticians/mathematicians can help to create a better understanding of the subject problem that hopefully will lead to the formulation of legislative policies.

Thus, to scientifically make an effort to understand "Global Warming," we must study the marriage of CO_2 and atmosphere temperature, individually and together, using available historical data. Here we shall briefly present some parametric statistical analysis, forecasting models for CO_2 and atmospheric temperature, T_a along with a differential equation, that give the rate of change of CO_2 as a function of time. Scientists can utilize these preliminary analysis and models to further the study of Global Warming. Additional information can be found in Tsokos (2007a, b), and Tsokos 2008b.

Atmospheric Temperature, T_a

Here we shall utilize historical temperature data recorded in the Continental United States from 1895 to 2007, to parametrically identify the probability density of the subject data and to develop a forecasting model to predict short and long term values of T_a.

The probability density function, pdf, of T_a is the three-parameter lognormal pdf. It is given by

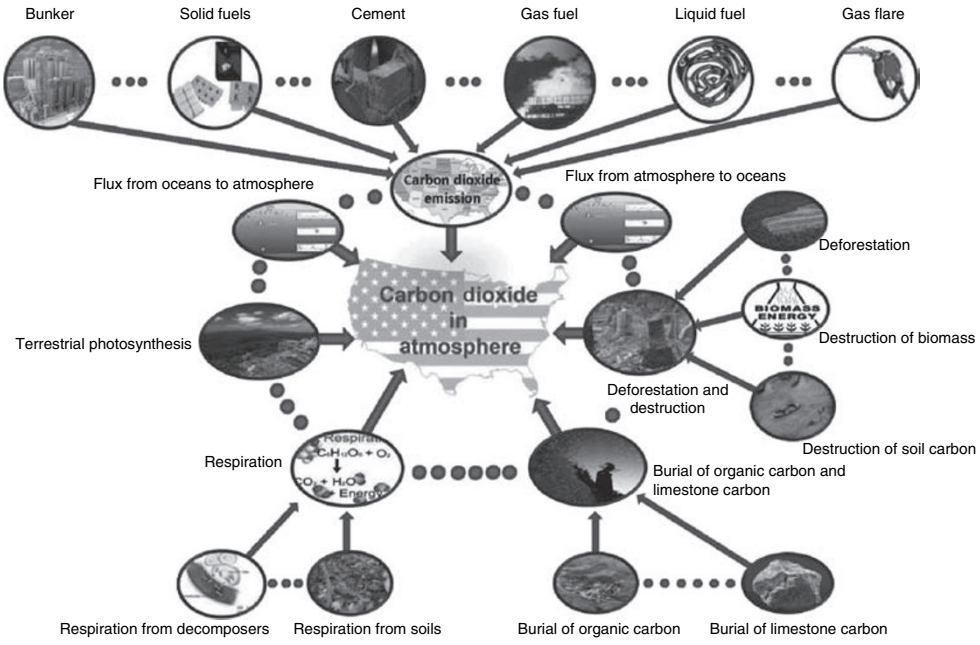
$$f(t;\mu,\theta,\sigma) = \frac{\exp\left\{-\frac{1}{2}\left[\ln(t-\theta)-\mu\right]^2\right\}}{(t-\theta)\sigma\sqrt{2\pi}},\quad t \geq \theta, \sigma, \mu > 0, \tag{1}$$

where μ, σ and θ, are the scale, shape and location parameters, respectively.

For the given T_a data the maximum likelihood estimation of population parameter, μ, σ and θ are $\hat{\mu} = 3.59, \hat{\sigma} = 0.019$ and $\hat{\theta} = 0.195$. Thus, the actual pdf that we will be working with is given by

$$f(t;\hat{\mu},\hat{\theta},\hat{\sigma}) = \frac{\exp\left\{-\frac{1}{2}\left[\ln(t-0.195)-2.59\right]^2\right\}}{(t-0.195)\cdot 0.019\sqrt{2\pi}},\quad t \geq 0.195. \tag{2}$$

Having identified the pdf that probabilistically characterizes the behavior of the atmospheric T_a, we can obtain the expected value of T_a, all the useful basic statistics along with being able to obtain confidence limits on the true T_a.



Mathematical and Statistical Modeling of Global Warming. Fig. 1 Carbon dioxide (CO_2) in the atmosphere in USA "A Schematic View"

Such a pdf should be applicable in other countries around the world.

The subject data, T_a, is actually a stochastic realization and is given as nonstationary time series. The development of the multiplicative seasonal autoregressive integrated moving average, ARIMA model is defined by

$$\Phi_p(B^s)\phi(1-B)^d(1-B^s)^D x_t = \theta_q(B)\Gamma_Q(B^s)\varepsilon_t, \qquad (3)$$

where p is the order of the autoregressive process; d is the order of regular differencing; q is the order of the moving average process; P is the order of the seasonal autoregressive process; D is the order of the seasonal differencing; Q is the order of the seasonably moving average process; and s refers to the seasonal period, and

$$\begin{aligned}
\phi_p(B) &= (1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)\\
\theta_q(B) &= (1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q)\\
\Phi_P(B^s) &= 1-\Phi_1 B^s-\Phi_2 B^{2s}-\cdots-\Phi_P B^{Ps}\\
\Gamma_Q(B^s) &= 1-\Gamma_1 B^s-\Gamma_2 B^{2s}-\cdots-\Gamma_Q B^{Qs}.
\end{aligned}$$

The developing process of (3) using the actual data is complicated and here we present the final useful form of the model. The reader is referred to Shih and Tsokos (2007, 2009) for details.

The estimated forecasting model for the atmospheric data is given by

$$\begin{aligned}
\hat{x}_t = {} & 1.0941x_{t-1} - 0.057x_{t-2} - 0.0371x_{t-3} + 0.9954x_{t-12}\\
& - 1.0891x_{t-13} + 0.0567x_{t-14} + 0.0369x_{t-15}\\
& + 0.0046x_{t-24} + 0.0895x_{t-25} - 0.0004x_{t-26}\\
& + 0.0017x_{t-27} - 0.9861\varepsilon_{t-1} - 0.9742\Gamma_1\varepsilon_{t-12}\\
& + 0.9607\varepsilon_{t-13}. \qquad (4)
\end{aligned}$$

The mean of the residuals, $\bar{r}$, the variance, S_r^2, the standard deviation, S_r, standard error, SE, and the mean square error, MSE, are presented below for one unit of time ahead forecasting.

$\bar{r}$	S_r^2	S_r	SE	MSE
−0.008512476	4.331902	2.081322	0.05673052	4.328756

These numerical results give an indication of the quality of the developed model.

Carbon Dioxide, CO_2

Parametric Analysis

The other most important entity in Global Warming is CO_2. The complexity of CO_2 in the atmosphere is illustrated by the schematic diagram that was introduced. To

better understand CO_2, we need to probabilistically determine the best probability distribution, pdf, that characterizes its behavior. Presently, scientists working on the subject matter make the assumption that CO_2 in the atmosphere follows the classical Gaussian pdf and that is not the best possible fit of the actual data and could lead to misleading decisions. The actual data that we are using was collected in the Island of Hawaii/Mauna Loa from 1990 to 2004. Through goodness-of-fit statistical testing, the best fit of the CO_2 data that we can study its behavior probabilistically is the three-parameter Weibull pdf. The cumulative three-parameter Weibull probability distribution is given by

$$F(x) = 1 - \exp\left\{-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right\},\ \gamma \le x < \infty, \delta > 0, \beta > 0 \quad (5)$$

where α, β, and γ are the shape, scale, and location parameter. The nth moment, mean and variance are given by

$$m_n = \beta^n \Gamma\left(1+\frac{n}{\alpha}\right),\ \mu = \beta\Gamma\left(1+\frac{1}{\alpha}\right) \text{ and } \sigma^2 = \beta^2\Gamma\left(1+\frac{2}{\alpha}\right) - \mu,$$

respectively, where Γ is the gamma function. The approximate maximum likelihood estimates of the true parameters, α, β and γ for the Hawaii data are given by

$$\hat{\alpha} = 2.108,\ \hat{\beta} = 17.092,\ \text{and } \hat{\gamma} = 349.6.$$

Thus, the cumulative pdf that we can use to probabilistically characterize the CO_2 behavior and answer related questions is given by:

$$F(x) = 1 - \exp\left\{-\left(\frac{x-349.6}{17.092}\right)^{2.108}\right\}. \quad (6)$$

For additional details of the subject area see Shih and Tsokos (2009).

Forecasting Model of CO_2

Here we present a forecasting model of CO_2 in the atmosphere. Having such a model will allow us to accurately predict the amount of CO_2 in the atmosphere, and make appropriate decisions as needed. The actual CO_2 data as a function of time results in a nonstationary time series. For details in the development of this model, see Shih and Tsokos (2009). The best forecasting model that we developed is an ARIMA model with second order autoregressive process, with a first order moving average process and a 12-month seasonal effect. Its final form is given by

$$\begin{aligned}\hat{CO}_{2_A} &= 0.6887x_{t-1} + 0.1989x_{t-2} + 0.1124x_{t-3} + 1.0759x_{t-12} \\ &\quad - 0.74097x_{t-13} - 0.213997x_{t-14} - 0.12093x_{t-15} \\ &\quad - 0.0683x_{t-24} + 0.047038x_{t-25} + 0.013585x_{t-26} \\ &\quad + 0.00768x_{t-27} - 0.00076x_{t-36} + 0.005234x_{t-37} \\ &\quad + 0.0015116x_{t-38} + 0.00085x_{t-39} - 0.8787\varepsilon_{t-12}.\end{aligned}$$

A similar statistical model can be developed for CO_2 emission, Shih and Tsokos (2009).

A Differential Equation of CO_2 in the Atmosphere

The main attributable variables in CO_2 in the atmosphere are:

E: CO_2 emission (fossil fuel combination)
D: Deforestation and destruction
R: Terrestrial plant respiration
S: Respiration
O: the flux from oceans to atmosphere
P: terrestrial photosynthesis
A: the flux from atmosphere to oceans
B: Burial of organic carbon and limestone carbon

One important question that we would like to know is the rate of change of CO_2 as a function of time. The general form of the differential equation of the subject matter is of the form:

$$\frac{d(CO_2)}{dt} = f(E, D, R, S, O, P, A, B)$$

or

$$CO_{2_A} = \int (E + D + R + S + (O - A) - P - B)dt.$$

Here, B, P and R are constants, thus

$$\begin{aligned}CO_{2_A} = \int &(k_E E + k_D D + k_R R + k_S S + k_{O-A}(O - A) \\ &+ k_P P - k_B B)dt.\end{aligned}$$

Using the available data we can estimate the functional analytical form of all the attributable variables that appear

in the integrand. Thus, the final working form of CO_2 in the atmosphere is given by

$$CO_2 = \begin{cases} k_E\{-593503t + 2.4755 \times 10^9 e^{-\frac{1}{1200}}\} \\ +k_D(10730.5t + 0.01625t^2) \\ +k_S\{-0.132(1995 + \frac{t}{12})^4 + 1054.4(1995 + \frac{t}{12})^3 \\ -315462(1995 + \frac{t}{12})^2 + 3 \times 10^8 t\} \\ +K_{A-O}\{42.814t - 4.2665t^2 \\ +0.0967t^3\} - k_P \int Pdt - k_B \int Bdt \end{cases}.$$

Having a workable form of the differential equation, we can develop the necessary algorithm to track the influence the attributable variables will have in estimating the change of rate of CO_2 as a function of time.

Conclusion

Finally, is the "Global Warming" phenomenon real? Yes. However, it is not as urgent as some environmentalists claim. For example, our statistical analytical models predict that in the next 10 years, 2019, we will have an increase of carbon dioxide in the atmosphere in the continental U.S. of approximately 7%. In developing a strategic legislative plan, we must address the economic impact it will have in our society. In our present global economic crisis, introducing legislation to address Global Warming issues will present additional critical economic problems. In a global context we must consider about 155 economic developing countries that have minimal to no strategic plans in effect that collect the necessary information that addresses the subject matter in their country. Furthermore, we have approximately 50 undeveloped countries that have minimum understanding about the concept of global warming. Thus, talking about developing global strategies and policies about "Global Warming" is quite premature.

Acknowledgments

This article is a revised and extended version of the paper published in *Hellenic News of America*, 23, 3, November 2009.

About the Author

Chris P. Tsokos is Distinguished University Professor of Mathematics and Statistics and Director of the Graduate Program in Statistics at the University of South Florida. He is the author/co-author of more than 285 research journal publications and more than 20 books plus special volumes. He has also directed more than 37 Ph.D. theses as a major professor. Dr. Tsokos is the recipient of many distinguished awards and honors, including Fellow of the American Statistical Association, USF Distinguished Scholar Award, Sigma Xi Outstanding Research Award, USF Outstanding Undergraduate Teaching Award, USF Professional Excellence Award, URI Alumni Excellence Award in Science and Technology, Pi Mu Epsilon, election to the International Statistical Institute, Sigma Pi Sigma, USF Teaching Incentive Program, and several humanitarian and philanthropic recognitions and awards. Professor Tsokos is an Editor/Chief-Editor/Co-Chief Editor of a number of journals including *International Journal of Environmental Sciences*, *International Journal of Mathematical Sciences*, *International Journal of Business Systems*, *International Journal of Nonlinear Studies*, and *Nonlinear Mathematics, Theory, Methods and Applications*. He also serves as an Associate Editor for a number of international journals.

"Professor Chris P. Tsokos' contributions to statistics, mathematical sciences, engineering and international education over a period of almost a half century are well-known, well-recognized and well-documented in the literature. In particular, his most notable work in the Bayesian reliability, stochastic dynamic systems and statistical modeling in a nonlinear and nonstationary world is well-recognized and well-established." (G. S. Ladde and M. Sambandham (2008). Professor Chris P. Tsokos: a brief review of statistical, mathematical and professional contributions and legacies, *Neural, Parallel & Scientific Computations*, 16 (1), Special issue in honor of Dr. Chris P. Tsokos.)

Cross References

►Environmental Monitoring, Statistics Role in
►Forecasting with ARIMA Processes
►Marine Research, Statistics in
►Statistics and Climate Change
►Time Series

References and Further Reading

Hachett K, Tsokos CP (2009) A new method for obtaining a more effective estimate of atmospheric temperature in the continental United States. Nonlinear Anal-Theor 71(12):e1153–e1159

Shih SH, Tsokos CP (2007) A weighted moving average procedure for forecasting. J Mod Appl Stat Meth 6(2):619–629

Shih SH, Tsokos CP (2008a) A temperature forecasting model for the continental United States. J Neu Par Sci Comp 16:59–72

Shih SH, Tsokos CP (2008b) Prediction model for carbon dioxide emission in the atmosphere (2008). J Neu Par Sci Comp 16: 165–178
Shih SH, Tsokos CP (2009) A new forecasting model for nonstationary environmental data. Nonlinear Anal-Theor 71(12):e1209–e1214
Tsokos CP (2007a) St. Petersburg Times, Response to "Global Warming: Meet Your News Adversary"
Tsokos CP (2007b) Global warming: MEDIA CHAOS: can mathematics/statistics help? International Conference on Dynamical Systems and Applications, Atlanta, GA
Tsokos CP (2008a) Statistical modeling of global warming. Proc Dyn Syst Appl 5:460–465
Tsokos CP (2008b) Global warming (2008). The Fifth World Congress of IFNA (July 2–9, Orlando, Florida)
Tsokos CP, Xu Y (2009) Modeling carbon dioxide emission with a system of differential equations. Nonlinear Anal-Theor 71(12):e1182–e1197
Wooten R, Tsokos CP (2010) Parametric analysis of carbon dioxide in the atmosphere. J Appl Sci 10:440–450

Maximum Entropy Method for Estimation of Missing Data

D. S. Hooda
Professor and Dean (Research)
Jaypee University of Engineering and Technology, Guna, India

In field experiments we design the field plots. In case we find one or more observations missing due to natural calamity or destroyed by a pest or eaten by animals, it is cumbersome to estimate the missing value or values as in field trials it is practically impossible to repeat the experiment under identical conditions. So we have no option except to make best use of the data available. Yates (1933) suggested a method: "Substitute x for the missing value and then choose x so as to minimize the error sum of squares."

Actually, the substituted value does not recover the best information, however, it gives the best estimate according to a criterion based on the least square method. For the randomized block experiment

$$x = \frac{pP + qQ - T}{(p-1)(q-1)}, \tag{1}$$

where

p = number of treatments;
q = number of blocks;
P = total of all plots receiving the same treatment as the missing plot;
Q = total of all plots in the same block as the missing plot; and
T = total of all plots.

For the Latin Square Design, the corresponding formula is

$$x = \frac{p(P_r + P_c + P_t) - 2T}{(p-1)(q-1)}, \tag{2}$$

where

p = number of rows or columns of treatments;
P_r = total of row containing the missing plot;
P_c = total of column containing the missing plot;
P_t = total of treatment contained in the missing plot; and
T = grand total.

In case more than one plot yields are missing, we substitute the average yield of available plots in all except one of these and substitute x in this plot. We estimate x by Yate's method and use this value to estimate the yields of other plots one by one.

Next we discuss the maximum entropy method. If $x_1, x_2, \ldots, x_n$ are known yields and x is the missing yield. We obtain the maximum entropy estimate refer to Kapur and Kesavan (1992) for x by maximizing:

$$-\sum_{i=0}^{n} \frac{x_i}{T+x} \log \frac{x_i}{T+x} - \frac{x}{T+x} \log \frac{x}{T+x}. \tag{3}$$

Thus we get

$$\hat{x} = \left[x_1^{x_1} x_2^{x_2} \ldots x_n^{x_n}\right]^{\frac{1}{T}}, \tag{4}$$

where $T = \sum_{i=1}^{n} x_i$.

The value given by (4) is called maximum entropy mean of $x_1, x_2, \ldots, x_n$.

Similarly, if two values x and y are missing, x and y are determined from

$$\hat{x} = \left[x_1^{x_1} x_2^{x_2} \ldots x_n^{x_n}\right]^{\frac{1}{T+y}}, \tag{5}$$

$$\hat{y} = \left[x_1^{x_1} x_2^{x_2} \ldots x_n^{x_n}\right]^{\frac{1}{T+x}}. \tag{6}$$

The solution of (5) and (6) is

$$\hat{x} = \hat{y} = \left[x_1^{x_1} x_2^{x_2} \ldots x_n^{x_n}\right]^{\frac{1}{T}}. \tag{7}$$

Hence all the missing values have the same estimate and this does not change if the missing values are estimated one by one.

There are three following drawbacks of the estimate given by (4)

(1) $\hat{x}$ is rather unnatural. In fact $\hat{x}$ is always greater than arithmetic mean of $x_1, x_2, \ldots, x_n$.
(2) If two values are missing, the maximum entropy estimated for each is the same as given by (7).
(3) This is not very useful for estimating missing values in design of experiments.

The first drawback can be overcome by using generalized measure of entropy instead of Shannon entropy. If we use Burg's measure given by

$$B(P) = \sum_{i=1}^{n} \log p_i. \tag{8}$$

Then we get the estimate

$$\hat{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \overline{x}. \tag{9}$$

In fact we choose a value $\hat{x}$, which is as equal to $x_1, x_2, \ldots, x_n$ as possible and so we maximize a measure of equality. Since there are many measures of equality, therefore our estimate will also depend on the measure of equality we choose.

The second drawback can be understood by considering the fact that the information theoretic estimate for a missing value depends on:

(a) The information available to us
(b) The purpose for which missing value is to be used.

As for the third drawback, according to the principle of maximum entropy, we should use all the information given to us and avoid scrupulously using any information not given to us. In design of experiments, we are given information about the structure of the design, which we are not using this knowledge in estimating the missing values. Consequently, the estimate is not accurate; however, information theoretic model defined and studied by Hooda and Kumar (2005) can be applied to estimate the missing value x_{ij} in contingency tables. Accordingly, the value x_{ij} is to be chosen to minimize the measure of dependence D.

About the Author

Professor D. S. Hooda is Vice President of the International Forum of Interdisciplinary Mathematics. He is General Secretary of Indian Society of Information Theory and Applications. He is an Elected member of the International Statistical Institute. American Biographical Institute, USA, chose him in 2004 for his outstanding research and conferred with honorary appointment to Research Board of Advisors of the institute. Indian Society of Information Theory has bestowed on him a prestigious award in 2005 for his outstanding contribution and research in information theory. He was Pro-Vice-Chancellor of Kurukshetra University. He has published about 80 papers in various journals and four books in mathematics and statistics. Presently, Professor Hooda is Dean (Research) Jaypee Institute of Engineering and Technology, Raghogarh, Guna.

Cross References

►Entropy
►Estimation
►Estimation: An Overview
►Nonresponse in Surveys
►Nonsampling Errors in Surveys
►Sampling From Finite Populations

References and Further Reading

Hooda DS, Kumar P (2005) Information theoretic model for analyzing independence of attributes in contingency table. Paper presented at the international conference held at Kuala Lumpur, Malayasia, 27–31 Dec 2005

Kapur JN, Kesavan HK (1992) Entropy optimization principles with applications. Academic, San Diego

Yates F (1933) The analysis of replicated experiments when the field experiments are incomplete. Empire J Exp Agr 1:129–142

Mean, Median and Mode

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

Mean, median and mode indicate central point of distribution or data set. Let P_X denotes distribution of a random variable X. Any reasonable rule $\mathcal{O} = \mathcal{O}(P_X)$ indicating a point $\mathcal{O}$ to be the center of P_X should satisfy the following postulates:

A1 If $P(a \le X \le b) = 1$ then $a \le \mathcal{O}(P_X) \le b$
A2 $\mathcal{O}(P_{X+c}) = \mathcal{O}(P_X) + c$ for any constant c [transitivity]
A3 $\mathcal{O}(P_{cX}) = c\mathcal{O}(P_X)$ for any constant c [homogeneity]

The *mean* is a synonym of the first moment, i.e. the expected value EX. For a continuous random variable X it may be expressed in terms of density function $f(x)$, as the integral $EX = \int_{-\infty}^{+\infty} xf(x)dx$. In discrete case it is defined as the sum of type $EX = \sum_i x_i p_i$, where x_i is a possible value of X, $i \in I$, while $p_i = P(X = x_i)$ is its probability. The mean fulfils all the above postulates and, moreover, an extra condition

AM $E(X-EX)^2 \le E(X-c)^2$ for any $c \in R$

It is worth to add that mean may not exist.

The *median* $Me = Me(X)$ is a scalar α defined by conditions $P_X(X \le \alpha) \ge \frac{1}{2}$ and $P_X(X \ge \alpha) \ge \frac{1}{2}$. In terms of the cumulative distribution function $F = F_X$ it means that $F(\alpha) \ge \frac{1}{2}$ and $\lim_{x \uparrow \alpha} F(\alpha) \le \frac{1}{2}$. In particular, if X is continuous with density f, then the desired conditions reduces to $\int_{-\infty}^{\alpha} f(x)dx \ge \frac{1}{2}$ and $\int_{\alpha}^{\infty} f(x)dx \ge \frac{1}{2}$. In discrete case it can be expressed in the form $\sum_{\{i:x_i \le \alpha\}} p_i \ge \frac{1}{2}$ and $\sum_{\{i:x_i \ge \alpha\}} p_i \ge \frac{1}{2}$. The median also satisfies the conditions $A1 - A3$ and, moreover

AMe $E|X - MeX| \le E|X - c|$ for any $c \in R$.

The mode $Mo = Mo(X)$ of a random variable X is defined in terms of its density function f (continuos case) or its probability mass function $p_i = P(X = x_i)$ (discrete case). Namely, $Me(X) = \arg\max f(x)$, or is an element x in the set of possible values $\{x_i : i \in I\}$ that $P(X = x) = \max\{p_i : i \in I\}$. The mode also satisfies the conditions $A1 - A3$. It is worth to add that mode may not be unique. There exist bimodal and multimodal distributions. Moreover the set of possible modes may be interval.

In the context of data set, represented by a sequence $x = (x_1, \ldots, x_n)$ of observations, the postulates $A1 - A3$ may be reformulated as follows:

S1 $\mathcal{O}(x_{i_1}, \ldots, x_{i_n}) = \mathcal{O}(x_1, \ldots, x_n)$ for any permutation $i_1, .., i_n$ of the indices $1, \ldots, n$
S2 $\min\{x_1, \ldots, x_n\} \le \mathcal{O}(x_1, \ldots, x_n) \le \max\{x_1, \ldots, x_n\}$
S3 $\mathcal{O}(x_1 + c, \ldots, x_n + c) = \mathcal{O}(x_1, \ldots, x_n) + c$
S4 $\mathcal{O}(cx_1, \ldots, cx_n) = c\mathcal{O}(x_1, \ldots, x_n)$.

In this case the mean, median and mode are defined as follows.

The mean of the data $x = (x_1, \ldots, x_n)$, denoted usually by $\overline{x}$, is the usual arithmetic average $\overline{x} = \frac{1}{n}\sum x_i$. The mean not only satisfies all conditions $S1 - S4$ but also possesses the property

SM $\sum_{i=1}^n (x_i - \overline{x})^2 \le \sum_{i=1}^n (x_i - c)^2$ for all $c \in R$.

Now let us arrange the elements of the sequence $x = (x_1, \ldots, x_n)$ in the not decreasing order $x_{[1]} \le x_{[2]} \le \ldots \le x_{[n]}$. The median of the data set $x = (x_1, \ldots, x_n)$ is defined by the formula

$$Me(x) = \begin{cases} x_{[\frac{n+1}{2}]}, & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]}\right) & \text{if } n \text{ is even.} \end{cases}$$

The median satisfies the conditions $S1 - S4$ and, moreover,

SMe $\sum_{i=1}^n |x_i - Me(x)| \le \sum_{i=1}^n |x_i - c|$ for all $c \in R$.

The mode of the data $x = (x_1, \ldots, x_n)$, denoted by $Mo(x)$, is the value in the set that occurs most often. For instance if $x = (7, 3, 18, 24, 9, 3, 18)$ then $x \uparrow = (3, 7, 9, 13, 18, 18, 24)$. For such data $Me(x) = x_{[4]} = 13$ and $Mo(x) = 18$.

It is worth to add that the mean is very sensitive for outlying observations.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Asymptotic Relative Efficiency in Estimation
►Expected Value
►Geometric Mean
►Harmonic Mean
►Mean, Median, Mode: An Introduction
►Random Variable
►Robust Statistical Methods
►Sampling Distribution
►Skewness

References and Further Reading

Cramér H (1946) Mathematical methods of statistics. Princeton University Press, Princeton
Joag-Dev K (1989) MAD property of median. A simple proof. Am Stat 43:26–27
Prokhorov AW (1982a) Expected value. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 3. Soviet Encyclopedia, Moscow, pp 600–601 (in Russian)
Prokhorov AW (1982b) Mode. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 3. Soviet Encyclopedia, Moscow p 763 (in Russian)

Mean, Median, Mode: An Introduction

S. N. Gupta
University of South Pacific, Suva, Fiji

Introduction

Mean, median and mode are three statistical measures commonly used to summarize data sets. They are known by the common name *average*. In its broadest sense, an *average* is simply any single value that is representative of

many numbers. Averages are also called *measures of central tendency* because an average is usually located near the center of the data set. Some examples: average age of the players of a cricket team, average reaction time of a particular chemical, average amount spent by a customer in a shopping mall, etc.

The Mean

The *mean*, also known as *arithmetic mean*, is the most widely used average and is defined as the sum of the observations divided by the number of observations. The formula for computing mean is: $\bar{x} = (\sum x)/n$, where $\bar{x}$ is the symbol for mean (pronounced "x-bar"), x is the *symbol* for variable, $\sum x$ is the *sum* of observations (i.e., the sum of the values of the variable x) and n is the *number* of observations.

Although, there are also other kinds of means (such as the ▶harmonic mean and the ▶geometric mean), the arithmetic mean is by far the most popular. For this reason, the word arithmetic is rarely used in practice and we simply refer to the "mean."

Example 1 The ages (in weeks) of five babies are 5, 9, 8, 6 and 10. Find the mean.

Solution: The mean of the set is given by $\bar{x} = \frac{1}{n}\sum x = \frac{5+9+8+6+10}{5} = \frac{38}{5} = 7.6$ weeks.

Calculation of Mean for Discrete Frequency Distribution

Sometimes, it is convenient to represent the data in form of a frequency distribution. In such cases the formula for mean is: $\bar{x} = \frac{\sum fx}{\sum f}$, where f is the frequency, $\sum f$ is the sum of the frequencies, $\sum fx$ is the sum of each observation multiplied by its frequency.

Example 2 Data for numbers of children in 35 families are given below. Find the mean.

No. of children (x):	0	1	2	3	4
Frequency (f):	2	9	11	8	5

Solution:

x	0	1	2	3	4	
f	2	9	11	8	5	$\sum f = 35$
fx	0	9	22	24	20	$\sum fx = 75$

The mean $\bar{x} = \frac{\sum fx}{\sum f} = \frac{75}{35} = 2.1$ children per family.

Calculation of Mean for Grouped Frequency Distribution

It is not possible to calculate exact mean in grouped frequency distribution, because some information is lost when the data are grouped. So, only an approximate value of mean is obtained based on the assumption that all observations in a class interval occur at the *midpoint* (x_m) of that interval. Thus, the formula of Example 2 can be used after replacing x by x_m.

Example 3 The following is the distribution of the number of fish caught by 50 fishermen in a village. Find the mean number of fish caught by a fisherman.

No. of fish caught:	11–15	16–20	21–25	26–30
No. of fishermen:	12	14	13	11

Solution:

No. of fish caught	Midpoint (x_m)	f	fx_m
11–15	13	12	156
16–20	18	14	252
21–25	23	13	299
26–30	28	11	308
		$\sum f = 50$	$\sum fx_m = 1015$

Therefore, the mean is $\bar{x} = \frac{\sum fx_m}{\sum f} = \frac{1015}{50} = 20.3$ fish per fisherman.

Weighted Mean

When *weights* (measures of relative importance) are assigned to observations, weighted means are used. If an observation x is assigned a weight w, the weighted mean is given by $\bar{x} = \sum wx / \sum w$.

The Median

The *median* is another kind of average. It is defined as the centre value when the data are arranged in order of magnitude. Thus, the median is a value such that 50% of the data are below median and 50% are above median.

Calculation of Median for Raw Data

The observations are first arranged in ascending order of magnitude. If there are n observations, the median is

1. The value of the $[(n+1)/2]$th observation, *when n is odd.*
2. The mean of the $[n/2]$th and $[(n/2)+1]$th observations, *when n is even.*

Example 4 Find the median for the following data set:

16, 32, 20, 13, 13, 24, 10.

Solution: Arranging the data in ascending order we have

10, 13, 13, 16, 20, 24, 32.

Here, $n = 7$, which is odd. Therefore, median = $\frac{n+1}{2}$ th score=$\frac{7+1}{2}$th score=4th score=16.

Example 5 Find the median for the data:

17, 18, 26, 30, 19, 24, 20, 22, 29, 25.

Solution: Here, n = 10, which is even. Arranging the data in ascending order we have

17, 18, 19, 20, 22, 24, 25, 26, 29, 30.

$$\text{Therefore, median} = \frac{1}{2}\left[\frac{n}{2}\text{th score} + \left(\frac{n}{2}+1\right)\text{th score}\right]$$
$$= \frac{1}{2}\left[\frac{10}{2}\text{th score} + \left(\frac{10}{2}+1\right)\text{th score}\right]$$
$$= \frac{1}{2}[5\text{th score} + 6\text{th score}]$$
$$= \frac{1}{2}[22 + 24] = 23.$$

Calculation of Median for Discrete Frequency Distribution
The same basic formulae as used for raw data are used, but cumulative frequencies are calculated for convenience of locating the observations at specific numbers.

Example 6 Data for the number of books purchased by 28 customers are given below. Find the median.

No. of books (x):	1	2	3	4
No. of customers (f) :	5	9	8	6

Solution:

No. of books (x)	1	2	3	4
No. of customers (f)	5	9	8	6
Cumulative frequency ($c.f.$)	5	14	22	28

Here $n = \sum f = 28$ (even). Therefore,

$$\textit{median} = \frac{1}{2}\left[\frac{28}{2}\textit{th score} + \left(\frac{28}{2}+1\right)\text{th score}\right]$$
$$= \frac{1}{2}[14\text{th score} + 15\text{th score}] = \frac{1}{2}[2+3] = 2.5$$

Calculation of Median for Grouped Frequency Distribution
In a grouped distribution, exact median cannot be obtained because some information is lost in grouping. Here, we first locate the *median class* and then obtain an estimate of the *median* by the formula:

$$\textit{median} = l_1 + \frac{\left(\frac{n}{2} - c\right)}{f}(l_2 - l_1)$$

where, l_1, l_2 are the lower and upper boundaries of the median class, f is the frequency of the median class, n is the sum of all frequencies and c is the cumulative frequency of the class immediately preceding the median class.

Example 7 Find the median for the data of Example 3 above.

Solution: Construct a table for class boundaries and cumulative frequencies:

Class	Class boundaries	f	$c.f.$
11–15	10.5–15.5	12	12
16–20	15.5–20.5	14	26
21–25	20.5–25.5	13	39
26–30	25.5–30.5	11	50
		n = 50	

Here, $n/2$ = 25. The median will lie in the class having cumulative frequency ($c.f.$) just larger than 25. The median class is 16–20. Thus, l_1 = 15.5, l_2 = 20.5, c = 12, f = 14.

Hence, *median* = $15.5 + \left(\frac{25-12}{14}\right) \times 5 = 15.5 + 4.64 =$ 20.14.

The Mode

The *mode* is the most *frequent* value i.e., the value that has the largest frequency. A major drawback of mode is that a data set may have more than one mode or no mode at all. Also the mode may not always be a central value as in the Example 8(a) below.

Example 8 Find mode in the following data sets:

(a) 5, 5, 6, 7, 7, 8, 8, 9, 9, 9, 9.

(b) 12, 14, 15, 15, 15, 19, 19, 19, 20, 20.

(c) 11, 15, 16, 19, 21, 23, 26, 27, 29, 30.

Solution

(a) One mode at 9, (b) Two modes at 15 and 19, (c) No mode as each value occurs only once. For grouped frequency distribution, the mode can be estimated by taking the mid-point of the *modal class* corresponding to the

largest frequency. One advantage of mode is that it can be calculated for both kinds of data, qualitative and quantitative, whereas mean and median can be calculated for only quantitative data. E.g., A group consists of five Hindus, six Muslims and nine Christians. Here, Christianity is most frequent and so it is the mode of this data set.

Remarks If a distribution is symmetrical then mean = median = mode. For skewed distributions a thumb rule (though not without exceptions) is that if the distribution is skewed to the right then mean > median > mode and the inequalities are reversed if the distribution is skewed to the left.

To sum up, there is no general rule to determine which average is most appropriate for a given situation. Each of them may be better under different situations. Mean is the most widely used average followed by median. The median is better when the data set includes ▶outliers or is open ended. Mode is simple to locate and is preferred for finding the most popular item e.g. most popular drink or the most common size of shoes etc.

Cross References

▶Geometric Mean
▶Harmonic Mean
▶Mean Median and Mode
▶Skewness

References and Further Reading

Bluman AG (2007) Elementary statistics: a step by step approach, 6th edn. McGraw Hill, New York
Croucher JS (2002) Statistics: making business decisions. McGraw Hill/Irwin, New York
Mann PS (2006) Introductory statistics, 6th edn. Wiley, New York

Mean Residual Life

Jonathan C. Steele[1], Frank M. Guess[2], Timothy M. Young[2], David J. Edwards[3]
[1]Minitab, Inc., State College, PA, USA
[2]Professor
University of Tennessee, Knoxville, TN, USA
[3]Assistant Professor
Virginia Commonwealth University, Richmond, VA, USA

Theories and applications that use Mean Residual Life (MRL) extend across a myriad of helpful fields, while the methods differ considerably from one application to the next. Accelerated stress testing, fuzzy set engineering modeling, mixtures, insurance assessment of human life expectancy, maintenance and replacement of bridges, replacement of safety significant components in power plants, and evaluation of degradation signals in systems are just a few examples of applications of MRL function analysis. Note that MRL is also called "expected remaining life," plus other phrase variations. For a random lifetime X, the MRL is the conditional expectation $E(X - t|X > t)$, where $t \geq 0$. The MRL function can be simply represented with the reliability function $R(t) = P(X > t) = 1 - F(t)$ as:

$$e(t) = E(X - t|X > t) = \frac{\int_t^\infty R(x)dx}{R(t)}$$

where $R(t) > 0$ for $e(t)$ to be well defined. When $R(0) = 1$ and $t = 0$, the MRL equals the average lifetime. When $R(t) = 0$, then $e(t)$ is defined to be 0. The empirical MRL is calculated by substituting either the standard empirical estimate of $R(t)$ or, when censoring occurs, by substituting the Kaplan-Meier estimate of $R(t)$ (see ▶Kaplan-Meier Estimator). To use the Kaplan-Meier estimate when the final observation is censored requires a modification to define the empirical reliability function as eventually 0.

The reliability function can also be represented as a function of the MRL as:

$$R(t) = \left(\frac{e(0)}{e(t)}\right) \exp^{-\int_0^t \left[\frac{1}{e(x)}\right]dx}.$$

Note that the MRL function can exist, while the hazard rate function might not exist, or vice versa, the hazard rate function can exist while the MRL function might not. Compare Guess and Proschan (1988) plus Hall and Wellner (1981) for comments. When both functions exist, and the MRL function is differentiable, the hazard rate function is a function of the MRL:

$$h(t) = \frac{1 + e'(t)}{e(t)}$$

where $e'(t)$ is the first derivative of the MRL function.

The breadth of applications for the MRL function is astounding. As examples, Chiang (1968) and Deevey (1947) cite the use of the MRL for annuities via expected life tables (see ▶Life Table) in ancient Roman culture. Bhattacharjee (1982) suggests how to use the MRL to decide when to sell an item that has maintenance costs, which has copious natural applications, such as to real estate. Steele (2006) and Guess et al. (2006) illustrate a confidence interval for the range of values where one MRL function dominates

another and use it to reveal an opportunity to increase the profitability of a process that manufactures engineered medium density fiberboard. See also the insightful results on MRL functions of mixtures, ►order statistics, and coherent systems from Navarro and Hernandez (2008). Another topic of extensive research over the years is testing classes of MRL functions. For more on those tests, see references in Hollander and Proschan (1984), Hollander and Wolfe (1999) or Anis et al. (2004), for example. A brief list of other MRL papers, among many wide-ranging papers available, includes Peiravi and Dehqanmongabadi (2008), Zhao and Elsayed (2005), Bradley and Gupta (2003), Asadi and Ebrahimi (2000), Oakes and Dasu (1990), Berger et al. (1988), Guess and Park (1988), and Guess et al. (1986). We would recommend many other useful papers, but space severely limits our list.

While we do not give a complete inventory, note that R packages like *evd*, *ismev*, and *locfit* possess capabilities such as MRL plotting and/or computing the MRL for censored data; compare Shaffer et al. (2008). Another freeware, Dataplot, the software for the NIST website, does a MRL plot, but calls it a "conditional mean exceedance" plot, see Heckert and Filliben (2003). For-profit statistical software, such as JMP, MINITAB, PASW (formerly SPSS), SAS, etc., can be appropriately utilized for computing the MRL, using the basic formulas above (PASW and others use the phrase "life tables," which often contain a column for MRL). Pathak et al. (2009) illustrate the use of MATLAB for computing several different lifetime data functions including the MRL. Steele (2006) computes MRL via Maple.

Cross References

►Conditional Expectation and Probability
►Hazard Ratio Estimator
►Kaplan-Meier Estimator
►Life Expectancy
►Life Table

References and Further Reading

Anis MZ, Basu SK, Mitra M (2004) Change point detection in MRL function. Indian Soc Probab Stat 8:57–71

Asadi M, Ebrahimi N (2000) Residual entropy and its characterizations in terms of hazard function and mean residual life function. Stat Probab Lett 49(3):263–269

Berger RL, Boos DD, Guess FM (1988) Tests and confidence sets for comparing two mean residual life functions. Biometrics 44(1):103–115

Bhattacharjee MC (1982) The class of mean residual lives and some consequences. J Algebra Discr 3(1):56–65

Bradley DM, Gupta RC (2003) Limiting behaviour of the mean residual life. Ann I Stat Math 55(1):217–226

Chiang CL (1968) Introduction to stochastic processes in biostatistics. Wiley, New York

Deevey ES (1947) Life tables for natural populations of animals. Q Rev Biol 22:283–314

Guess FM, Hollander M, Proschan F (1986) Testing exponentiality versus a trend change in mean residual life. Ann Stat 14(4): 1388–1398

Guess FM, Park DH (1988) Modeling discrete bathtub and upside-down bathtub mean residual-life functions. IEEE T Reliab 37(5):545–549

Guess FM, Proschan F (1988) MRL: theory and applications. In: Krishnaiah PR, Rao CR (eds) Handbook of statistics 7: quality control and reliability. North Holland, Amsterdam, pp 215–224

Guess FM, Steele JC, Young TM, León RV (2006) Applying novel mean residual life confidence intervals. Int J Reliab Appl 7(2):177–186

Hall WJ, Wellner JA (1981) Mean residual life. In: Csörgö ZM et al (eds) Statistics and related topics. North Holland, Amsterdan, pp 169–184

Heckert NA, Filliben JJ (2003) CME plot. In: NIST handbook 148: DATAPLOT reference manual, volume I: commands, National Institute of Standards and Technology Handbook Series, pp 2-45–2-47. For more details see link: http://www.itl.nist.gov/div898/software/dataplot/document.htm

Hollander M, Proschan F (1984) Nonparametric concepts and methods in reliability. In: Krishnaiah PR, Sen PK (eds) Handbook of statistics 4: nonparametric methods. North Holland, Amsterdam, pp 613–655

Hollander M, Wolfe D (1999) Nonparametric statistical methods, 2nd edn. Wiley, New York

Navarro J, Hernandez PJ (2008) Mean residual life functions of finite mixtures, order statistics and coherent systems. Metrika 67(3):277–298

Oakes D, Dasu T (1990) A note on residual life. Biometrika 77(2):409–410

Pathak R, Joshi S, Mishra DK (2009) Distributive computing for reliability analysis of MEMS devices using MATLAB. In: Proceedings of the international conference on advances in computing, communication and control (Mumbai, India, January 23–24, 2009). ACM, New York, pp 246–250

Peiravi A, Dehqanmongabadi N (2008) Accelerated life testing based on proportional mean residual life model for multiple failure modes. J Appl Sci 8(22):4166–4172

Shaffer LB, Young TM, Guess FM, Bensmail H, León RV (2008) Using R software for reliability data analysis. Int J Reliab Appl 9(1):53–70

Steele JC (2006) "Function domain sets" confidence intervals for the mean residual life function with applications in production of medium density fiberboard. Thesis at University of Tennessee, Knoxville, TN. Available at link: http://etd.utk.edu/2006/SteeleJonathanCody.pdf

Zhao WB, Elsayed EA (2005) Modelling accelerated life testing based on mean residual life. Int J Syst Sci 36(11): 689–696

Measure Theory in Probability

Milan Merkle
Professor, Faculty of Electrical Engineering
University of Belgrade, Belgrade, Serbia

Foundations of Probability: Fields and Sigma-Fields

Since Kolmogorov's axioms, Probability theory is a legitimate part of Mathematics, with foundations that belong to Measure theory. Although a traditional probabilist works solely with countably additive measures on sigma fields, the concepts of countable additivity and infinite models are by no means natural. As Kolmogorov [1956 p. 15] points out, "... in describing any observable random process we can obtain only finite fields of probability. Infinite fields of probability occur only as idealized models of real random processes."

To build a probability model, we need first to have a non-empty set Ω which is interpreted as a set of all possible outcomes of a statistical experiment. Then we define which subsets of Ω will be assigned a probability. The family $\mathcal{F}$ of all such subsets has to satisfy

(1) $\Omega \in \mathcal{F}$,
(2) $B \in \mathcal{F} \implies B' \in \mathcal{F}$,
(3) $B_1, B_2 \in \mathcal{F} \implies B_1 \cup B_2 \in \mathcal{F}$,

and then we say that $\mathcal{F}$ is a field. If (3) is replaced by stronger requirement

(3') $B_1, B_2, \ldots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} B_i \in \mathcal{F}$

then we say that $\mathcal{F}$ is a sigma field.

The family $\mathcal{P}(\Omega)$ of all subsets of Ω is a field, and it is the largest field that can be made of subsets of Ω – it clearly contains all other possible fields. The smallest such field is $\mathcal{F}_0 = \{\emptyset, \Omega\}$; it is a subset of any other field.

The intersection of any family of fields is again a field. The union of a family of fields need not be a field. Both statements hold for sigma-fields, too.

Given a collection $\mathcal{A}$ of subsets of Ω, the intersection of all fields (sigma-fields) that contain $\mathcal{A}$ is called a field (sigma-field) *generated by* $\mathcal{A}$.

Having a non-empty set Ω and a field $\mathcal{F}$ of its subsets, a finitely additive probability measure is a function $P: \mathcal{F} \to \mathbb{R}_+$ such that

(a) $P(\Omega) = 1$.
(b) $P(A) \geq 0$ for every $A \in \mathcal{F}$.
(c) $P(A \cup B) = P(A) + P(B)$ whenever $A, B \in \mathcal{F}$ and $A \cap B = \emptyset$ (*finite additivity*).

If (c) is replaced by the condition of *countable additivity*

(c') For any countable collection $A_1, A_2, \ldots$ of sets in $\mathcal{F}$, such that $A_i \cap A_j = \emptyset$ for any $A_i \neq A_j$ and such that $A_1 \cup A_2 \cup \cdots \in \mathcal{F}$ (the latter condition is needless if $\mathcal{F}$ is a sigma-field):

$$P\left(\bigcup_{i=1}^{+\infty} A_i\right) = \sum_{i=1}^{+\infty} P(A_i)$$

then P is called (a countably additive) *probability measure*, or just *probability*. The triplet $(\Omega, \mathcal{F}, P)$ is called a *probability space*. By Carathéodory extension theorem, any countably additive probability measure P defined on a field $\mathcal{F}$ extends uniquely to a countably additive probability measure on the sigma field generated by $\mathcal{F}$; hence, if P is countably additive, we may always assume that $\mathcal{F}$ is a sigma-field.

A set $B \subset \Omega$ is called a *null set* if $B \subset A$ for some $A \in \mathcal{F}$ with $P(A) = 0$. Let $\mathcal{N}$ be a collection of all null sets in $(\Omega, \mathcal{F}, P)$. If $\mathcal{N} \subset \mathcal{F}$, the sigma-field $\mathcal{F}$ is called *complete*. For any sigma-field $\mathcal{F}$ there exists a complete sigma-field $\bar{\mathcal{F}}$, called a *completion* of $\mathcal{F}$, and defined as the sigma field generated by $\mathcal{F} \cup \mathcal{N}$.

A general positive measure μ is a set function defined on $(\Omega, \mathcal{F})$ with values in $\mathbb{R}_+ \cup \{+\infty\}$, which satisfies (b), (c) or (c'), and $\mu(\emptyset) = 0$. If $\mu(\Omega) < +\infty$, the measure is called *finite* and can be normalized to a probability measure by $P(A) = \mu(A)/\mu(\Omega)$ for all $A \in \mathcal{F}$. If Ω can be represented as a countable union of measurable sets of finite measure, then a measure is called *sigma-finite*. The most commonly used measure in Mathematics is the Lebesgue measure λ on $\mathbb{R}$, with the property that $\lambda([a, b]) = b - a$ for any $a < b$. This measure is not finite, as $\lambda(\mathbb{R}) = +\infty$, but it is sigma-finite.

If there exists a countable set $S \subset \Omega$ such that $\mu(S') = 0$, the measure μ is called *discrete*. Unless the measure is discrete, the sigma-field $\mathcal{F}$ is usually taken to be strictly smaller than $\mathcal{P}(\Omega)$, to ensure that it will be possible to assign some value of the measure to each set in $\mathcal{F}$. This is motivated by existence of non-measurable sets in $\mathbb{R}$ (sets that cannot be assigned any value of Lebesgue measure). Non-measurable sets cannot be effectively constructed and their existence is a consequence of Axiom of Choice [see Solovay (1970)]. The described construction of a probability space ensures that a probability can be assigned to all sets of interest.

The countable (vs. finite) additivity has a role to exclude from consideration measures that are too complicated, and also to enable applicability of fundamental theorems (for details on finitely additive measures see Yosida and Hewitt (1952)). Within axioms (a)-(b)-(c), the countable additivity is equivalent to *continuity of probability*, a property that can be described in two dual (equivalent) forms:

1. If $A_1 \subset A_2 \subset \cdots \subset \ldots$, then $P\left(\bigcup_{n=1}^{+\infty} A_n\right) = \lim_{n\to+\infty} P(A_n)$;
2. If $A_1 \supset A_2 \supset \cdots \supset \ldots$, then $P\left(\bigcap_{n=1}^{+\infty} A_n\right) = \lim_{n\to+\infty} P(A_n)$;

Random Variables and Their Distributions

Let $(\Omega, \mathcal{F}, P)$ be a probability space (usually called *abstract probability space*). Let X be a mapping from Ω to some other space S. A purpose of introducing such mappings can be twofold. First, in some simple models like tossing a coin, we prefer to have a numerical model that can also serve as a model for any experiment with two outcomes. Hence, instead of $\Omega = \{H, T\}$, we can think of $S = \{0,1\}$ as a set of possible outcomes, which are in fact labels for any two outcomes in a real world experiment. Second, in large scale models, we think of Ω as being a set of possible states of a system, but to study the whole system can be too difficult task, so by mapping we wish to isolate one or several characteristics of Ω.

While Ω can be a set without any mathematical structure, S is usually a set of real numbers, a set in $\mathbb{R}^d$, or a set of functions. To be able to assign probabilities to events of the form $\{\omega \in \Omega \mid X(\omega) \in B\} = X^{-1}(B)$, we have to define a sigma-field $\mathcal{B}$ on S, that will accommodate all sets B of interest. If S is a topological space, usual choices are for $\mathcal{B}$ to be generated by open sets in S (Borel sigma-field), or to be generated by all sets of the form $f^{-1}(U)$, where $U \subset S$ is an open set and f is a continuous function $S \mapsto \mathbb{R}$ (Baire sigma-field). Since for any continuous f and open U, the set $f^{-1}(U)$ is open, the Baire field is a subset of corresponding Borel field. In metric spaces (and, in particular, in $\mathbb{R}^d$, $d \geq 1$) the two sigma fields coincide.

A mapping $X : \Omega \mapsto S$ is called $(\Omega, \mathcal{F}) - (S, \mathcal{B})$ – measurable if $X^{-1}(B) \in \mathcal{F}$ for any $B \in \mathcal{B}$. The term *random variable* is reserved for such a mapping in the case when S is a subset of $\mathbb{R}$. Otherwise, X can have values in $\mathbb{R}^d$, when it is called a *random vector*, or in some functional space, when it is called a *random process*, where trajectories $X(\omega) = f(\omega, \cdot)$ depend on a numerical argument usually interpreted as time, or a *random field* if trajectories are functions of arguments that are not numbers. In general, X can be called *a random element*.

The central issue in a study of random elements is the probability measure $\mu = \mu_X$ induced by X on the space $(S, \mathcal{B})$ by $\mu_X(B) = P(X^{-1}(B))$, $B \in \mathcal{B}$, which is called the *probability distribution of* X. In fact, X is considered to be defined by its distribution; the mapping by itself is not of interest in Probability. In this way, each random element X is associated with two probability triplets: $(\Omega, \mathcal{F}, P)$ and $(S, \mathcal{B}, \mu)$. If a model considers only random variables that map Ω into S, then the first triplet can be discarded, or more formally, $(\Omega, \mathcal{F}, P)$ can be identified with $(S, \mathcal{B}, \mu)$.

The collection of sets $\{X^{-1}(B)\}_{B\in\mathcal{B}}$ is a sigma-field contained in $\mathcal{F}$, which is called a *sigma-field generated by* X, in notation $\sigma(X)$. It is considered in applications as a complete information about X, as it contains all relevant events in Ω from whose realizations we may deduce whether or not $X \in B$, for any $B \in \mathcal{B}$. In particular, if $\mathcal{B}$ contains all singletons $\{x\}$, then we know the value of X.

If there is another sigma-field $\mathcal{G}$ such that $\sigma(X) \subset \mathcal{G} \subset \mathcal{F}$, then we say that X is $\mathcal{G}$-measurable. In particular, if X is $\sigma(U)$-measurable, where U is another random element and if $\sigma(X)$ contains all sets of the form $X^{-1}(\{s\})$, $s \in S$, then X is a function of U.

The definition of a sigma-field does not provide any practical algorithm that can be used to decide whether or not a particular set belongs to a sigma field. For example, suppose that we have a Borel sigma-filed $\mathcal{B}$ on some topological space S, and we need to know whether or not $B \in \mathcal{B}$, for a given $B \subset S$. Then we need to either produce a formula that shows how to get B as a result of *countably many* unions, intersections and complements starting with open and closed sets, or to prove that such a formula does not exist. This is rarely obvious or straightforward, and sometimes it can require a considerable work. In cases when we want to show that a certain family of sets belongs to a given sigma-fields, the Dynkin's so-called "$\pi - \lambda$ theorem" is very useful. A collection $\mathcal{C}$ of subsets of a set S is called a π-system if $A \in \mathcal{C}, B \in \mathcal{C} \implies A \cap B \in \mathcal{C}$. It is called a λ-system if it has the following three properties: (1) $S \in \mathcal{C}$; (2) $A, B \in \mathcal{C}$ and $B \subset A \implies A \backslash B \in \mathcal{C}$; (3) For any sequence of sets $A_n \in \mathcal{C}$ with $A_n \subset A_{n+1}$ (increasing sets), it holds that $\sum_{i=1}^{+\infty} A_n \in \mathcal{C}$. Then we have the following.

Dynkin's $\pi - \lambda$ Theorem Let $\mathcal{A}$ be a π-system, $\mathcal{B}$ a λ-system and $\mathcal{A} \subset \mathcal{B}$. Then $\sigma(\mathcal{A}) \subset \mathcal{B}$.

Integration

Let X be a random variable that maps $(\Omega, \mathcal{F}, P)$ into $(\mathbb{R}, \mathcal{B}, \mu)$, where $\mathbb{R}$ is the set of reals, $\mathcal{B}$ is a Borel

sigma-algebra and μ is the distribution of X. The expectation of X is defined as

$$E\,X = \int_\Omega X(\omega)\mathrm{d}P(\omega) = \int_{\mathbb{R}} x\mathrm{d}\mu(x),$$

provided the integrals exist in the Lebesgue sense. By the construction of Lebesgue integral, E X exists if and only if E $|X|$ exists; in that case we say that X is integrable. To emphasize that the expectation is with respect to measure P, the notation $\mathrm{E}_P X$ can be used.

Let f be a measurable function $\mathbb{R} \to \mathbb{R}$ (in $\mathbb{R}$ we assume the Borel sigma-field if not specified otherwise). Then $f(X)$ is again a random variable, that is, the mapping $\omega \mapsto f(X(\omega))$ is $(\Omega, \mathcal{F}) - (\mathbb{R}, \mathcal{B})$ -measurable, and

$$Ef(X) = \int_\Omega f(X(\omega))\mathrm{d}P(\omega) = \int_{\mathbb{R}} f(x)\mathrm{d}\mu(x),$$

if the integral on the right hand side exists, and then we say that f is integrable. Expectations can be defined in the same way in more general spaces of values of f or X, for instance in $\mathbb{R}^d, d > 1$ or in any normed vector space.

Radon-Nikodym Theorem Suppose that P and Q are positive countably additive and sigma-finite measures (not necessarily probabilities) on the same space $(\Omega, \mathcal{F})$. We say that P is absolutely continuous with respect to Q (in notation $P \ll Q$) if $P(B) = 0$ for all $B \in \mathcal{F}$ with $Q(B) = 0$.

If $P \ll Q$, then there exists a non-negative measurable function f such that

$$P(A) = \int_\Omega I_A(\omega) f(\omega)\mathrm{d}Q(\omega), \quad \text{and}$$
$$\int_\Omega g(\omega)\mathrm{d}P(\omega) = \int_\Omega g(\omega) f(\omega)\mathrm{d}Q(\omega),$$

for any measurable g. The function f is called a *Radon-Nikodym derivative*, in notation $f = \frac{\mathrm{d}P}{\mathrm{d}Q}$, and it is Q-almost surely unique.

If Q is the Lebesgue measure and P a probability measure on $\mathbb{R}$, then the function f is called a *density* of P or of a corresponding random variable with the distribution P; distributions P on $\mathbb{R}$ that are absolutely continuous with respect to Lebesgue measure are called *continuous distributions*.

If both P and Q are probabilities and $P \ll Q$, then the ►Radon-Nikodym theorem yields that there exists a random variable $\Lambda \geq 0$ with $\mathrm{E}_Q \Lambda = 1$ such that

$$P(A) = \mathrm{E}_Q I_A \Lambda \qquad \text{and} \qquad \mathrm{E}_P X = \mathrm{E}_Q X\Lambda$$

for any random variable X.

Cross References

►Axioms of Probability
►Foundations of Probability
►Probability Theory: An Outline
►Radon–Nikodým Theorem
►Random Variable
►Stochastic Processes

References and Further Reading

Kolmogorov AN (1956) Foundations of the theory of probability, 2nd English edn. Chelsea, New York
Solovay RM (1970) A model of set-theory in which every set of reals is Lebesgue measurable. Ann Math Second Ser 92:1–56
Yosida K, Hewitt E (1952) Finitely additive measures. Trans Am Math Soc 72:46–66

Measurement Error Models

Alexander Kukush
Professor
National Taras Shevchenko University of Kyiv,
Kyiv, Ukraine

A (nonlinear) measurement error model (MEM) consists of three parts: (1) a *regression model* relating an observable regressor variable z and an unobservable regressor variable ξ (the variables are independent and generally vector valued) to a response variable y, which is considered here to be observable without measurement errors; (2) a *measurement model* relating the unobservable ξ to an observable surrogate variable x; and (3) a *distributional model* for ξ.

Parts of MEM

The *regression model* can be described by a conditional distribution of y given (z, ξ) and given an unknown parameter vector θ. As usual this distribution is represented by a probability density function $f(y|z, \xi; \theta)$ with respect to some underlying measure on the Borel σ-field of $\boldsymbol{R}$. We restrict our attention to distributions that belong to the exponential family, i.e., we assume f to be of the form

$$f(y|z, \xi; \beta, \varphi) = \exp\left(\frac{y\eta - c(\eta)}{\varphi} + a(y, \varphi)\right) \qquad (1)$$

with

$$\eta = \eta(z, \xi; \beta). \qquad (2)$$

Here β is the regression parameter vector, φ a scalar dispersion parameter such that $\theta = (\beta^T, \varphi)^T$, and a, c, and η are known functions. This class comprises the class of generalized linear models, where $\eta = \eta(\beta_0 + z^T\beta_z + \xi^T\beta_\xi)$, $\beta = (\beta_0, \beta_x^T, \beta_\xi^T)^T$.

The *classical measurement model* assumes that the observed variable x differs from the latent ξ by a measurement error variable δ that is independent of z, ξ, and y:

$$x = \xi + \delta \tag{3}$$

with $\mathbf{E}\delta = 0$. Here we assume that $\delta \sim N(0,\ \Sigma_\delta)$ with Σ_δ known. The observable data are independent realizations of the model $(x_i, y_i),\ \ i = 1, \ldots, n$.

Under the *Berkson measurement model*, the latent variable ξ differs from the observed x by a centered measurement error δ that is independent of z, x, and y:

$$\xi = x + \delta. \tag{4}$$

Thus, the values of x are fixed in advance, whereas the unknown true values, ξ, are fluctuating.

The *distributional model* for ξ either states that the ξ are unknown constants (*functional case*) or that ξ is a random variable (*structural case*) with a distribution given by a density $h(\xi; \gamma)$, where γ is a vector of nuisance parameters describing the distribution of ξ. In the structural case, we typically assume that

$$\xi \sim N(\mu_\xi,\ \Sigma_\xi), \tag{5}$$

although sometimes it is assumed that ξ follows a mixture of normal distributions. In the sequel, for the structural case we assume γ to be known. If not, it can often be estimated in advance (i.e., pre-estimated) without considering the regression model and the data y_i. For example, if ξ is normal, then μ_ξ and Σ_ξ can be estimated by $\overline{x}$ and $S_x - \Sigma_\delta$, respectively, where $\overline{x}$ and S_x are the empirical mean vector and the empirical covariance matrix of the data x_i, respectively.

The goal of measurement error modeling is to obtain nearly unbiased estimates of the regression parameter β by fitting a model for y in terms of (z, x). Attainment of this goal requires careful analysis. Substituting x for ξ in the model (1) – (2), but making no adjustments in the usual fitting methods for this substitution, leads to estimates that are biased, sometimes seriously.

In the structural case, the *regression calibration* (RC) estimator can be constructed by substituting $\mathbf{E}(\xi|x)$ for unobservable ξ. In both functional and structural cases, another, the simulation-extrapolation (*SIMEX*) estimator, becomes very popular. These estimators are not consistent in general, although they often reduce the bias significantly; see Carroll et al. (2006).

Polynomial and Poisson Model

We mention two important examples of the classical MEM (1) – (3) where for simplicity the latent variable is scalar and the observable regressor z is absent. The *polynomial model* is given by

$$y = \beta_0 + \beta_1 \xi + \ldots + \beta_k \xi^k + \varepsilon,$$

where $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$ and ε is independent of ξ. Here

$$\eta = \sum_{r=0}^{k} \beta_r \xi^r,\ \ c(\eta) = \frac{1}{2}\eta^2,$$

and $\varphi = \sigma_\varepsilon^2$. Both cases are possible: (a) the measurement error variance σ_δ^2 is known and (b) the ratio $\sigma_\varepsilon^2/\sigma_\delta^2$ is known; for the latter case see Shklyar (2008). In the particular case of $k = 1$, we obtain the *linear model*; an overview of methods in this MEM is given in Cheng and Van Ness (1999).

In the *loglinear Poisson model* we have $y \sim Po(\lambda)$ with $\lambda = \exp(\beta_0 + \beta_1 \xi)$; then $\eta = \log \lambda,\ c(\eta) = e^\eta$, and $\varphi = 1$.

Methods of Consistent Estimation in Classical MEM

Now, we deal with the general model (1) – (3). We distinguish between two types of estimators, functional and structural. The latter makes use the distribution of ξ, which therefore must be given, at least up to the unknown parameter, vector γ. The former does not need the distribution of ξ and works even when ξ is not random (functional case).

Functional Method: Corrected Score

If the variable ξ were observable, one could estimate β (and also φ) by the method of maximum likelihood (ML). The corresponding likelihood score function for β is given by

$$\psi(y, z, \xi; \beta, \varphi) = \frac{\partial \log f(y|z, \xi; \beta, \varphi)}{\partial \beta} = \frac{y - c'(\eta)}{\varphi} \frac{\partial \eta}{\partial \beta}.$$

We want to construct an unbiased estimating function for β in the observed variables. For this purpose, we need to find functions g_1 and g_2 of z, x, and β such that

$$\mathbf{E}[g_1(z, x; \beta)|z, \xi] = \frac{\partial \eta}{\partial \beta}, \quad \mathbf{E}[g_2(z, x; \beta)|z, \xi] = c'(\eta)\frac{\partial \eta}{\partial \beta}.$$

Then

$$\psi_C(y, z, x; \beta) = y g_1(z, x; \beta) - g_2(z, x; \beta)$$

is termed the corrected score function. The *Corrected Score* (CS) estimator $\hat{\beta}_C$ of β is the solution to

$$\sum_{i=1}^{n} \psi_C(y_i, z_i, x_i; \hat{\beta}_C) = 0.$$

The functions g_1 and g_2 do not always exist. Stefanski (1989) gives the conditions for their existence and shows how to find them if they exist. The CS estimator is consistent in

both functional and structural cases. It was first proposed by Stefanski (1989) and Nakamura (1990).

An alternative functional method, particularly adapted to ▶generalized linear models, is the conditional score method; see Stefanski and Carroll (1987).

Structural Methods: Quasi-Likelihood and Maximum Likelihood

The conditional mean and conditional variance of y given (z, ξ) are, respectively,

$$\mathbf{E}(y|z,\xi) = m^*(z,\xi;\beta) = c'(\eta),\ \mathbf{V}(y|z,\xi) = v^*(z,\xi;\beta) = \varphi c''(\eta).$$

Then the conditional mean and conditional variance of y given the observable variables are

$$m(z,x;\beta) = \mathbf{E}(y|z,x) = E[m^*(z,\xi;\beta)|x],$$
$$v(z,x;\beta) = \mathbf{V}(y|z,x) = \mathbf{V}[m^*(z,\xi;\beta)|x] + \mathbf{E}[v^*(z,\xi;\beta)|x].$$

For the quasi-likelihood (QL) estimator, we construct the quasi-score function

$$\psi_Q(y,z,x;\beta) = [y - m(z,x;\beta)]v(z,x;\beta)^{-1}\frac{\partial m(z,x;\beta)}{\partial \beta}.$$

Here we drop the parameter φ considering it to be known. We also suppress the nuisance parameter γ in the argument of the functions m and v, although m and v depend on γ. Indeed, in order to compute m and v, we need the conditional distribution of ξ given x, which depends on the distribution of ξ with its parameter γ. For instance, assume (5) where the elements of μ_ξ and Σ_ξ make up the components of the parameter vector γ. Then $\xi|x \sim N(\mu(x), T)$ with

$$\mu(x) = \mu_\xi + \Sigma_\xi(\Sigma_\xi + \Sigma_\delta)^{-1}(x - \mu_\xi),$$
$$T = \Sigma_\delta - \Sigma_\delta(\Sigma_\xi + \Sigma_\delta)^{-1}\Sigma_\delta.$$

The QL estimator $\hat{\beta}_Q$ of β is the solution to

$$\sum_{i=1}^{n} \psi_Q(y_i, z_i, x_i; \hat{\beta}_C) = 0.$$

The equation has a unique solution for large n, but it may have multiple roots if n is not large. Heyde and Morton (1998) develop methods to deal with this case.

Maximum likelihood is based on the conditional joint density of x, y given z. Thus, while QL relies only on the error-free mean and variance functions, ML relies on the whole error-free model distribution. Therefore, ML is more sensitive than QL with respect to a potential model misspecification because QL is always consistent as long as at least the mean function (along with the density of ξ) has been correctly specified. In addition, the likelihood function is generally much more difficult to compute than the quasi-score function. This often justifies the use of the relatively less efficient QL instead of the more efficient ML method.

Efficiency Comparison

For CS and QL, $\hat{\beta}$ is asymptotically normal with asymptotic covariance matrix (ACM) Σ_C and Σ_Q, respectively. In the structural model, it is natural to compare the relative efficiencies of $\hat{\beta}_C$ and $\hat{\beta}_Q$ by comparing their ACMs. In case there are no nuisance parameters, it turns out that

$$\Sigma_C \geq \Sigma_Q \tag{6}$$

in the sense of the Loewner order for symmetric matrices. Moreover, under mild conditions the strict inequality holds.

These results hold true if the nuisance parameters γ are known. If, however, they have to be estimated in advance, (6) need not be true anymore. For the Poisson and polynomial structural models, Kukush et al. (2007) prove that (6) still holds even if the nuisance parameters are preestimated. Recently Kukush et al. (2009) have shown that QL can be modified so that, in general, $\Sigma_C \geq \Sigma_Q$; for this purpose the γ must be estimated together with β and not in advance.

Estimation in Berkson Model

Now, we deal with the model (1), (2), and (4). Substituting x for ξ in the regression model (1) – (2) is equivalent to RC. Therefore, it leads to estimates with a typically small bias.

A more precise method is ML. The conditional joint density of x and y given z has a simpler form compared with the classical MEM. That is why ML is more reliable in the Berkson model.

Nonparametric Estimation

We mention two nonparametric problems overviewed in Carroll et al. (2006), Ch. 12: the estimation of the density ρ of a random variable ξ, and the nonparametric estimation of a regression function f, both when ξ is measured with error. In these problems under normally distributed measurement error, the best mean squared error of an estimator of $\rho(x_0)$ or $f(x_0)$ converges to 0 at a rate no faster than the exceedingly slow rate of logarithmic order. However, under a more heavy-tailed measurement error, estimators can perform well for a reasonable sample size.

About the Author

Dr. Alexander Kukush is a Professor, Department of Mechanics and Mathematics, National Taras Shevchenko University of Kyiv, Ukraine. He is an Elected member of the International Statistical Institute (2004). He has authored and coauthored more than 100 papers on statistics and a book: *Theory of Stochastic Processes With Applications to Financial Mathematics and Risk Theory* (with D. Gusak, A. Kulik, Yu. Mishura, and A. Pilipenko, Problem Books in Mathematics, Springer, 2009). Professor Kukush has received the Taras Shevchenko award for a cycle of papers on regression (National Taras Shevchenko University of Kyiv, 2006).

Cross References

►Astrostatistics
►Bias Analysis
►Calibration
►Estimation
►Likelihood
►Linear Regression Models
►Nonparametric Estimation
►Normal Distribution, Univariate
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Probability Theory: An Outline

References and Further Reading

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models, 2nd edn. Chapman and Hall, London

Cheng CL, Van Ness JW (1999) Statistical regression with measurement error. Arnold, London

Heyde CC, Morton R (1998) Multiple roots in general estimating equations. Biometrika 85:967–972

Kukush A, Malenko A, Schneeweiss H (2007) Comparing the efficiency of estimates in concrete errors-in-variables models under unknown nuisance parameters. Theor Stoch Proc 13(29):4, 69–81

Kukush A, Malenko A, Schneeweiss H (2009) Optimality of the quasi score estimator in a mean-variance model with applications to measurement error models. J Stat Plann Infer 139:3461–3472

Nakamura T (1990) Corrected score functions for errors-in-variables models: Methodology and application to generalized linear models. Biometrika 77:127–137

Shklyar SV (2008) Consistency of an estimator of the parameters of a polynomial regression with a known variance relation for errors in the measurement of the regressor and the echo. Theor Probab Math Stat 76:181–197

Stefanski LA (1989) Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. Commun Stat A – Theor 18:4335–4358

Stefanski LA, Carroll RJ (1987) Conditional scores and optimal scores in generalized linear measurement error models. Biometrika 74:703–716

Measurement of Economic Progress

Marat Ibragimov[1], Rustam Ibragimov[2]
[1]Associate Professor
Tashkent State University of Economics, Tashkent, Uzbekistan
[2]Associate Professor
Harvard University, Cambridge, MA, USA

Broadly defined, measurement of economic progress focuses on quantitative analysis of the standard of living or quality of life and their determinants. The analysis concerns many elements of the standard living such as its material components, human capital, including education and health, inequality and other factors [see, among others, Barro and Sala-i Martin (2004), Howitt and Weil (2008), Steckel (2008), and references therein].

Theoretical foundation for empirical analysis of determinants of economic growth is provided by the Solow growth model. The human capital-augmented version of the model with the Cobb-Douglas production function [see Mankiw et al. (1992)] assumes that, for country i at time t, the aggregate output $Y_i(t)$ satisfies $Y_i(t) = K_i(t)^{\alpha} H_i(t)^{\beta} (A_i(t)L_i(t))^{1-\alpha-\beta}$, where $K_i(t)$ is physical capital, $H_i(t)$ is human capital, $L_i(t)$ is labor supply and $A_i(t)$ is a productivity parameter (the efficiency level of each worker or the level of technology). The variables L and A are assumed to obey $L_i(t) = L_i(0)e^{n_i t}$ and $A(t) = A(0)e^{gt}$, where n_i and g are, respectively, the population growth rate and the rate of technological progress. Physical and human capital are assumed to follow continuous-time accumulation equations $dK_i(t)/dt = s_{K,i}Y_i(t) - \delta K_i(t)$ and $dH_i(t)/dt = s_{H,i}Y_i(t) - \delta H(t)$ with the depreciation rate δ and the savings rates $s_{K,i}$ and $s_{H,i}$. Under the above assumptions, the growth model leads to the regressions $\gamma_i = a_0 + a_1 \log y_i(0) + a_2 \log(n_i + g + \delta) + a_3 \log s_{K,i} + a_4 \log s_{H,i} + \epsilon_i$, where $\gamma_i = (\log y_i(t) - \log y_i(0))/t$ is the growth rate of output per worker $y_i(t) = Y_i(t)/L_i(t)$ between time 0 and t [see, among others, Barro and Sala-i Martin (2004), Durlauf et al. (2005)]. Cross-country growth regressions typically include additional regressors Z_i and focus on estimating models in the form $\gamma_i = \mathbf{a}\mathbf{X}_i + \mathbf{b}\mathbf{Z}_i + \epsilon_i$, where $\mathbf{a} = (a_0, a_1, ..., a_4) \in \mathbf{R}^5$, $\mathbf{b} = (b_1, b_2, ..., b_m) \in \mathbf{R}^m$, the components of $\mathbf{X}_i = (1, \log y_i(0), \log(n_i + g + \delta), \log s_{K,i}, \log s_{H,i})'$ are the growth determinants in the Solow model and $\mathbf{Z}_i \in \mathbf{R}^m$ is the vector of growth determinants outside the Solow growth theory.

The statistical analysis of economic progress and its determinants presents a number of challenges due to

the necessity of using proxy measures and corresponding weights for different components of the standard of living and factors affecting it. The material standard of living is typically measured as per capita Gross Domestic Product (GDP) adjusted for changes in price levels. Proxies for education and human capital used in growth economics include school-enrollment rates at the secondary and primary levels, literacy rates, average years of secondary and higher schooling and outcomes on internationally comparable examinations. Many works in the literature have also used student-teacher ratios as a measure of quality of education. The two most widely used measures of health are life expectancy at birth or age 1 and average height used as a proxy for nutritional conditions during the growing years.

Barro (1991) and Barro and Sala-i Martin (2004) find that the growth rate of real per capita GDP is positively related to initial human capital, including education and health, proxied by school-enrollment rates, upper-level schooling and life expectancy and negatively related to the initial level of real per capita GDP. The results in Barro (1991) also indicate statistically significant negative effects of political instability (measured using the number of revolutions and coups per year and the number of political assassinations per million population per year) on growth. Other factors used in the analysis in Barro (1991) and Barro and Sala-i Martin (2004) include fertility and the ratio of real government consumption to real GDP (with statistically significant negative effects on growth), investment ratio, inflation rate as well as proxies for market distortions, maintenance of the rule of law, measures for democracy, international openness, the terms of trade, indicators for economic systems and countries in sub-Saharian Africa and Latin America and other variables.

A number of works in theoretical and empirical growth economics have focused on the development and analysis of performance of models with endogenous technological progress. Many recent studies have also studied the factors that lead to the observed differences in the determinants of economic growth in different countries, including capital components, technology and efficiency. In particular, several works have emphasized the role of geographical differences, cultural factors, economic policies and institutions as fundamental causes of the differences in growth determinants (Howitt and Weil 2008).

Statistical study of economic growth determinants is complicated by relatively small samples of available observations, measurement errors in key variables, such as GDP, heterogeneity in observations and estimated parameters, dependence in data and large number of potential growth regressors under analysis. Related issues in the analysis of economic growth concern difficulty of causal interpretation of estimation results, robustness of the conclusions to alternative measures of variables in the analysis, and open-endedness of growth theories that imply that several key factors matter for growth at the same time. Levine and Renelt (1992) focus on the analysis of robustness of conclusions obtained using cross-country growth regressions. They propose assessing the robustness of the variable Z of interest using the variation of the coefficient b in cross-country regressions $\gamma_i = \mathbf{a}\mathbf{X}_i + bZ_i + \mathbf{c}\mathbf{V}_i + \epsilon_i$, where $\mathbf{X}_i$ is the vector of variables that always appear in the regressions (e.g., the investment share of GDP, initial level of income, a proxy for the initial level of human capital such as the school enrollment rate, and the rate of population growth in country i), and $\mathbf{V}_i$ is a vector of additional control variables taken from the pool of variables available. Departing from the extreme bounds approach in Levine and Renelt (1992) that requires the estimate of the coefficient of interest b to be statistically significant for any choice of control variables $\mathbf{V}$, several recent works [see Sala-i Martin et al. (2004), Ch. 12 in Barro and Sala-i Martin (2004), and references therein] propose alternative less stringent procedures to robustness analysis. Several recent works on the analysis of economic growth and related areas emphasize importance of models incorporating disasters and crises and probability distributions generating ►outliers and extreme observations, such as those with heavy-tailed and power-law densities [see Barro (1991), Gabaix (2009) and Ibragimov (2009)].

Acknowledgments

Marat Ibragimov gratefully acknowledges support by a grant R08-1123 from the Economics Education and Research Consortium (EERC), with funds provided by the Global Development Network and the Government of Sweden. Rustam Ibragimov gratefully acknowledges partial support by the National Science Foundation grant SES-0820124.

Cross References

►Composite Indicators
►Econometrics
►Economic Growth and Well-Being: Statistical Perspective
►Economic Statistics

References and Further Reading

Barro RJ (1991) Economic growth in a cross section of coutries. Q J Econ 106:407–443
Barro RJ, Sala-i Martin X (2004) Economic growth. MIT, Cambridge, MA

Durlauf S, Johnson P, Temple J (2005) Growth econometrics. In: Aghion P, Durlauf S (eds) Handbook of economic growth. North-Holland, Amsterdam

Gabaix X (2009) Power laws in economics and finance. Annu Rev Econ 1:255–293

Howitt P, Weil DN (2008) Economic growth. In: Durlauf SN, Blume LE (eds) New palgrave dictionary of economics, 2nd edn. Palgrave Macmillan, Washington, DC

Ibragimov, R (2009) Heavy tailed densities, In: *The New Palgrave Dictionary of Economics Online*, (Eds. S. N. Durlauf and L. E. Blume), Palgrave Macmillan. http://www.dictionaryofeconomics.com/article?id=pde2008_H000191

Levine R, Renelt D (1992) A sensitivity analysis of cross-country growth regressions. Am Econ Rev 82:942–963

Mankiw NG, Romer D, Wiel DN (1992) A contribution to the empirics of economic growth. Q J Econ 42:407–437

Sala-i Martin X, Doppelhofer G, Miller RI (2004) Determinants of long-term growth: A Bayesian averaging of classical estimates (bace) approach. Am Econ Rev 94:813–835

Steckel RH (2008) Standards of living (historical trends). In: Durlauf SN, Blume LE (eds) New palgrave dictionary of economics, 2nd edn. Palgrave Macmillan, Washington, DC

Measurement of Uncertainty

K. R. Muraleedharan Nair
Professor
Cochin University of Science and Technology, Cochin, India

The measurement and comparison of uncertainty associated with a random phenomenon have been a problem attracting a lot of researchers in Science and Engineering over the last few decades. Given a system whose exact description is unknown its ►entropy is the amount of information needed to exactly specify the state of the system. The Shannon's entropy, introduced by Shannon (1948), has been extensively used in literature as a quantitative measure of uncertainty. If $A_1, A_2, \ldots, A_n$ are mutually exclusive events, with respective probabilities $p_1, p_2, \ldots, p_n$, the Shannon's entropy is defined as

$$H_n(P) = -\sum_{i=1}^{n} p_i \log p_i. \tag{1}$$

Earlier development in this area was centered on characterizing the Shannon's entropy using different sets of postulates. The classic monographs by Ash (1965), Aczel and Daroczy (1975) and Behra (1990) review most of the works on this aspect. Another important aspect of interest is that of identifying distributions for which the Shannon's entropy is maximum subject to certain restrictions on the underlying random variable. Depending on the conditions imposed, several maximum entropy distributions have been derived. For instance, if X is a random variable in the support of the set of non-negative real numbers, the maximum entropy distribution under the condition that the arithmetic men is fixed is the exponential distribution. The book by Kapur (1989) covers most of the results in this area.

For a continuous non-negative random variable X with probability density function $f(x)$ the continuous analogue of (1) takes the form

$$H(f) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx. \tag{2}$$

Several modifications of the Shannon's entropy has been proposed and extensively studied. Renyi (1961) define the entropy of order α as

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \frac{\sum_{i=1}^{n} p_i^\alpha}{\sum_{i=1}^{n} p_i}, \ \alpha \neq 1, \alpha > 0 \tag{3}$$

where $P = (P_{1,\ldots\ldots}P_n)$ is such that $p_i \geq 0$, *and* $\sum_{i-1}^{n} p_i = 1$.

As $\alpha \to 1$, (3) reduces to (1). Khinchin (1957) generalized the Shannon's entropy by choosing a convex function $\varphi(.)$, with $\varphi(1) = 0$ and defined the measure

$$H_\varphi(f) = -\int_{-\infty}^{\infty} f(x) \varphi[f(x)] dx. \tag{4}$$

Nanda and Paul (2006) studied (4) for two particular choices of φ in the form

$$H_1^\beta(f) = \frac{1}{\beta - 1}\left[1 - \int_0^\alpha f^\beta(x) dx\right] \tag{5}$$

and

$$H_2^\beta(f) = \frac{1}{1-\beta}\left[\log \int_0^\infty f^\beta(x) dx\right] \tag{6}$$

where the support of f is the set of non-negative reals and $\beta > 0$ with $\beta \neq 1$. As $\beta \to 1$, (5) and (6) reduces to the Shannon's entropy given in (2).

Recently Rao et al. (2004) introduced cumulative residual entropy defined by

$$E(X) = -\int_0^\infty \overline{F}(x) \log \overline{F}(x) dx$$

which is proposed as an alternative measure of uncertainty based on the cumulative survival function $\overline{F}(x) = P(X > x)$. For various properties and applications of this measure we refer to Rao (2005) and Asadi and Zohrevand (2007).

There are several other concepts closely related to the Shannon's entropy. Kullback and Leibler (1951) defines the directed divergence (also known as relative entropy or cross entropy) between two distributions $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$ with

$$p_i, q_i \geq 0 \sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$$

as

$$D_n(P,Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}. \tag{7}$$

Kannappan and Rathie (1973) and Mathai and Rathie (1975) have obtained characterization results based on certain postulates which naturally leads to (7). The continuous analogue of (7) turns out to be

$$D(f,g) = \int_{-\infty}^{\alpha} f(x) \log \frac{f(x)}{g(x)} dx \tag{8}$$

where $f(x)$ and $g(x)$ are probability density functions corresponding to two probability measures P and Q.

The concept of affinity between two distributions was introduced and studied in a series of works by Matusita [see Matusita (1961)]. This measure has been widely used as a useful tool for discrimination among distributions. Affinity is symmetric in distributions and has direct relationship with error probability when classification or discrimination is concerned. For two discrete distributions P and Q considered above the Matusita's affinity (Mathai and Rathie 1975) between P and Q is defined as

$$\delta(P,Q) = \sum_{i=1}^{n} (p_i q_i)^{1/2}. \tag{9}$$

If X and Y are non-negative random variables and if $f(x)$ and $g(x)$ are the corresponding probability density functions, the affinity between f and g takes the form

$$\delta(f,g) = \int_0^{\infty} \sqrt{f(x)g(x)} dx \tag{10}$$

$\delta(f,g)$ lies between 0 and 1.

Majernik (2004) has shown that

$$H(f,g) = 2[1 - \delta(f,g)]$$

where $H(f,g)$ is the Hellinger's distance defined by

$$H(f,g) = \int_0^{\infty} \left[\sqrt{f(x)} - \sqrt{g(x)}\right]^2 dx. \tag{11}$$

Affinity is a special case of the Chernoff distance considered in Akahira (1996) defined by

$$C(F,G) = -\log\left[\int f^{\alpha}(x) g^{1-\alpha} dx\right], 0 < \alpha < 1. \tag{12}$$

It may be noticed that when $\alpha = \frac{1}{2}$ (12) reduces to $-\log \delta(f,g)$, where $\delta(f,g)$ is the affinity defined in (10).

The concept of inaccuracy was introduced by Kerridge (1961). Suppose that an experimenter asserts that the probability for the i^{th} eventuality is q_i whereas the true probability is p_i, then the inaccuracy of the observer, as proposed by Kerridge, can be measured by

$$1(P,Q) = -\sum_{i=1}^{n} p_i \log q_i \tag{13}$$

where P and Q are two discrete probability distributions, considered earlier.

Nath (1968) extended the Kerridge's concept to the continuous situation. If $F(x)$ is the actual distribution function corresponding to the observations and $G(x)$ is the distribution assigned by the experimenter and $f(x)$ and $g(x)$ are the corresponding density functions the inaccuracy measure is defined as

$$1(F,G) = -\int_0^{\alpha} f(x) \log g(x) dx. \tag{14}$$

This measure has extensively been used as a useful tool for measurement of error in experimental results. In expressing statements about probabilities of various events in an experiment, two kinds of errors are possible: one resulting from the lack of enough information or vagueness in experimental results and the other from incorrect information. In fact, (14) can be written as

$$1(F,G) = -\int_0^{\infty} f(x) \log f(x) dx + \int_0^{\infty} f(x) \log \frac{f(x)}{f(x)} dx. \tag{15}$$

The first term on the right side of (15) represents the error due to uncertainty which is the Shannon's entropy while the second term is the Kullback–Leibler measure, defined in (8) representing the error due to wrongly specifying the distribution as $G(x)$. In this sense the measure of inaccuracy can accommodate the error due to lack of information as well as that due to incorrect information.

In many practical situations, complete data may not be observable due to various reasons. For instance, in lifetime studies the interest may be on the life time of a unit after a specified time, say t. If X is the random variable representing the life time of a component the random variable of interest is $X - t | X > t$. Ebrahimi (1996) defines the residual entropy function as the Shannon's entropy associated with the residual life distribution, namely

$$H(f,t) = -\int_t^{\infty} \frac{f(x)}{\overline{F}(t)} \log \frac{f(x)}{\overline{F}(x)}, \overline{F}(t) > 0. \tag{16}$$

In terms of the hazard rate $h(x) = \frac{f(x)}{F(x)}$, (16) can also be written as

$$H(f,t) = 1 - \frac{1}{\bar{F}(t)} \int_t^\infty f(x) \log h(x) dx. \qquad (17)$$

Ebrahimi points out that (16) can be used as a potential measure of stability of components in the reliability context. The problem of ordering life time distributions using this concept has been addressed in Ebrahimi and Kirmani (1996). Belzunce et al. (2004) has shown that the residual entropy function determines the distributions uniquely if $H(f,t)$ is increasing in t. Characterization of probability distributions using the functional form of the residual entropy function have been the theme addressed in Nair and Rajesh (1998), Sankaran and Gupta (1999), Asadi and Ebrahimi (2000) and Abraham and Sankaran (2005).

Recently Nanda and Paul (2006) has extended the definition of the Renyi entropy defined by (5) and (6) to the truncated situation. It is established that under certain conditions the Renyi's residual entropy function determines the distribution uniquely. They have also looked into the problem of characterization of probability distributions using the same.

Ebrahimi and Kirmani (1996) has modified the definition of the Kullback–Leibler measure to the truncated situation to accommodate the current age of a system. Recently Smitha et al. (2008) have extended the definition of affinity to the truncated situation and has obtained characterization results for probability distributions under the assumption of proportional hazard model. Nair and Gupta (2007) extended the definition of the measure of inaccuracy to the truncated situation and has characterized the generalized Pareto distributions using the functional form of the inaccuracy measure.

About the Author

Dr. K.R. Muraleedharan Nair is a senior Professor in the Department of Statistics of the Cochin University of Science and Technology, India. He had been teaching Statistics at the post graduate level for the past 39 years. He has served the University as the Head of the Department (2004–2007) and as the Controller of examinations (2000–2003). He is currently the Vice President of the Indian Society for Probability and Statistics, besides being reviewer for certain reputed journals. He has published 28 papers in international journals besides several conference papers. He is a member of the Board of Studies as well as Faculty of Science in some of the Indian Universities.

Cross References

►Diversity
►Entropy
►Entropy and Cross Entropy as Diversity and Distance Measures
►Kullback-Leibler Divergence
►Maximum Entropy Method for Estimation of Missing Data
►Probability Theory: An Outline
►Role of Statistics
►Statistical View of Information Theory

References and Further Reading

Abraham B, Sankaran PG (2005) Renyi's entropy for residual lifetime distributions, Stat Papers 46:17–30
Aczel J, Daroczy Z (1975) On measures of information and their characterization, Academic, New York. Ann Inst Stat Math 48:349–364
Akahira M (1996) Loss of information of a statistic for a family of non-regular distributions. Ann Inst Stat Math 48:349–364
Asadi M, Ebrahimi N (2000) Residual entropy and its characterizations in terms of hazard function and mean residual life function. Stat and Prob Letters 49:263–269
Asadi M, Zohrevand Y (2007) On the dynamic cumulative residual entropy. J Stat Plann Infer 137:1931–1941
Ash RB (1965) Information theory. Wiley, New York
Behra M (1990) Additive and non-additive measures of entropy. Wiley Eastern, New York
Belzunce F, Navarro J, Ruiz JM, del Aguila Y (2004) Some results on residual entropy function. Metrica 59:147–161
Ebrahimi N (1996) How to measure uncertainty in the residual life time distribution. Sankhya A 58:48–56
Ebrahimi N, Kirmani SUNA (1996) Some results on ordering survival function through uncertainty. Stat Prob Lett 29:167–176
Kannappan PI, Rathie PN (1973) On characterization of directed divergence. Inform Control 22:163–171
Kapur JN (1989) Maximum entropy models in science and engineering. Wiley Eastern, New Delhi
Kerridge DF (1961) Inaccuracy and inference. J R Stat Soc Series B, 23:184–194
Khinchin AJ (1957) Mathematical foundation of information theory. Dover, New York
Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86
Majernik K (2004) A dissimilarity measure for an arbitrary number of probability distributions. Int J Gen Sys 33(6):673–678
Mathai AM, Rathie PN (1975) Basic concepts in information theory and statisticsaxiomatic foundations and applications. Wiley, New York
Matusita K (1961) Interval estimation based on the notion of affinity. Bull Int Stat Inst 38(4):241–244
Nanda AK, Paul P (2006) Some results on generalized residual entropy. Inform Sci 176:27–47
Nair KRM, Rajesh G (1998) Characterization of probability distribution using the residual entropy function. J Ind Stat Assoc 36:157–166
Nair NU, Gupta RP (2007) Characterization of proportional hazard models by properties of information measures. Int J Stat 6(Special Issue):223–231
Nath P (1968) Inaccuracy and coding theory. Metrika 13:123–135
Rajesh G, Nair KRM (1998) Residual entropy function in discrete time. Far East J Theor Stat 2(1):1–10

Rao M, Chen Y, Vemuri BC, Wang F (2004) Cumulative residual entropy: a new measure of information. IEE Trans Inform Theor 50(6):1220–1228
Rao M (2005) More on a concept of entropy and information. J Theor Probab 18:967–981
Renyi A (1961) On measures of entropy and information, Proceedings of Fourth Berkley Symposium on Mathematics, Statistics and Probability, 1960, University of California Press, vol 1, pp 547–561
Sankaran PG, Gupta RP (1999) Characterization of life distributions using measure of uncertainty. Cal Stat Assoc Bull 49:154–166
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 279–423:623–656
Smitha S, Nair KRM, Sankaran PG (2008) On measures of affinity for truncated distribution. Cal Stat Assoc Bull 59:151–162

Measures of Agreement

Elisabeth Svensson
Örebro University, Örebro, Sweden

Agreement in repeated assessments is a fundamental requirement for quality of data from assessments on ►rating scales. Scale assessments produce ordinal data, the ordered categories representing only a rank order of the intensity of a particular variable and not a numerical value in a mathematical sense, even when the assessments are numerically labeled.

The main quality concepts of scale assessments are *reliability* and *validity*. *Reliability* refers to the extent to which repeated measurements of the same object yield the same result, which means agreement. In *intra-rater reliability* studies the agreement in test-retest assessments is evaluated. *Inter-rater reliability* refers to the level of agreement between two raters judging the same object.

The *percentage agreement* (*PA*) in assessments is the basic agreement measure and is also called *overall agreement* or *raw agreement*. When $PA < 100\%$ the reasons for disagreement can be evaluated by a statistical approach by Svensson that takes account of the rank-invariant properties of ordinal data. The approach makes it possible to identify and measure systematic disagreement, when present, separately from disagreement caused by individual variability in assessments. Different frequency distributions of the two sets of ordinal assessments indicate that the two assessments disagree systematically regarding the use of the scale categories. When higher categories are more frequently used in one set of assessments, X, than in the other, Y, there is a systematic disagreement in position. The measure *Relative Position, RP,* estimates the parameter of a systematic disagreement in position defined by $\gamma = P(X < Y) - P(Y < X)$.

A systematic disagreement in how the two assessments are concentrated to the scale categories is measured by the *Relative Concentration, RC,* estimating the parameter of a systematic shift in concentration $\delta = P(X_{l_1} < Y_k < X_{l_2}) - P(Y_{l_1} < X_k < Y_{l_2})$.

The measure of individual variability, the relative rank variance, $0 \le RV \le 1$ is defined $RV = \frac{6}{n^3}\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}\left[\overline{R}_{ij}^{(X)} - \overline{R}_{ij}^{(Y)}\right]^2$ where $\overline{R}_{ij}^{(X)}$ is the mean augmented rank of the observations in the ijth cell of an $m \times m$ square contingency table according to the assessments X. In the aug-rank approach $\overline{R}_{i,j-1}^{(X)} < \overline{R}_{i,j}^{(X)}$ and $\overline{R}_{i-1,j}^{(Y)} < \overline{R}_{i,j}^{(Y)}$. $RV = 0$ means that the observed disagreement is completely explained by the measures of systematic disagreement. In that case the two sets of aug-ranks are equal and the paired distribution is the *rank-transformable pattern of agreement* (see ►Ranks).

The advantage of separating the observed disagreement in the components of systematic and individual disagreements is that it is possible to improve the rating scales and/or the users of the scale. Systematic disagreement is population based and reveals a systematic change in conditions between test-.retest assessments or that raters interpret the scale categories differently. Large individual variability is a sign of poor quality of the rating scale as it allows for uncertainty in repeated assessments.

The Cohen's *coefficient kappa* (κ) is a commonly used measure of agreement adjusted for the chance expected agreement. There are limitations with kappa. The maximum level of kappa, $\kappa = 1$, requires equally skilled raters, in other words lack of systematic disagreement (bias). The value of weighted kappa depends on the choice of weights, and the weighting procedure ignores the rank-invariant properties of ordinal data. The kappa value increases when the number of categories decreases, and depends also on how the observations are distributed on the different categories, the prevalence. Therefore kappa values from different studies are not comparable.

The calculations of Cronbach's alfa and other so-called reliability coefficients are based on the assumption of quantitative, normally distributed data, which is not achievable in data from rating scales.

There is also a widespread misuse of correlation in reliability studies. The correlation coefficient measures the *degree of association* between two variables and does not measure the level of agreement, see Fig. 1. The PA is 12%, and the observed disagreement is mainly explained by a systematic disagreement in position. The negative RP value

A. The observed pattern							B. The rank-transformable pattern of agreement					
X / Y	C_1	C_2	C_3	C_4	total		X / Y	C_1	C_2	C_3	C_4	total
C_4			1	1	2		C_4				2	2
C_3		2	2	14	18		C_3			1	17	18
C_2	1	1	11	3	16		C_2			16		16
C_1	2	8	3	1	14		C_1	3	11			14
total	3	11	17	19	50			3	11	17	19	50

Measures of Agreement. Fig. 1 The frequency distribution of 50 pairs of assessments on a scale with four ordered categories, $C_1 < C_2 < C_3 < C_4$ and the corresponding rank-transformable pattern of agreement, defined by the marginal distributions

(-0.48) and the constructed RTPA shows that the assessments Y systematically used a lower category than did X. A slight additional individual variability, $RV = 0.08$ is observed. The Spearman rank-order correlation coefficient is 0.66 in A and 0.97 in B, ignoring the fact that the assessments are systematically biased and unreliable. The same holds for the coefficient kappa (-0.14).

About the Author

For biography *see* the entry ►Ranks.

Cross References

►Kappa Coefficient of Agreement
►Ranks
►Rating Scales

References and Further Reading

Svensson E (1997) A coefficient of agreement adjusted for bias in paired ordered categorical data. Biometrical J 39:643–657
Svensson E (1998) Application of a rank-invariant method to evaluate reliability of ordered categorical assessments. J Epidemiol Biostat 3(4):403–409

Measures of Dependence

Reza Modarres
Head and Professor of Statistics
The George Washington University, Washington, DC, USA

Let X and Y be continuous random variables with joint distribution function (DF) H and marginal DFs F and G. Three well-known measures of dependence are

1. Pearson's correlation:
$$\rho = \frac{1}{\sigma_X \sigma_Y} Cov(X, Y) = \frac{1}{\sigma_X \sigma_Y} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [H(x,y) - F(x)G(y)]dxdy$$
where σ_x, σ_y and $Cov(X, Y)$ are the standard deviations and covariance of X and Y, respectively
2. Spearman's correlation: $s = 12 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [H(x,y) - F(x)G(y)]dF(x)dG(y)$,
3. Kendall's correlation: $\tau = 4 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x,y)dH(x,y) - 1$

Pearson correlation measures the strength of linear relationship between X and Y and has well-studied theoretical properties. However, it can be unduly influenced by ►outliers, unequal variances, non-normality, and nonlinearity. Spearman's correlation reflects the monotone association between X and Y and measures the correlation between $F(X)$ and $G(Y)$. Kendall's correlation is the probability of concordance minus the probability of discordance. Spearman's and Kendall's correlations remain invariant under a monotone transformation. However, Pearson's correlation remains only invariant under a location and scale change.

Using the probability integral transformations $u = F(x)$ and $v = G(y)$, the copula (see also ►Copulas) of X and Y is defined as $C(u, v) = H(F^{-1}(u), G^{-1}(v))$. Hence,

$$\rho = \frac{1}{\sigma_X \sigma_Y} \iint_{I^2} [C(u,v) - uv]dF^{-1}(u)dG^{-1}(v),$$

$$s = 12 \iint_{I^2} [C(u,v) - uv]dudv,$$

$$\tau = 4 \iint_{I^2} C(u,v)dC(u,v) - 1$$

where I^2 is the unit square. Schweizer and Wolff (1981) note that $C(u, v) - uv$ is the signed volume between the surface $z = C(u, v)$ and $Z = uv$ (the independence copula).

Copula representation of ρ clearly shows its dependence on the marginal distributions. Therefore, it is not a measure of nonparametric dependence. Daniels (1950) shows that $-1 \le 3\tau - 2s \le 1$. Nelsen (1991) studies the relationship between s and τ for several families of copulas and Fredricks and Nelsen (2007) show that the ratio τ/s approaches 2/3 as H approaches independence.

Hoeffding (1940) and Frechét (1951) show that for all $(x,y) \in R^2$ the joint DF is bounded: $H_1(x,y) \le H(x,y) \le H_2(x,y)$ where $H_1(x,y) = max(0, F(x) + G(y) - 1)$ and $H_2(x,y) = min(F(x), G(y))$ are distribution functions. Perfect negative correlation is obtained when H_1 is concentrated on the line $F(x) + G(y) = 1$ whereas perfect positive correlation is obtained when H_2 is concentrated on the line $F(x) = G(y)$. In fact, $H_0(x,y) = F(x)G(y)$ for all $(x,y) \in R^2$ reflects independence of X and Y. Let $C_1(x,y) = max(0, u + v - 1)$, $C_2(x,y) = min(u,v)$ and $C_0(x,y)$ denote the Frechét lower, upper and independence copulas, respectively. Similarly, $C_1(u,v) \le C(u,v) \le C_2(u,v)$.

Using Hoeffding lemma (1948)

$$Cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[H(x,y) - F(x)G(y)]dxdy,$$

one can show $\rho_1 \le \rho \le \rho_2$ where ρ_1 and ρ_2 are the correlation coefficients associated with H_1 and H_2, respectively. Depending on the marginal distributions the range of ρ may be much smaller than $|\rho| \le 1$. For example, for the bivariate log-normal distribution with unit variances, one can show $\rho \in (-0.368, 1)$. Lancaster (1958) uses Chebyshev-Hermite polynomial to obtain the correlation coefficient of transformed bivariate random vectors. Freeman and Modarres (2005) obtain the form of the correlation after a ►Box-Cox transformation.

Moran (1967) states that the necessary and sufficient conditions for ρ to assume extreme values of +1 and −1 are

1. $X \overset{d}{=} aY + b$ for constants
2. $F(\mu + x) = 1 - F(\mu - x)$ where μ is the mean of X. Normal, uniform, double exponential and logistic distributions satisfy these conditions

Rényi (1959) considers a set of conditions that a symmetric nonparametric measure of dependence should satisfy. Schweizer and Wolff (1981) note that Rényi's conditions are too strong and suggest that any suitably normalized distance measure such as the L_p distance provides a symmetric measure of nonparametric dependence. They show that these distances, according to a modified set of Rényi conditions, enjoy many useful properties. Let $L_p = (K_p \int_{I^2} |C(u,v) - uv|^p dudv)^{1/p}$ where K_p is chosen such that L_p remains in $(0,1)$. We have

1. $L_1 = 12 \int_{I^2} |C(u,v) - uv| dudv$
2. $L_2 = \left(90 \int_{I^2} (C(u,v) - uv)^2 dudv\right)^{1/2}$
3. $L_\infty = 4\, Sup_{I^2} |C(u,v) - uv|$

In fact Hoeffding (1948) and Blum et al. (1961) base a nonparametric test of independence between X and Y on L_∞. Modarres (2007) studies several tests of independence, including a measure based on the likelihood of cut-points.

About the Author

Dr. Reza Modarres is a Professor and Head, Department of Statistics, George Washington University, Washington DC. He is an elected member of International Statistical Society. He has authored and co-authored more than 50 papers and is on the editorial board of several journals.

Cross References

►Bivariate Distributions
►Copulas: Distribution Functions and Simulation
►Correlation Coefficient
►Kendall's Tau
►Statistics on Ranked Lists
►Tests of Independence

References and Further Reading

Blum JR, Kiefer J, Rosenblatt M (1961) Distribution free tests of independence based on the sample distribution function. Ann Math Stat 32:485–498
Daniels HE (1950) Rank correlation and population models. J R Stat Soc B 12:171–181
Frechét M (1951) Sur les tableaux de corrélation dont les marges sont données. Ann Univ Lyon Sec A 14:53–57
Fredricks GA, Nelsen RB (2007) On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. J Stat Plan Infer 137:2143–2150
Freeman J, Modarres R (2005) Efficiency of test for independence after Box-Cox transformation. J Multivariate Anal 95: 107–118
Hoeffding W (1940) Masstabinvariante korrelations-theorie. Schriften Math Inst Univ Berlin 5:181–233
Hoeffding W (1948) A nonparametric test of independence. Ann Math Stat 19:546–557
Lancaster HO (1958) The structure of bivariate distributions. Ann Math Stat 29:719–736
Modarres R (2007) A test of independence based on the likelihood of cut-points. Commun Stat Simulat Comput 36:817—825
Moran PAP (1967) Testing for correlation between non-negative variates. Biometrika 54:385–394
Nelsen RB (1991) Copulas and association. In: Dall'Aglio G, Kotz S, Salinetti G (eds) Advances in probability distributions with given marginals. beyond copulas. Kluwer Academic, London
Rényi A (1959) On measures of dependence. Acta Math Acad Sci Hunger 10:441–451
Schweizer B, Wolff EF (1981) On nonparametric measures of dependence for random variables. Ann Stat 9(4):879–885

Median Filters and Extensions

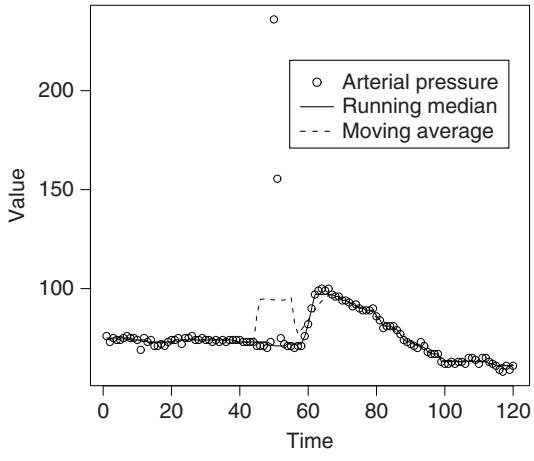
Roland Fried[1], Ann Cathrice George[2]
[1]Professor
TU Dortmund University, Dortmund, Germany
[2]TU Dortmund University, Dortmund, Germany

De-noising a time series, that is a sequence of observations of a variable measured at equidistant points in time, or an image, that is a rectangular array of pixels, is a common task nowadays. The objective is to extract a varying level (a "signal") representing the path followed by the time series or the true image which is overlaid by irrelevant noise.

Linear filters like moving averages are computationally simple and eliminate normal noise efficiently. However, their output is heavily affected by strongly deviating observations (called ▶outliers, spikes or impulses), which can be caused for instance by measurement artifacts. Moreover, linear filters do not preserve abrupt changes (also called step changes or jumps) in the signal or edges in an image. Tukey (1977) suggests median filters, also called running medians, for these purposes.

We focus on the time series setting in the following. Let $y_1, \ldots, y_N$ be observations of a variable at equidistant points in time. De-noising these data for extraction of the time-varying mean level underlying these data (the signal) can be accomplished by moving a time window $y_{t-k}, \ldots, y_t, \ldots, y_{t+k}$ of length $n = 2k+1$ through the series for estimation of the level μ_t in the center of the window. Whereas a moving average calculates the arithmetic average of the data in the time window for this, a running median uses the median of these values. If the window width is fixed throughout, we get estimates of the levels $\mu_{k+1}, \ldots, \mu_{N-k}$ at instances not very close to the start or the end of the time series. The levels at the start or the end of the time series can be estimated for instance by extrapolation of the results from the first and last window or by adding the first and the last observed value a sufficient number of times.

Figure 1 depicts observations of the arterial blood pressure of a patient in intensive care measured once a minute, as well as the outputs of a moving average and a running median, both with window width $n = 11$. The moving average is strongly affected by a few measurement artifacts, and it smooths the sudden increase at $t = 60$. The running median eliminates the spikes and preserves the shift.



Median Filters and Extensions. Fig. 1 Measurements of the arterial blood pressure of a patient and outputs of a running median and a moving average, both with window width $n = 11$

A possible disadvantage of running medians is that they implicitly rely on the assumption that the level is almost constant within each time window. While increasing the window width improves the reduction of noise if the signal is locally constant, this is no longer the case in trend periods. Davies et al. (2004) investigate application of robust regression to a moving time window to improve the approximation of trends in the presence of ▶outliers. Many further refinements of robust filters for signal extraction from time series or images and different rules for choosing a (possibly locally adaptive) window width from the data have been suggested in the literature. See Gather et al. (2006) for an overview on robust signal extraction from time series.

Cross References

▶Moving Averages
▶Outliers
▶Smoothing Techniques
▶Statistical Signal Processing
▶Time Series

References and Further Reading

Davies L, Fried R, Gather U (2004) Robust signal extraction for online monitoring data. J Stat Plan Infer 122:65–78

Gather U, Fried R, Lanius V (2006) Robust detail-preserving signal extraction. In: Schelter B, Winterhalder M, Timmer J (eds) Handbook of time series analysis. Wiley, New York, pp. 131–158

Tukey JW (1977) Exploratory data analysis (preliminary edition 1971). Addison-Wesley, Reading MA

Medical Research, Statistics in

B. S. Everitt
Professor Emeritus
Institute of Psychiatry, King's College, London, UK

Statistical science plays an important role in medical research. Indeed a major part of the key to the progress in medicine from the 17th century to the present day has been the collection and valid interpretation of empirical evidence provided by the application of statistical methods to medical studies. And during the last few decades, the use of statistical techniques in medical research has grown more rapidly than in any other field of application. Indeed, some branches of statistics have been especially stimulated by their applications in medical investigations, notably the analysis of ►survival data (see, for example, Collett 2003). But why has statistics (and statisticians) become so important in medicine? Some possible answers are:

- Medical practice and medical research generate large amounts of data. Such data can be full of uncertainty and variation and extracting the "signal," i.e. the substantive medical message in the data, form the 'noise' is usually anything but trivial.
- Medical research often involves asking questions that have strong statistical overtones, for example: 'How common is a particular disease?'; 'Which people have the greatest chance of contracting some condition or other?'; 'What is the probability that a patient diagnosed with breast cancer will survive more than five years?'
- The evaluation of competing treatments or preventative measures relies heavily on statistical concepts in both the design and analysis phase.

In a short article such as this it is impossible to cover all areas of medicine in which statistical methodology is of particular importance and so we shall concentrate on only three namely, clinical trials, imaging and molecular biology. (For a more comprehensive account of the use of statistics in medicine see Everitt and Palmer (2010)).

Clinical Trials

If a doctor claims that a certain type of psychotherapy will cure patients of their depression, or that taking large doses of vitamin C can prevent and even cure the common cold, how should these claims be assessed? What sort of evidence do we need to decide that claims made for the efficacy of clinical treatments are valid? One thing is certain: We should *not* rely either on the views of 'experts' unless they provide sound empirical evidence (measurements, observations, i.e., *data*) to support their views, nor should we credit the anecdotal evidence of people who have had the treatment and, in some cases, been 'miraculously' cured. (And it should be remembered that the plural of anecdote is not evidence.) Such 'wonder' treatments, which are often exposed as ineffectual when exposed to more rigorous examination, are particularly prevalent for those complaints for which conventional medicine has little to offer (see the discussion of alternative therapies in Chapter 13 of Everitt 2008).

There is clearly a need for some form of carefully controlled procedure for determining the relative effects of different treatments and this need has been met in the 20th and 21st centuries by the development of the clinical trial, a medical experiment designed to evaluate which (if any) of two or more treatments is the more effective. The quintessential components of a clinical trial, the use of a control group and, in particular the use of ►randomization as a way of allocating participants in the trial to treatment and control groups, were laid down in the first half of the 20th century. The randomization principle in clinical trials was indeed perhaps the greatest contribution made by arguably the greatest statistician of the 20th century, Sir Ronald Aylmer Fisher. Randomization achieves the following:

- It provides an impartial method, free of personal bias, for the assignment of participants to treatment and control groups. This means that treatment comparisons will not be invalidated by the way the clinician might chose to allocate the participants if left to his or her own judgment.
- It tends to balance treatment groups in terms of extraneous factors that might influence the outcome of treatment, even in terms of those factors the investigator may be unaware of.

Nowadays some 9,000–10,000 clinical trials are undertaken in all areas of medicine form the treatment of acne to the prevention of cancer and the randomized controlled clinical trial is perhaps the outstanding contribution of statistics to 20th century medical research. And in the 21st century statisticians have applied themselves to developing methods of analysis for such trials that can deal with the difficult problems of patient drop-out, the longitudinal aspects of most trials and the variety of measurement types used in such trials (see Everitt and Pickles 2004).

Imaging

Examples of medical imaging systems include conventional radiology (X-rays), positron-emission tomography (PET), magnetic resonance imaging (MRI) and functional magnetic resonance imaging (fMRI). A significant advantage often claimed for medical imaging is its ability to visualize structures or processes in the patient without the need for intrusive procedures, for example, surgery; but this may also be a disadvantage and the question that may need to be asked is how well do the conclusions from an imaging experiment correspond to the physical properties that might have been found from an intrusive procedure?

Imaging studies generate large amounts of data and a host of statistical techniques have been employed to analyze such data and to extract as much information as possible from what is in many cases very 'noisy' data. Autoregressive models, linear mixed effects models, finite mixture models and Gaussian random field theory have all been applied to mixture data with varying degrees of success. Some important references are Besag (1986), Silverman et al. (1990) and Lange (2003).

Molecular Biology

Molecular biology is the branch of biology that studies the structure and function of biological macromolecules of a cell and especially their genetic role. A central goal of molecular biology is to decipher the genetic information and understand the regulation of protein synthesis and interaction in the cellular process. Advances in biotechnology have allowed the cloning and sequencing of DNA and the massive amounts of data generated have given rise to the new field of ►bioinformatics which deals with the analysis of such data. A variety of statistical methods have been used in this area; for example, hidden Markov models have been used to model dependencies in DNA sequences and for gene finding (see Schliep et al. 2003) and data mining techniques (see ►Data Mining), in particular, cluster analysis (see, for example, Everitt et al. 2010) have been used to identify sets of genes according to their expression in a set of samples, and to cluster samples (see ►Cluster Sampling) into homogeneous groups (see Toh and Honimoto 2002).

Statistical methods are an essential part of all medical studies and increasingly sophisticated techniques now often get a mention in papers published in the medical literature. Some of these have been mentioned above but others which are equally important are Bayesian modeling (see Congdon 2001) and generalized estimating equations (see Everitt and Pickles 2004). In these days of evidence-based medicine (Sackett et al. 1996), collaboration between medical researchers and statisticians is essential to the success of almost all research in medicine.

About the Author

Brian Everitt retired from his post as Head of the Department of Computing and Statistics at the Institute of Psychiatry, King's College, London in 2005. He is the author (or joint author) of about 100 journal papers and 60 books. In retirement he continues to write and with colleagues has nearly completed the 5th edition of *Cluster Analysis*, first published in 1974. Apart from writing his interests are playing classical guitar (badly), playing tennis, walking and reading.

Cross References

►Biopharmaceutical Research, Statistics in
►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest
►Medical Statistics
►Research Designs
►Role of Statistics
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences
►Statistics Targeted Clinical Trials Stratified and Personalized Medicines
►Statistics: Nelder's view
►Survival Data
►Time Series Models to Determine the Death Rate of a Given Disease

References and Further Reading

Besag J (1986) On the statistical analysis of dirty pictures (with discussion). J Roy Stat Soc Ser B 48:259–302
Collett D (2003) Survival data in medical research. CRC/Chapman and Hall, London
Congdon P (2001) Bayesian statistical modelling. Wiley, Chichester
Everitt BS (2008) Chance rules, 2nd edn. Springer, New York
Everitt BS, Landau S, Leese M, Stahl D (2010) Cluster analysis, 5th edn. Wiley, Chickester, UK
Everitt BS, Palmer CR (2010) Encyclopaedic companion to medical statistics, 2nd edn. Wiley, Chichester, UK
Everitt BS, Pickles A (2004) Statistical aspects of the design and analysis of clinical trials. Imperial College Press, London
Lange N (2003) What can modern statistics offer imaging neuroscience? Stat Methods Med Res 12(5):447–469
Sackett DL, Rosenberg MC, Gray JA, Haynes RB, Richardson W (1996) Evidence-based medicine: what it is and what it isn't. Brit Med J 312:71–72
Schliep A, Schonhuth A, Steinhoff C (2003) Using hidden Markov models to analyze gene expression data. Bioinformatics 19: 255–263
Silverman BW, Jones MC, Wilson JD, Nychka DW (1990) A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). J Roy Stat Soc Ser B 52:271–324
Toh H, Honimoto K (2002) Inference of a genetic network by a combined approach to cluster analysis and graphical Gaussian modelling. Bioinformatics 18:287–297

Medical Statistics

Vern T. Farewell[1], Daniel M. Farewell[2]
[1]Associate Director
Medical Research Council, Biostatistics Unit, Cambridge, UK
[2]School of Medicine, Cardiff University, Cardiff, UK

Historical Background

The term statistics has at least three, related, meanings. It may refer to data in raw form, or to summaries thereof, or to the analysis of uncertainty associated with data. The phrase medical statistics, therefore, may reasonably be applied to the specialization to medical science of any of these understandings of statistics.

Raw medical statistics date back at least to the London Bills of Mortality, collected weekly between 1603 and 1836 in order to provide an early warning of plague. The early demographic work of John Graunt (1620–1674) was based on these Bills. The summaries of vital statistics undertaken by William Farr (1807–1883), working at the General Registry Office of England and Wales, became the basis of many important health reforms. However, the founding editors of the journal *Statistics in Medicine* described modern medical statistics as "the deployment of the ideas, principles and methods of statistics to stimulate deeper understanding in medicine" (Colton et al. 1982), emphasizing the third understanding of the term.

The history of the link between statistics and medicine includes key figures in the development of statistics itself. For example, Arbuthnot (1667–1753) and Bernoulli (1700–1782), often cited in the early use of significance tests, were each qualified in both mathematics and in medicine. Many individuals have contributed to the emergence of medical statistics as a scientific discipline in its own right. The French writers, Pinel (1745–1826), Louis (1787–1872) and Gavarret (1809–1890) and the Danish physician, Heiberg (1868–1963) provided early impetus. Subsequently, Pearl (1879–1940) and Greenwood (1880–1949) established research programmes in medical statistics in the USA and the UK respectively. In 1937, Hill (1897–1991) published the highly influential book, *Principles of Medical Statistics*, Hill (1937), of which twelve editions were published over the next 55 years. Two other important contributions of Hill were arguably the first modern randomized clinical trial on the effect of streptomycin in tuberculosis, and his discussion of criteria for causality in epidemiological studies. A useful source for information on the history of medical statistics is the Lind Library [http://www.jameslindlibrary.org].

The Nature of Medical Statistics

Much activity in medical statistics is necessarily collaborative. Over the course of a career, statisticians engaged in medical research are likely to work closely with physicians, nurses, laboratory scientists and other specialists. Communication across disciplines can present challenges but, in addition to its scientific merit, also frequently stimulates worthwhile methodological and theoretical research. Further, since medical research often raises ethical issues, these too must be considered by medical statisticians. Hill (1936) stressed that the statistician "cannot sit in an armchair, remote and Olympian, comfortably divesting himself of all ethical responsibility."

A dominant characteristic of the statistical methods arising in medical statistics is that they must make allowance for known variability. Comparisons of groups should adjust for systematic discrepancies between groups, for instance in terms of demographics. This has been reflected for many years by the high profile given to regression methodology, which allows multiple explanatory variables to be incorporated. A more recent manifestation is in the monitoring of medical performance, where quality control procedures developed for industrial application have been modified to allow for predictable heterogeneity in medical outcomes (Grigg et al. 2003).

Illustrative Methodological Developments

In 1984, Cox identified three important periods in the development of modern statistical methodology. The first was linked to developments in agriculture, the second to industrial applications, and the third to medical research. Developments linked to medical research flourished in the 1970s; where earlier statistical methodology placed particular emphasis on normally distributed data, there was a need for methods more suited to survival (or time-to-event) and categorical data. A distinguished example of the former is Cox's own pioneering paper (Cox 1972), presenting a semiparametric regression model for ▶survival data that did not require full specification of an underlying survival distribution. In addition, and in contrast to virtually all other regression methods then available, this model allowed the incorporation of explanatory variables that varied over time. A wealth of subsequent extensions to this already very general methodology followed, many facilitated by Aalen's (1978) reformulation of the problem in a counting process framework [see also Andersen et al. (1993)].

An important application of statistical models for categorical data was to ▶case- control studies. These epidemiological investigations of the relationship between a disease

D and exposure E, a possible risk factor, involve separate sampling of diseased and disease-free groups, from which information on E and other disease risk factors is obtained. Binary ►logistic regression would seem to provide a natural tool for the analysis of these studies, but for the fact that it focuses on $\mathrm{pr}(D|E)$ whereas the sampling is from the distribution $\mathrm{pr}(E|D)$. Building on a series of earlier papers, Prentice and Pyke (1979) established how a prospective logistic regression model for $\mathrm{pr}(D|E)$ could be used with case-control data to provide valid estimates of the odds-ratio parameters. This rapidly became the standard methodology for the analysis of case-control studies (Breslow 1996).

Study Design

The design of medical studies is also a major area of activity for medical statisticians. The paradigmatic design is perhaps the Phase III clinical trial, of which a key aspect is often randomized treatment assignment. While ►randomization can provide a basis for statistical inference, its primary motivation in trials is to enable statements of causality, critical for Phase III trials where the aim is to establish treatment efficacy. Nevertheless, the need for, and methods of, randomization continue to generate discussion, since randomization can be seen to sacrifice potential individual advantage for collective gain. Other design questions arise in Phase I trials that establish the tolerability of treatments and basic pharmacokinetics, and Phase II trials aimed at finding potentially efficacious treatments or dosages.

For ethical reasons, ongoing monitoring of data during a clinical trial is often needed, and this has been an area of methodological investigation within medical statistics since the pioneering work of Armitage (1975) (a comprehensive discussion may be found in Jennison and Turnbull (2000)). There is also an increasing role for statisticians on formal committees that monitor trial data and safety, where their expertise is combined with that of physicians, ethicists, and community representatives to ensure the ethical conduct of trials more generally.

In the 1980s, two important variations on the standard case-control design emerged, namely case-cohort studies (Prentice 1986) and two stage case-control designs (Breslow and Cain 1988); both have proved very useful in epidemiology. Epidemiological cohorts where individuals are followed to observe disease incidence, or clinical cohorts for which information on patients with specified conditions is collected routinely – both usually implemented over long periods of time – also continue to present design and analysis challenges to the medical statistician.

More Recent Topics of Interest

Typically, medical studies are conducted not only to discover statistical associations, but also in the hopes of suggesting interventions that could benefit individuals or populations. This has led to a preference for investigations incorporating randomization or multiple waves of observation, based on the idea that cause should precede effect. Randomized or not, information gathered repeatedly on the same subjects is known as longitudinal data, and its analysis has become a major subdiscipline within medical statistics. Two distinct approaches to longitudinal data analysis have risen to prominence: likelihood-based models (incorporating both classical and Bayesian schools of thought) and estimating-equation techniques.

A consequence of this emphasis on studies monitoring subjects over several months (or even years) has been an increased awareness that data, as collected, are often quite different from what was intended at the design stage. This may be due to subjects refusing treatment, or choosing an alternate therapy, or dropping out of the investigations altogether. Likelihood approaches to longitudinal data may be extended to incorporate an explicit model for the observation process (Henderson et al. 2000), while estimating equations can be modified with subject- or observation-specific weights (Robins et al. 1995) to account for departures from the study design. Non-compliance, dynamic treatment regimes, and incomplete data are all areas of active methodological research within medical statistics.

Two other major areas of current interest are meta-analysis and genetic or genomic applications. Meta-analysis is often taken to refer to the technical aspects of combining information from different studies that address the same research question, although the term is sometimes used to describe the more general systematic review, which includes broader issues such as study selection. Study heterogeneity is an important aspect of ►meta-analysis that the statistician must address. The size and complexity of genetic and genomic data present major statistical and computational challenges, notably due to hypothesis test multiplicity.

Conclusion

Medicine remains a major area of application driving methodological research in statistics, and the demand for medical statisticians is considerable. A comprehensive introduction to the area can be found in Armitage et al. (2002) and a less technical introduction is Matthews and Farewell (2007).

About the Author

Prior to moving to the MRC Bistatistics Unit, Vern Farewell held professorial positions at the University of Washington, the University of Waterloo and University College London. He has published over 200 papers in the statistical and medical literature and is co-author of the four editions of the book *Using and Understanding Medical Statistics*. Since 2007, he has been Editor of *Statistics in Medicine*.

Cross References

►Biostatistics
►Case-Control Studies
►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest
►Hazard Regression Models
►Logistic Regression
►Medical Research, Statistics in
►Meta-Analysis
►Modeling Survival Data
►Psychiatry, Statistics in
►Statistical Analysis of Longitudinal and Correlated Data
►Statistical Genetics
►Statistical Methods in Epidemiology
►Statistics, History of
►Statistics: An Overview
►Survival Data

References and Further Reading

Aalen OO (1978) Nonparametric inference for a family of counting processes. Ann Stat 6:701–726

Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Armitage P (1975) Sequential medical trials. Blackwell, Oxford

Armitage P, Berry G, Matthews JNS (2002) Statistical methods in medical research. Blackwell Science, Oxford

Breslow NE (1996) Statistics in epidemiology: the case control study. J Am Stat Assoc 91:14–28

Breslow NE, Cain KC (1988) Logistic regression for two-stage case-control data. Biometrika 75:11–20

Colton T, Freedman L, Johnson T (1982) Editorial. Stat Med 1:1–3

Cox DR (1972) Regression models and life tables (with discussion). J R Stat Soc B 34:187–220

Cox DR (1984) Present position and potential developments: some personal views: design of experiments and regression. J R Stat Soc A 147:306–315

Grigg OA, Farewell VT, Spiegelhalter DJ (2003) Use of risk adjusted CUSUM and RSPRT charts for monitoring in medical contexts. Stat Meth Med Res 12:147–170

Henderson R, Diggle P, Dobson A (2000) Joint modelling of repeated measurements and event time data. Biostatistics 1:465–480

Hill AB (1936) Medical ethics and controlled trials. Br Med J 3: 1043–1049

Hill AB (1937) Principles of medical statistics. Lancet, London

Jennison C, Turnbull BW (2000) Group sequential methods with applications to clinical trials. Chapman and Hall/CRC, New York

Matthews DE, Farewell VT (2007) Using and understanding medical statistics. Karger, Basel

Prentice RL (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73:1–12

Prentice RL, Pyke R (1979) Logistic disease incidence models and case-control studies. Biometrika 66:403–411

Robins JM, Rotnitsky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc 90:106–121

Meta-Analysis

Elena Kulinskaya[1], Stephan Morgenthaler[2], Robert G. Staudte[3]

[1]Professor, Aviva Chair in Statistics
University of East Anglia, Norwich, UK

[2]Professor, Chair of Applied Statistics
Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[3]Professor and Head of Department of Mathematics and Statistics
La Trobe University, Bundoora, VIC, Australia

Introduction

Given several studies on the same topic, a *meta-analysis* synthesizes the information in them so as to obtain a more precise result. The proper procedure of conducting a *systematic review* of literature, the selection of which studies to include and the issues of *publication bias* and other possible biases are important aspects not covered here and we refer the interested reader to Cooper and Hedges (1994) and Higgins and Green (2008). We assume all studies estimate the same *effect*, which is often a comparison of outcomes for control and treatment groups via clinical trials. Examples for two binomial samples with parameters $(n_1, p_1), (n_2, p_2)$ are the *risk difference* $p_1 - p_2$, *relative risk* p_2/p_1 and *odds ratio* $\{p_2/(1-p_2)\}/\{p_1/(1-p_1)\}$. Other examples comparing normal samples are the difference in means $\mu_1 - \mu_2$, or *effect sizes* such as the *standardized mean difference, or Cohen's-d* $d = (\mu_1 - \mu_2)/\sigma$ from Cohen (1988), where σ^2 is an assumed common variance, and Glass's $g = (\mu_1 - \mu_2)/\sigma_1$ from Glass (1976), where σ_1^2 is the variance of the control group.

Traditional Meta-Analysis Methodology

We are given K independent studies, in which the estimated effects $\hat{\theta}_k$ based on N_k observations are asymptotically normal such that $\hat{\theta}_k$ is for large enough N_k approximately normally distributed with mean θ_k and variance σ_k^2/N_k. This is denoted $\hat{\theta}_k \sim AN\left(\theta_k, \sigma_k^2/N_k\right)$ for each $k = 1, \ldots, K$. Examples satisfying the above assumptions are the risk difference, the log-relative risk, the log-odds ratio and the Cohen's-d. The goal is to combine the estimators $\hat{\theta}_k$ in some way so as to estimate a representative θ for all K studies, or even more ambitiously, for all potential studies of this type. Thus there is the conceptual question of how to define a representative θ, and the inferential problem of how to find a confidence interval for it.

Confidence Intervals for Effects

Note that for each individual study, one can already form large sample confidence intervals for individual θ_k, $k = 1, \ldots, K$. For *known* σ_k, a $100(1-\alpha)\%$ large-sample confidence interval for θ_k is $[L_k, U_k] = \left[\hat{\theta}_k - z_{1-\alpha/2}\sigma_k/N_k^{1/2},\ \hat{\theta}_k + z_{1-\alpha/2}\sigma_k/N_k^{1/2}\right]$, where $z_\beta = \Phi^{-1}(\beta)$ is the β quantile of the standard normal distribution. If σ_k is *unknown*, and there exists estimators $\hat{\sigma}_k$ with $\hat{\sigma}_k/\sigma_k \to 1$ in probability as $N_k \to \infty$, then the same can be said for $[L_k, U_k] = \left[\hat{\theta}_k - z_{1-\alpha/2}\hat{\sigma}_k/N_k^{1/2}, \hat{\theta}_k + z_{1-\alpha/2}\hat{\sigma}_k/N_k^{1/2}\right]$.

Unequal Fixed Effects Model (UFEM)

Standard meta-analysis proceeds by choosing a weight w_k for each study and combines the estimated $\hat{\theta}_k$ through weighted means. If we interpret θ_k as the true effect for the study k and if this effect is of interest in its own right, then the following definition can be adopted. Consider a representative effect for the K studies defined by $\theta_w = \sum_k w_k\theta_k/W$ with $W = \sum_j w_j$. This *weighted effect* is the quantity that we want to estimate by meta-analysis. There is a good dose of arbitraryness in this procedure, because the weighted effect does not necessarily have a readily interpreted meaning. An exception occurs if the weights are all equal to one, in which case θ_w is simply the average of the study effects.

The weights are, however, often chosen to be proportional to the reciprocals of the variances in order to give more weight to θ_k that are estimated more accurately. If this is the choice, it follows that $w_k = N_k/\sigma_k^2$ and $\hat{\theta}_w = \sum_k w_k\hat{\theta}_k/W$ satisfies $\hat{\theta}_w \sim AN(\theta_w, W^{-1})$. Therefore a $100(1-\alpha)\%$ large-sample confidence interval for θ_w is given by $[L, U] = [\hat{\theta}_w - z_{1-\alpha/2}W^{-1/2},\ \hat{\theta}_w + z_{1-\alpha/2}W^{-1/2}]$.

In practice the weights usually need to be estimated, (w_k by $\hat{w}_k$ and W by $\hat{W} = \sum_k \hat{w}_k$), but a large sample confidence interval for θ_w can be obtained by substituting $\hat{\theta}_{\hat{w}}$ for $\hat{\theta}_w$ and $\hat{W}$ for W in the above interval.

Fixed Effects Model (FEM)

When statisticians speak of the fixed effects model they usually mean *equal* fixed effects which makes the very strong assumption that all $\theta_k = \theta$. This has the appeal of simplicity. The UFEM just described includes the FEM as a special case. In particular the target parameter θ_w reduces to $\theta_w = \theta$ and thus becomes a meaningful quantity no matter what weights are chosen.

However, one of the preferred choices still uses the weights inversely proportional to the variance, because in this case $\sum_k w_k\hat{\theta}_k/W$ has the smallest asymptotic variance amongst all unbiased (for θ) linear combinations of the individual study estimators of θ. The same confidence interval given above for θ_w is used for θ. The methodology for the UFEM and FEM models is the same, but the target parameter θ_w of the UFEM has a different interpretation.

Random Effects Model (REM)

The REM assumes that the true effects θ_k, $k = 1, \ldots, K$ are the realized values of sampling from a normal population with mean θ and variance γ^2 for some unknown inter-study variance γ^2, and further that the above results for the UFEM are all *conditional* on the given θ_k, $k = 1, \ldots, K$. The justification for this assumption is that the K studies are a 'random sample' of all possible studies on this topic. Inference for θ can now be interpreted as saying something about the larger population of possible studies.

Formally, the REM assumes $\theta_1, \ldots, \theta_K$ are a sample from $N(\theta, \gamma^2)$, with both parameters unknown; and $\hat{\theta}_k|\theta_k \sim AN\left(\theta_k, \sigma_k^2/N_k\right)$ for each k. If the *conditional* distribution of $\hat{\theta}_k$, given θ_k, were exactly normal, then the *unconditional* distribution of $\hat{\theta}_k$ would be exactly $\hat{\theta}_k \sim N\left(\theta, \gamma^2 + \sigma_k^2/N_k\right)$. However, in general the unconditional distributions are only asymptotically normal $\hat{\theta}_k \sim AN\left(\theta, \gamma^2 + \sigma_k^2/N_k\right)$. It is evident that one needs an estimate $\hat{\gamma}^2$ of γ^2 in order to use the inverse variance weights approach described earlier, and this methodology will be described below.

Choosing between Fixed and Random Effects Models

Qualitative Grounds

If one assumes the K studies are a random sample from a larger population of potential studies and that the true effects θ_k are each $N(\theta, \gamma^2)$ then θ is the target effect, and γ^2 is a measure of inter-study variability of the effect. In

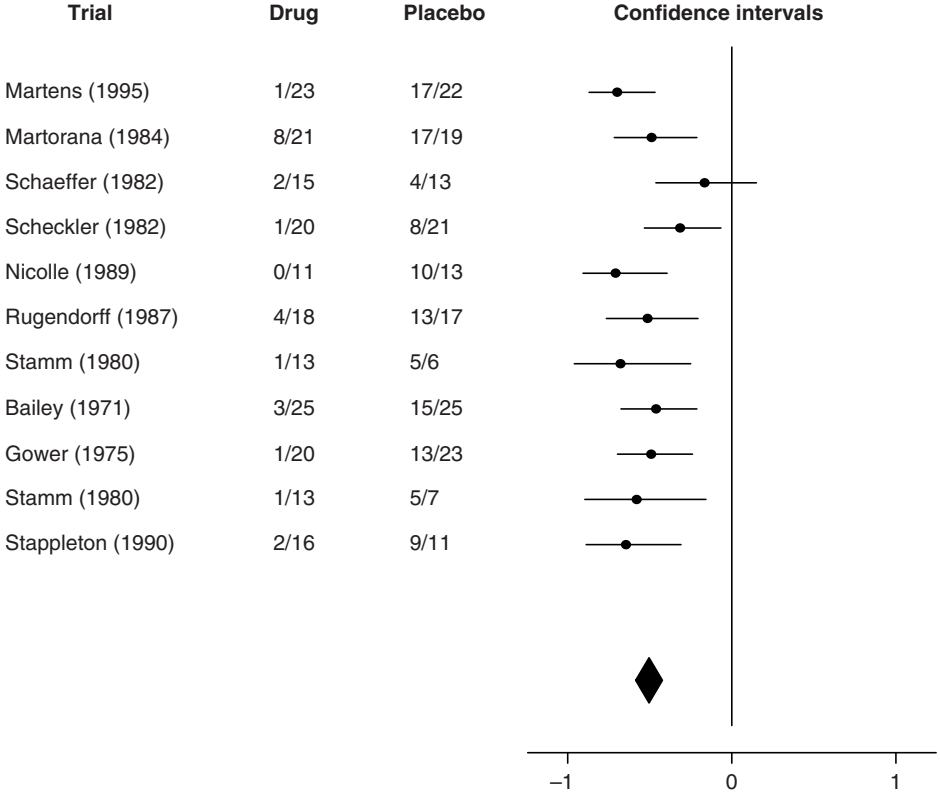
this case choose the REM. If there is reason to believe that the θ_k are different, but not the result of random sampling, then use the UFEM. In this case, it may be possible to explain a good part of the variation in the effects θ_k by *meta-regression*. The differences between the studies can sometimes be captured by variables that describe the circumstances of each study and by regressing the $\hat{\theta}_k$ on such variables, these differences can be explained and corrected. Meta-regression may thus turn a UFEM into a FEM. In both models, the target is $\theta_w = \sum_k w_k \theta_k / W$. If there is reason to believe all $\theta_k = \theta$, (the *homogeneous case*), use the FEM with target θ. For the FEM and UFEM inferential conclusions only apply to the K studies.

Quantitative Grounds

It is clear that if $\gamma^2 = 0$ in the REM, or all $\theta_k = \theta$ in the UFEM, one obtains the FEM. It is a special case of both. One way to test the null hypothesis of homogeneity (all $\theta_k = \theta$) is to use Cochran's Q, defined by $Q = \sum_k w_k (\hat{\theta}_k - \hat{\theta}_w)^2$, where w_k are the inverse variance weights and $\hat{\theta}_w = \sum_k w_k \hat{\theta}_k / W$. One can show that under the null hypothesis of homogeneity, and when each $\hat{\theta}_k$ is normally distributed, $Q \sim \chi^2_{K-1}$, so a level α test of homogeneity rejects when $Q \geq \chi^2_{K-1,1-\alpha}$. Further, under the UFEM model, the statistic Q has a non-central chisquared distribution $Q \sim \chi^2_{K-1}(\lambda)$, where $\lambda = \sum_k w_k (\theta_k - \theta_w)^2$. This result and others allowing for the weaker assumption $\theta_k \sim AN\left(\theta_k, \sigma_k^2 / N_k\right)$ and estimated weights are derived in Sect. 24.1, Kulinskaya et al. (2008). In the asymptotic case, the χ^2 distributions are only approximate. Testing for heterogeneity is strongly discouraged in Higgins and Green (2008) in favor of the quantification of inherently present heterogeneity.

Inference for the REM

Let $M_r = \sum_k w_k^r$ for inverse variance weights w_k, and $a = M_1 - M_2/M_1$. It can be shown that for this model $\mathrm{E}[Q] = K - 1 + a\gamma^2$. This "justifies" the DerSimonian and Laird (1986) estimator $\hat{\gamma}^2_{DL} = \{Q - (K-1)\}^+ / a$, where $\{\ldots\}^+$ means set the quantity in brackets equal to 0 if it is negative and otherwise leave it. Using this estimator and $\hat{\theta}_k \sim AN\left(\theta, \gamma^2 + w_k^{-1}\right)$, we have new weights $w_k^* = \left(\gamma^2 + w_k^{-1}\right)^{-1}$



Meta-Analysis. Fig. 1 The data of eleven independent studies of antibiotic treatment to prevent recurrent urinary tract infection are presented in this forest plot. The confidence intervals for the individual studies are shown on the right-hand side. The lozenge at the bottom shows the combined confidence interval, the result of the meta-analysis

and estimator $\hat{\theta}^* = \sum_k w_k^* \hat{\theta}_k / W^* \sim AN(\theta, \{W^*\}^{-1})$, where $W^* = \sum_k w_k^*$. In practice w_k^* is usually estimated by $\hat{w}_k^* = 1/\left(\hat{\gamma}_{DL}^2 + \hat{w}_k^{-1}\right)$. Anotherestimator of γ^2 is proposed in Biggerstaff and Tweedie (1997).

Meta-Regression

In some cases there is information regarding the K studies which may explain the inter-study variance. In this case the estimated effects $\hat{\theta}_k$ can be considered as responses to be regressed on explanatory variables $x_1, \ldots, x_p$, also called *moderators*. Thus one has $y_k = \beta_0 + \beta_1 x_{k1} + \cdots + \beta_p x_{kp} + \epsilon_k$, where y_k is the estimated effect $\hat{\theta}_k$ (or a transformed effect), and ϵ_k is the random error in the kth study, $k = 1, \ldots, K$. Weighted least squares (with known or estimated weights) can be used to estimate the coefficients. When the variance stabilizing transformation is applied to estimated effects, generalized linear models techniques (see ►Generalized Linear Models) with Gaussian family of distributions can be used, see Chap. 14 of Kulinskaya et al. (2008).

Example

As illustration, consider a series of 11 studies of antibiotic treatment to prevent recurrent urinary tract infection. The sources of the data, the data themselves, and the confidence intervals are shown in Fig. 1. These studies are part of those reviewed by Albert et al. (2004) and have been discussed in Chap. 19 (p. 158) of Kulinskaya et al. (2008). The total sample sizes range from $N = 19$ to $N = 50$. The parameter of interest is the risk difference $p_1 - p_2$ between the placebo group and the treated groups. The studies show a more or less strong benefit of the treatment, while the meta-analysis gives a fairly convincing result. This depiction of results is known as a *forest plot*.

Additional Literature

The traditional approach is general, only requiring asymptotically normal effects and estimates for the weights. However the methodology is overly simple, because it assumes known weights, when in fact they usually need to be estimated. Recent studies indicate that typical sample sizes are woefully inadequate in order for the approximations that assume known weights to be reliable (Malzahn et al. 2000; Viechtbauer 2007). One way of overcoming this problem is to employ variance stabilization of the estimated effects before applying the traditional approach, see Kulinskaya et al. (2008). For further reading we recommend the classical work Hedges and Olkin (1985), as well as the recent books Böhning et al. (2008), Borenstern et al. (2009), Hartung et al. (2008) and Whitehead (2002).

About the Authors

Prof. Elena Kulinskaya is a recently appointed Aviva Chair in Statistics, University of East Anglia. Previously she has been Director of the Statistical Advisory Service at Imperial College London (2004–2010). She is also a Visiting Professor at The Center for Lifespan and Chronic Illness Research (CLiCIR), University of Hertfordshire. She has a long standing interest in statistical evidence and its applications in meta-analysis. She has authored and co-authored 78 papers, including numerous theoretical and applied papers on meta-analysis, and a recent book on meta analysis (*Meta-analysis: A Guide to Calibrating and Combining Statistical Evidence*, Wiley, 2008) co-authored with Stephan Morgenthaler and R.G. Staudte and dedicated to a new approach based on variance stabilization.

Dr. Stephan Morgenthaler is Professor of Applied Statistics in the Institute of Mathematics Ecole Polytechnique Fédérale de Lausanne in Switzerland. He has authored, co-authored and edited more than 80 papers and eight books. He is a member of the ISI and a Fellow of the American Statistical Association. He served as a vice-president of ISI from 2004 to 2008.

Dr. Robert G. Staudte is Professor and Head, Department of Mathematics and Statistics, La Trobe University, Melbourne, Australia. He has authored and co-authored more than 50 papers and four books, including *Robust Estimation and Testing*, Wiley 1990, co-authored with Professor Simon J. Sheather; and *Meta Analysis: a Guide to Calibrating and Combining Statistical Evidence*, Wiley 2008, co-authored with Professors Elena Kulinskaya and Stephan Morgenthaler. He was Associate Editor of the *Journal of Statistical Planning and Inference* (1995–1998).

Cross References

►Clinical Trials: Some Aspects of Public Interest
►Effect Size
►Forecasting Principles
►Medical Statistics
►Psychology, Statistics in
►P-Values, Combining of
►Time Series Models to Determine the Death Rate of a Given Disease

References and Further Reading

Albert X, Huertas I, Pereiró I, Sanfelix J, Gosalbes V, Perrota C (2004) Antibiotics for preventing recurrent urinary tract infection in non-pregnant women (Cochran Review). In: The Cochran Library, Issue 3. Wiley, Chichester, UK

Biggerstaff BJ, Tweedie RL (1997) Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. Statistics in Medicine 16:753–768

Böhning D, Kuhnert R, Rattanasiri S (2008) Meta-analysis of Binary data using profile likelihood. Chapman and Hall/CRC Statistics. CRC, Boca Raton, FL
Borenstern M, Hedges LV, Higgins JPT, Rothstein H (2009) Introduction to meta analysis. Wiley, London
Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Earlbaum Associates, Hillsdale, NJ
Cooper H, Hedges LV (eds) (1994) The handbook of research synthesis. Russell Sage Foundation, New York
DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. control Clin Trials 7:177–188
Glass GV (1976) Primary, secondary and meta-analysis of research. Educ Res 5:3–8
Hartung J, Knapp G, Sinha BK (2008) Statistical meta analysis with applications. Wiley, Chichester
Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic, Orlando
Higgins JPT, Green S (eds) (2008) Cochrane handbook for systematic review of interventions version 5.0.1. The Cochrane Collaboration: available on www.cochrane-handbook.org
Kulinskaya E, Morgenthaler S, Staudte RG (2008) Meta analysis: a guide to calibrating and combining statistical evidence. Wiley, Chichester
Malzahn U, Bohning D, Holling H (2000) Nonparametric estimation of heterogeneity variance for the standardized difference used in meta-analysis. Biometrika 87(3):619–632
Viechtbauer W (2007) Hypothesis tests for population heterogeneity in meta-analysis. Br J Math Stat Psychol 60:29–60
Whitehead A (2002) Meta-analysis of controlled clinical trials. Applied statistics. Wiley, Chichester

Method Comparison Studies

Graham Dunn
Professor of Biomedical Statistics and Head of the Health Methodology Research Group
University of Manchester, Manchester, UK

We are here concerned with the comparison of the performance to two or more measurement devices or procedures. At its simplest, a method comparison study involves the measurement of a given characteristic on a sample of subjects or specimens by two different methods. One possible question is then whether measurements taken by the two different methods are interchangeable. Another is whether one of the two methods is more or less precise than the other. A third, more difficult task, is to calibrate one set of fallible measurements (using Device A, for example) against another set of fallible measurements produced by device B. A potentially-serious problem in all of these situations is the possibility that the measurement errors arising from the use of these two devices may be correlated. A slightly more complicated study involves replication of each of the sets of measurements taken using the two different procedures or devices, usually carried out on the naïve assumption that the measurement errors of the within-device replicates will be uncorrelated and that replication will enable the investigator to obtain an unbiased estimate of the instruments' precisions (based on the standard deviations of the replicates).

Let's return to the simplest situation – measurement of a given characteristic on a sample of subjects by two different methods that are assumed to provide independent measurement errors. Are the two methods interchangeable? How closely do the measurements agree with each other? Is this agreement good enough for all our practical purposes? A method suggested by Bland and Altman (1986) is to determine *limits of agreement*. One simply subtracts the measurement arising from one method from the corresponding measurement using the other. The average of these differences tells us about the possibility of relative bias (and the so-called Bland-Altman plot – a graph of the difference against the average of the two measurements – may tell us that the bias is changing with the amount of the characteristic being measured, but it is not 100% fool-proof since a relationship between the difference between and the average of the two measures may arise from differences in the instruments' precisions). The standard deviation of the differences tells us about the variability of the difference of the two measurement errors. The 95% limits of agreement are simply defined as the range of differences between the 2.5th and 97.5th percentiles or, assuming normality, approximately two standard deviations either side of the mean. If the measurement errors for the two methods are positively correlated then the variability of the differences will be less than one would expect if they were uncorrelated and the limits of agreement will be too small. If the measurement methods use different scales (comparison of temperatures in °C and °F, for example) then this simple procedure will break down and the limits of agreement will fail to tell the investigator that the two methods are interchangeable (after suitable rescaling).

One might be tempted to plot results using one of the methods (in °F, for example) against the other (in °C) and carry out a simple regression to calibrate one against the other. But the hitch is that both methods are subject to error (the classical errors-in-variables problem) and the estimate of the regression coefficient would be biased (attenuated towards zero). If one knows the ratio of the variances of the measurement errors for the two methods then it is possible to use orthogonal regression, widely-known as Deming's regression, to solve the problem. The

catch is that one does not normally have an unbiased estimate of the ratio of these two variances – the problem again arising from the lack of independence (i.e., correlation) of any replicate measures used to determine these variances (Carroll and Ruppert 1996).

A third relatively simple approach is to look for and make use of an instrumental variable (IV) through IV or ►two-stage least squares (2SLS) regression methods. Here we need a variable (not necessarily a third measurement of the characteristic, but it may be) that is reasonably highly correlated with the characteristic being measured but can be justifiably assumed to be uncorrelated with the associated measurement errors. If we label the measurements using the two methods as X and Y, and the corresponding values of the instrumental variable as Z, then the instrumental variable estimator of the slope of Y on X is given by the ratio Cov(Y,Z)/Cov(X,Z) – see Dunn (2004, 2007). From here it's a relatively simple move into factor analysis models for data arising from the comparison of three or methods (Dunn 2004).

Statistical analyses for the data arising from more the informative designs, with more realistic measurement models (heteroscedasticity of measurement errors, for example), is beyond the scope of this article but the methods are described in considerable detail in Dunn (2004). The methods typically involve software developed for covariance structure modelling. Analogous methods for the comparison of binary measurements (diagnostic tests) can also be found in Dunn (2004).

About the Author

For biography *see* the entry ►Psychiatry, Statistics in.

Cross References

►Calibration
►Instrumental Variables
►Measurement Error Models
►Two-Stage Least Squares

References and Further Reading

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310

Carroll RJ, Ruppert D (1996) The use and misuse of orthogonal regression in linear errors-in-variables models. Am Stat 50:1–6

Dunn G (2004) Statistical evaluation of measurement errors. Arnold, London

Dunn G (2007) Regression models for method comparison data. J Biopharm Stat 17:739–756

Methods of Moments Estimation

Martin L. Hazelton
Chair of Statistics
Massey University, Palmerston North, New Zealand

The method of moments is a technique for estimating the parameters of a statistical model. It works by finding values of the parameters that result in a match between the sample moments and the population moments (as implied by the model). This methodology can be traced back to Pearson (1894) who used it to fit a simple mixture model. It is sometimes regarded as a poor cousin of maximum likelihood estimation since the latter has superior theoretical properties in many settings. Nonetheless, the method of moments and generalizations thereof continue to be of use in practice for certain (challenging) types of estimation problem because of their conceptual and computational simplicity.

Consider a statistical model defined in terms of a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathsf{T}}$. We denote by $\mu_k = \mathsf{E}[X^k]$ the kth moment about zero of a random variable X generated by our model. This moment will be a function of $\boldsymbol{\theta}$, and so we will write $\mu_k = \mu_k(\boldsymbol{\theta})$ to emphasize this dependence.

Suppose that we have a (univariate) random sample $X_1, \ldots, X_n$ from the model, which we want to use to estimate the components of $\boldsymbol{\theta}$. From this we can compute the kth sample moment, $\hat{\mu}_k = n^{-1} \sum_{i=1}^{n} X_i^k$. The rationale for the method of moments is that the sample moments are natural estimators of the corresponding model-based moments, and so a good estimate of $\boldsymbol{\theta}$ will reproduce these observed moments. In practice it is usual (although not essential) to use moments of the lowest possible orders in order to obtain parameter estimates. The method of moments estimator $\hat{\boldsymbol{\theta}}$ is hence defined to be the solution of the system of equations

$$\mu_k(\boldsymbol{\theta}) = \hat{\mu}_k \qquad k = 1, 2, \ldots, q$$

where q is the smallest integer for which this system has a unique solution.

As an example, suppose that $X_1, \ldots, X_n$ are drawn from a ►gamma distribution with shape parameter α and scale parameter β. Then $\mu_1 = \alpha\beta$ and $\mu_2 = \alpha(\alpha + 1)\beta^2$. The method of moments estimators $\hat{\alpha}$ and $\hat{\beta}$ therefore satisfy the pair of equations

$$\hat{\alpha}\hat{\beta} = \hat{\mu}_1$$
$$\hat{\alpha}(\hat{\alpha} + 1)\hat{\beta}^2 = \hat{\mu}_2.$$

Solving these we obtain

$$\hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} \quad \text{and} \quad \hat{\beta} = \frac{\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_1}.$$

Method of moments estimators are, in general, consistent. To see this, note that the (weak) law of large numbers ensures that the sample moments converge in probability to their population counterparts. It then follows that if $\mu_k(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ for $k = 1, \ldots, q$ then the method of moments estimators will converge in probability to their true values. However, method of moments estimators are less efficient than maximum likelihood estimators, at least in cases where standard regularity conditions hold and the two estimators differ. Furthermore, unlike maximum likelihood estimation, the method of moments can produce infeasible parameter estimates in practice. For example, if $X_1, \ldots, X_n$ are drawn from a uniform distribution (see ►Uniform Distribution in Statistics) on $[0, \theta]$ then the method of moments estimator is $\hat{\theta} = 2\bar{X}$, but this estimate is infeasible if $\max\{X_i\} > 2\bar{X}$.

Despite the theoretical advantages of maximum likelihood estimation, the method of moments remains an important tool in many practical situations. One reason for this is that method of moments estimates are straightforward to compute, which is not always the case for maximum likelihood estimates. (For example, the maximum likelihood estimators for the gamma distribution parameters considered above are only available implicitly as the solution to the non-linear likelihood equations.) Furthermore, estimation by the method of moments does not require knowledge of the full data generating process. This has led to various extensions of the basic method of moments that can be applied in complex modeling situations.

One such extension is the generalized method of moments Hansen (1982) which is a type of generalized estimating equation methodology, widely used in econometrics. This technique works by utilizing sample and population moment conditions (or "orthogonality conditions") of the statistical model, and can provide estimates of parameters of interest in a model even when other model parameters remain unspecified. Another useful extension is the simulated method of moments (e.g., Gelman 1995). This technique can be employed when the model is so complex that neither the density function for the data nor the theoretical moments are available in closed form. It therefore provides a means of fitting micro-simulation and mechanistic stochastic models (Diggle and Gratton 1984).

About the Author

Professor Hazelton was appointed to the Chair of Statistics at Massey University in 2006. His current research interests include modeling and inference for transport networks, and multivariate smoothing problems. Professor Hazelton is an Associate Editor of the *Journal of the Korean Statistical Society* and a member of the Editorial Advisory Board for *Transportation Research Part B*.

Cross References

►Estimation
►Estimation: An Overview
►Social Network Analysis
►Statistical Inference for Stochastic Processes
►Statistics of Extremes
►Univariate Discrete Distributions: An Overview

References and Further Reading

Diggle P, Gratton J (1984) Monte Carlo methods of inference for implicit statistical models. J R Stat Soc B 46:193–227
Gelman A (1995) Method of moments using Monte Carlo simulation. J Comput Graph Stat 3:36–54
Hansen LP (1982) Large sample properties of generalized method of moments estimators. Econometrica 50:1029–1054
Pearson K (1894) Contribution to the mathematical theory of evolution. Philos Tr R Soc S-A 185:71–110

M

Minimum Variance Unbiased

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

The term *minimum variance unbiased* refers to a property of statistical decision rules.

Idea. Any statistical experiment may be perceived as a random channel transforming a deterministic quantity θ (parameter) into a random quantity X (observation). *Point estimation* is a reverse process of regaining θ from X according to a rule $\widehat{\theta} = \delta(X)$ called *estimator*. Formally, estimator is a function from the set $\mathcal{X}$, of possible values of X, into the set Θ, of possible values of θ. As a measure of imprecision of such estimator one can use the function $R_\delta(\theta) = E_\theta(\delta(X) - \theta)^2$ called the Mean Squared Error. It may be rewritten in the form

$$var_\theta \delta(X) + [b(\theta)]^2, \text{ where } b(\theta) = E_\theta \delta(X) - \theta$$

is the bias of δ.

If $b(\theta) = 0$ for all θ then $\widehat{\theta} = \delta(X)$ is said to be *unbiased*. Minimizing the MSE among the unbiased estimators reduces to minimizing its variance. Any estimator δ_0 realizing this minimum (if such exists) is said to be a *minimum variance unbiased estimator* (MVUE). Searching for such estimator or verifying whether it is a MVUE needs some special statistical tools.

Example 1 (Urn problem). An urn contains N balls, where any ball is black or white, while the number θ of black balls is unknown. To search θ we draw without replacement n balls. Let k be the number of black balls in the sample. Estimate θ.

A potential number X of black balls in the sample has the hypergeometric distribution (see ►Hypergeometric Distribution and Its Application in Statistics) taking values k with probabilities

$$P_\theta(X = k) = p_{\theta,k} = \begin{cases} \frac{\binom{\theta}{k}\binom{N-\theta}{n-k}}{\binom{N}{n}} & \text{if } k \in [\max(0, n-N+\theta), \\ & \quad \min(n, \theta)] \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Since $EX = \frac{n\theta}{N}$, the rule $\widehat{\theta} = \frac{N}{n}X$ is an unbiased estimator of θ. This is, formally, not acceptable unless n is a divisor of N, because $\widehat{\theta}$ takes values outside the parameter set. Thus one can seek for an acceptable unbiased estimator. According to the formula (1) we get

$$p_{0,k} = \begin{cases} 1, & \text{if } k = 0 \\ 0, & \text{otherwise,} \end{cases}$$

and

$$p_{1,k} = \begin{cases} \frac{N-n}{N}, & \text{if } k = 0 \\ \frac{n}{N}, & \text{if } k = 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus any unbiased estimator $\widehat{\theta} = \widehat{\theta}(X)$ must satisfy the conditions $\widehat{\theta}(X) = 0$ if $X = 0$ and $\frac{N}{n}$ if $X = 1$. Therefore the desired estimator exists if and only if n is a divisor on N.

Basic Concepts. Let $X = (X_1, \ldots, X_n)$ be a random vector, interpreted as a potential observation in a statistical experiment. Assume that distribution P of the vector belongs to a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where θ is an unknown parameter identifying P. Thereafter by distribution we shall mean density or probability mass function. Any potential estimator of θ is a function $T = t(X)$ called a statistic. If T involves the entire information on θ then one can reduce the problem by considering only these estimators which depends on X through T.

We say that a statistic T is *sufficient* for θ if the conditional probability $P_\theta(X/T)$ does not depend on θ. Determining a sufficient statistic directly from this definition may be a laborious task. It may be simplified by the well known Fisher-Neyman factorization criterion. A statistic $T = t(X)$ is sufficient for θ, if and only if, P_θ may be presented in the form $P_\theta(x) = g_\theta[t(x)]h(\theta)$. A sufficient statistic T is minimal if it is a function of any other sufficient statistic. In particular, the vector statistic $T = [t_1(X), \ldots, t_k(X)]$ in so called exponential family $P_\theta(x) = C(\theta)\exp\left[\sum_{j=1}^{k} Q_j(\theta)t_j(x)\right]h(x)$, for $\theta \in \Theta$, is sufficient.

We say that a statistic T is *complete* if for any (measurable) function f the condition $E_\theta f(T) = 0$ for all θ implies that $P[f(T) = 0] = 1$. It is known that any complete sufficient statistic (if exists) is minimal but a minimal sufficient statistic may not be complete. Moreover the above sufficient statistic in the exponential family distributions is complete providing Θ contains a k-dimensional rectangle.

Now let us consider a family of densities $\{p(x, \theta) : \theta \in \Theta\}$, where Θ is an open interval of a real line, satisfying some regularity conditions. Function $I = I(\theta)$ defined by the formula $I(\theta) = E\left[\frac{\partial \log p(X,\theta)}{\partial \theta}\right]^2$ is said to be Fisher information.

Advanced Tools. Let $X = (X_1, \ldots, X_n)$ be a random vector with a distribution P belonging to a family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and let $T = t(X)$ be a sufficient statistic for θ. In searching MVUE's one can use the following results.

►*Rao-Blackwell theorem*: If $U = u(X)$ is an unbiased estimator of a parametric function $g(\theta)$ then the conditional expectation $E[U/T]$ is also unbiased and its variance is not greater than $var(U)$.

Lehmann-Scheffé theorem: If T is, moreover, complete then any statistic $h(T)$ is a MVUE of its expectation. This MVUE is unique (with probability 1).

Rao-Cramer inequality: Let $\{p(x, \theta) : \theta \in \Theta\}$, where Θ is an open interval of a real line, be a family of densities satisfying some regularity conditions, such that $I(\theta) > 0$ for all θ. Then for any statistic $U = u(X)$ the inequality $var_\theta(U) \geq \frac{1}{I(\theta)}$ is met.

It is worth to add that the equality in the Rao-Cramer inequality is attained if and only if the family $\mathcal{P}$ of distributions is exponential. However this condition is not necessary for existing a MVUE; for instance, if $X_1, \ldots, X_n$ are i.i.d. according to the normal law $N\left(\alpha^{\frac{1}{3}}, 1\right)$. In this case the attainable minimum variance is $\frac{9\alpha^4}{n} + \frac{18\alpha^2}{n^2} + \frac{6}{n^3}$ while $\frac{1}{I(\theta)} = \frac{9\alpha^4}{n}$.

Example 2 (Bernoulli trials). Let $X_1, \ldots, X_n$ be independent and identically distributed zero-one distributions with probability $P(X_i = 1) = \theta$, where θ is unknown for $i = 1, \ldots, n$. In this case the family $\mathcal{P} = \{P_\theta : \theta \in (0,1)\}$ is exponential with complete sufficient statistic $\overline{X} = \frac{1}{n}\sum_i X_i$. Since $E\overline{X} = \theta$, the statistic $\overline{X}$ is the unique MVUE of θ. In this case the Fisher information takes the form $I(\theta) = \frac{n}{\theta(1-\theta)}$ while $var_\theta(\overline{X}) = \frac{\theta(1-\theta)}{n}$. Thus the lower bound $\frac{1}{I(\theta)}$ in the Rao-Cramer inequality is attained. It is worth to note that, similarly as in Example 1, this unique MVUE takes, with positive probability, the values 0 and 1, which lie outside the parameter set $(0,1)$.

Minimum Variance Invariant Unbiased Estimator. If distribution of the observation vector depends on several parameters, some of them may be out of our interest and play the role of nuisance parameters. Such a situation occurs, for instance, in linear models. In this case the class of all unbiased estimators is usually too large for handle. Then we may seek for an estimator which is invariant with respect to a class of transformations of observations or its variance does not depend on the nuisance parameters. An estimator minimizing variance in such a reduced class is called a minimum variance invariant unbiased estimator.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Best Linear Unbiased Estimation in Linear Models
►Cramér–Rao Inequality
►Estimation
►Properties of Estimators
►Rao–Blackwell Theorem
►Sufficient Statistics
►Unbiased Estimators and Their Applications

References and Further Reading

Cramér H (1946) Mathematical methods of statistics, Princeton University Press, Princeton, NJ
Kadec MN (1979) Sufficient statistic. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 2. Soviet Encyclopedia, Moscow, pp 375–377 (in Russian)
Nikulin MS (1984) Rao-Cramer inequality. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 4, Soviet Encyclopedia, Moscow, pp 867–868, (in Russian)
Nikulin MS (1993) Unbiased estimator. In: Hazewinkel M (ed) Encyclopaedia of mathematics. vol 9, pp 305–307
Lehmann EL (1983) Theory of point estimation. Wiley, New York
Rao CR (1973) Linear statistical inference, 2nd edn. Wiley, New York

Misuse and Misunderstandings of Statistics

Atsu S. S. Dorvlo
Professor
Sultan Qaboos University, Muscat, Sultanate of Oman

Introduction

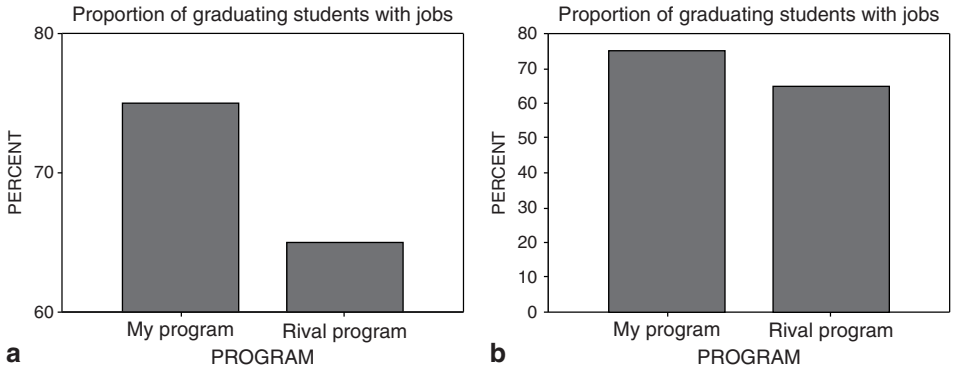
Because of the advent of high speed computers statistics has become more visible. Almost any discipline has an element of statistics in it. In fact one cannot publish in most journals when the statistics used or misused is not stated. Newspapers, magazines, etc are now awash with one form or other of "statistics". Now it is fashionable to take data, shove it into a computer and come out with nice tables, graphs and ►p-values. Clearly such practices are a gross ►misuse of statistics and do a disservice to the subject. There is no wonder we are in the company of "lies, damned lies and statistics."

So What Is Statistics?

There are several definitions of statistics, some not so flattering:

1. The American heritage dictionary says: Statistics is the mathematics of collection, organization and interpretation of numerical data.
2. Brase and Brase, in their beginning level statistics textbook define statistics as the science of how to collect, organize, analyze and interpret numerical information from data.
3. Evan Esar says statistics is the only science that enables different experts using the same figures to draw different conclusions.

The first two capture the essence of statistics. Ms. Esar captures the abuse that is possible. However, these definitions do not capture the true essence of statistics and that is: to make a deduction in the face of uncertainty. The true essence of statistics is captured when it is stated that statistics is the science that tells whether something we observe can be generalized or applied to a new or different but similar situation (the author of this statement is unknown). That is I observe a group of people in a community and found that 20% have cancer, can I generalized to say that the cancer rate in that community is 20%? Of course not without first saying how the sample was observed. The other definitions come into play then. I need to know how the data was collected/observed, how it was organized, analyzed, and then the interpretation.



In this author's opinion most of the problems, misunderstandings and misrepresentations in statistics originate from the observation – collection process. Invariably the data is observed/collected before thought is put in what to do with it. So therefore the inference which is finally made does not take account of how the data was observed in the first. Maybe in the everyday sense it is natural to observe first and then ask what to do with the data observed. However in complex tasks the research questions need to be asked first. Then thought put into how to collect the relevant data, organize and analyze it and make the inference supporting the research question or refuting it. Hence in large scale work, effort should be put in the "how to collect" the data stage. If this is done, only the relevant data will be collected, and there will be savings on resources, time and money.

In most instances the way data is collected, the data type collected determines the types of analysis that can be carried out. Data collection is an expensive, time consuming activity. It is unfortunate that lots of time and effort are wasted on collecting data only to find out that the data is not useful or the exercise could have been done in an easier and cheaper manner. Should 50 experiments be performed or can 10 be sufficient? Unfortunately more data does not necessarily equate to more valid or better results. In fact the opposite could be the case. Hence the design of the experiment or data collection, the estimation of the necessary sample sizes taking into consideration the error, precision and last but not least the use to which the results will be put, such as, will the results be generalized, should be well thought out at the very beginning of the study.

Another area where statistics has a bad name is the pictorial representation of results. The saying goes that "a picture is worth a thousand words." Simple clear graphs can help bring out the important aspects of the study. However there is room for abuse. More often than not attention is not paid to the scale of the graph. For example in comparing two teaching programs, what impression is graph (a) conveying? Are our students actually better? It is the duty of statisticians to point out at every opportunity the pitfalls that need to be avoided when reading graphs.

With the advent of fast computers computations that were near impossible or would take ages to accomplish a few years ago, now takes only seconds of computer time. Coupled with this is the fact that there are very good and easy to use software. Are computers taking the place of statisticians, especially applied statisticians? There is a lot more to data analysis than calculations. The computer is there to remove the drudgery out of number crunching. What calculations to perform, that is what analysis to do and foremost, the check of the validity of assumption under which the procedures are valid, is the domain of the statistician.

Conclusion

In my view statistics is simply whether one can generalize ones observation to a different or future situation. The difficulty is how the "observation" was obtained – data collection – and the generalization made – summarized, analyzed and interpreted. In all these the expert input of a statistician is invaluable.

Cross References

►Misuse of Statistics

References and Further Reading

Brase C, Brase C (2008) Understandable statistics, 9th edn. Brooks-Cole

Evan Esar (1899–1995) Quotations www.quotationspage.com/quotes or Esar's Comic Dictionary

Misuse of Statistics

Chamont Wang
Professor
The College of New Jersey, Ewing, NJ, USA

Statistics as an academic discipline is widely held as a science that is related to experiments and the quantification of uncertainty. This is true, but if used without caution, statistics can add more uncertainty to an already murky problem. A rich source on this topic would be "*How to Lie with Statistics* Turns Fifty," a 56-page Special Section of *Statistical Science* (2005, p. 205–260).

Misuses of statistics at a non-technical level can be roughly grouped in the following three categories, often with the three types of misuses feeding each other in a complicated, dynamic fashion.

1. Data Quality: A complete statistical project consists of the following components: (a) data collection, (b) data preprocessing, (c) data exploration, (d) data analysis and statistical modeling, and (e) summary report. The process is not entirely linear and often goes from one middle step back to another, and roughly 60–95% of the project effort is needed on data quality to ensure that the entire process will not go off the rails.

   In their 2005 article, "How to Lie with Bad Data," De Veaux and Hand pointed out that "Data can be bad in an infinite variety of ways." This is not an exaggeration. Fortunately, statistical design of experiments and survey methodology, if done right, are capable of producing data with high-quality. In the real world, the problem is that the majority of data are collected in non-controlled environments without much statistical guidance. Consequently, data might have been corrupted, distorted, wrong-headed, ill-defined, and with loads of missing values – the list goes on forever. De Veaux and Hand (2005) provided suggestions on how to detect data errors and how to improve data quality. The suggestions are very useful for practitioners.

   In journals and real-world applications, statistical reports often shine with tremendous amounts of energy on exotic models but with questionable effort (and insufficient details) on data quality. Statistics as a science is supposed to provide a guiding light for research workers and decision-makers. Without good data, exotic statistical models are unlikely to help. The situation is like a person who is nearly blinded by cataracts and tries to sharpen the lenses for better vision. The effort will be futile unless an operation is conducted to take away the clouding.

   A related note on data quality is the ▶outliers and unusual numbers in the data. Resistant and robust statistical procedures are often used to handle this kind of problem. But if the data was not collected in controlled experiments, then the efforts are mostly misguided. Furthermore, outliers often are the most interesting numbers that may reveal surprising features of the study. Blind applications of ▶robust statistics thus can be counterproductive if not altogether misleading.
2. Statistical tests and ▶*p*-values: A continuing source of mistake is the confusing of *statistical significance* with *practical significance*. Mathematically, if the sample size increases indefinitely, then the power of the statistical test will increase as well. Consequently, even a tiny difference between observed and the predicted values can be statistically highly significant. Certain large scale examples regarding the confusion of *practical significance* are discussed in Wang (1993, pp. 1–2, 117–119, 128). Other cautions on the misuse of statistical tests can be found in Freedman et al. (2007) and in the "What Can Go Wrong" sections of De Veaux et al. (2009, pp. 523, 549, 570, 604–605, 634–635, 662–663, 708) which discuss "no peeking at the data" and other caveats on the tests of significance.

   Freedman (2008a) further pointed out a potential problem in research journals when publications are "driven by the search for significance." The problem can be rather acute when research grants or academic careers hinge on publications. In short, researchers may conduct many tests, ignore contradictory results and only submit findings that meet the 5% cutoff. A possibility to deal with this problem, according to Freedman (2008a), is a journal requirement to document search efforts in the research process.
3. Statistical Inference of Cause-and-Effect: Causal inference is a foundation of science and is indeed a very tricky business. As an example, Aristotle maintained that cabbages produce caterpillars daily – a well-known assertion only to be refuted by controlled experiments carried out by Francesco Redi in 1668. For new comers to the field of statistics, it may be baffling that much of the practice of modern statistics is still Aristotelian in nature. For instance, a rough estimate indicates that in clinical research, "80% of observational studies fail to replicate or the initial effects are much smaller on retest" (Young et al. 2009; a la Ioannidis 2005).

Freedman (2008a) further discussed the related controversies and a diverse set of large-scale contradictory studies. The problem should be a concern to the statistical community as our trade is indeed widely used. For example, in the study of coronary heart disease, there are more than 3,600 statistical articles published each year (Ayres 2007, p. 92), and this is only the tip of the iceberg.

A potential problem with statistical causality is the use of regression models, directed graphs, path analysis, structural equations, and other law-like relationships. Take the example of regression; on a two-dimensional scatterplot, it is easy to see that *mathematically* it does not matter whether we put a variable on the left or the right of the equation. Any software package would produce the estimates of the slope and the intercept, plus a host of diagnostic statistics that often says the model is an excellent fit. Compounding the problem of causal inference, a third variable may be the reason behind the phenomenon as displayed by the scatterplot. For instance, a scatterplot can be drawn to show that the incidence of polio (Y-variable) increases when soft-drink sales (X-variable) increases, but in fact a lurking variable (warm weather) is the driving force behind the rise (Freedman et al. 1978, p. 137).

The problem quickly turns worse in higher-dimensional spaces. Try the following example in a regression class: draw 20 or 30 right triangles and then measure the values of (X_1, X_2, Y), with X_1, X_2 being the adjacent sides of the 90° angle. The Pythagorean Theorem says that $Y = \sqrt{X_1^2 + X_2^2}$. In an experiment (Wang 1993, p. 73–77), students of regression came up with all kinds of equations with R^2 of 96–99.93%. The equations all passed stringent tests of diagnostic statistics, but none of them comes close to the Pythagorean equation. A further twist makes the problem statistically intractable when the legs of the triangles are not orthogonal (Wang 1993, p. 77–78).

For causal inference, the misgivings of statistical models happen not only in the observational studies, but also in the analysis of experimental data. In an in-depth discussion, Freedman (2008b) examined the ▶Kaplan-Meier estimator and proportional-hazards models which are frequently used to analyze data from randomized controlled experiments. Specifically, Freedman investigated journal papers on the efficacy of screening for lung cancer (*New England Journal of Medicine*), the impact of negative religious feelings on survival (*Archives of Internal Medicine*), and the efficacy of hormone replacement therapy (*New England Journal of Medicine* and *Journal of the American Medical Association*). Freedman discussed reverse causation plus a host of other issues such as measurements, omitted variables, and the justification of the models. Freedman concluded that "the models are rarely informative," that "as far as the model is concerned, the ▶randomization is irrelevant," that "randomization does not justify the model," and that it "is a mistake" to apply the models in the first place.

In yet another example, Freedman (2008c) investigated ▶logistic regression in the experimental setting for drawing conclusions on cause-and-effect. Again, Freedman noted that the model is not justified by randomization. He further questioned "Why would the logit specification be correct rather than the probit – or anything else? What justifies the choice of covariates? Why are they exogenous? If the model is wrong, what is $\widehat{\beta}_2$ supposed to be estimating?" Furthermore, in a summary of a vast variety of investigations, Freedman (2008a) concluded that "Experimental data are frequently analyzed through the prism of models. This is a mistake."

Taken together, Freedman et al. (1978, 1991, 1998, 2007), Freedman (2005, 2008a, b, c), Wang (1993, p. 72–79), and a very long list of references all indicate that sophisticated statistical models are often detached from the underlying mechanism that generated the data. In other words, many law-like equations produced by statistical models are as structure-less as *Amoeba Regression* (Wang 1993) and need to be viewed with caution. This is indeed a big disappointment to countless researchers who spend their lives on statistical models (see, e.g., Pearl 2009, p. 100), but this is a truth that we have to face.

Nevertheless, the models should be treasured for a number of reasons. To begin with, recall Newton's theory on celestial mechanics. The story is well-known and is relevant to statistical modeling in the following ways: (1) The Newtonian theory relies on observational studies, yet its prediction accuracy rivals most of the tightly controlled experiments. In other words, there is nothing wrong with observational studies, as long as they are accurate and they are consistent in subsequent studies. (2) Statistical models represent the intellectual accomplishment of the statistical community that may one day produce useful results on both experimental data and observational studies. History is the witness that ivory tower research often produces surprising results decades or hundreds of years later. And when the model is correct, the consequences can be enormous. Take the example of proportional-hazards model,

even Freedman (2008b, p. 116) acknowledged that "Precise measures of the covariates are not essential" and that if the model "is right or close to right, it works pretty well." (3) If used for descriptive or exploratory purposes, fancy statistical models may indeed reveal unexpected features in the data. For certain examples on non-parametric structural equations and counterfactual analysis, see references in Pearl (2009). For another example on hot spot detection, see Wang et al. (2008).

As a matter of fact, in the past 15 years or so, statistical models have taken a new life in the realm of ►data mining, predictive modeling, and statistical learning (see, e.g., Wang et al. 2008). In these applications, the concerns are not cause-and-effect or the specific mechanism that generates the data. Instead, the focus is the prediction accuracy that can be measured by profit, false positive, false negative, and by other criteria to assess the model utility. This is a sharp departure from causation to prediction. The great news is that the new applications have been ranked by the 2001 *MIT Technology Review* as one of the ten emerging technologies that will change the world – and it is arguable that the successes of this new technology will eventually feedback to traditional statistics for other breakthroughs. In fact, countless examples with ingenious twists have already happened (see, e.g., Ayres 2007). It is a triumph of statistical models.

A cautionary note is that statistical learning and the new breed of predictive modeling can easily go wrong and misinformation can propagate with unprecedented speed in the modern age of internet blogging and social networks. Newcomers to the field should consult, for examples, "Top 10 Data Mining Mistakes" (Elder 2009) and "Myths and Pitfalls of Data Mining" (Khabaza 2009). For unsupervised learning, one may want to read "The Practice of Cluster Analysis" (Kettenring, 2006) and "A Perspective on Cluster Analysis" (Kettenring 2008). For supervised learning, given a dozen or thousands of predictors, statistical tools are frequently used to generate predictor importance scores, but these scores are often wildly different from one algorithm to the next (see e.g., Wang et al. 2008, Sect. 4).

For yet another example, a model such as a Neural Network may produce higher profit and higher prediction accuracy than other tools, yet the model may also be more volatile in repeated uses and hence pose considerable hazards in the long run. ►Sensitivity analysis and similar techniques are thus needed to prevent misleading conclusions (see, e.g., Wang et al. 2009).

The hallmark of empirical science is its replicability. Much of the current statistical practice, unfortunately, does not really meet this criterion. Just look at how many authors are unwilling to disclose their data and how many journals are unwilling to archive the datasets and the code (see also Freedman, 2008a, c). Exceptions include *American Economic Review, American Economic Journals* and *Science*.

Data disclosure reduces the cost of research and cost of replicating results. It also deters unprofessional conduct and improves collective findings of the research community. Certain online journals (see e.g., http://www.bentley.edu/csbigs/csbigs-v1-n1.cfm) post both the research article and the data side-by-side. If more journals are willing to make available the datasets used in their publications, the situation of misuse and misconduct of statistics will be greatly improved.

About the Author

Dr. Chamont Wang received the Ph.D. degree in Statistics from Michigan State University, East Lansing (1983). He is Full Professor at the Department of Mathematics and Statistics, the College of New Jersey, serving as an Associate Editor of a research journal, CSBIGS (*Case Studies in Business, Industry and Government Statistics*), serving as an expert witness of a premier expert witness referral firm. He is author of the book, *Sense and nonsense of statistical inference: controversy, misuse, and subtlety* (Taylor and Francis, 1993), and also of journal papers in the field of Chaos and Dynamical Systems. He is a member of American Statistical Association, the Mathematical Association of America, and the Institute of Mathematical Statistics.

Cross References

►Discriminant Analysis: Issues and Problems
►Economic Growth and Well-Being: Statistical Perspective
►Fraud in Statistics
►Misuse and Misunderstandings of Statistics
►Role of Statistics
►Significance Tests: A Critique
►Statistical Fallacies
►Statistical Fallacies: Misconceptions, and Myths
►Statistics and the Law
►Statistics: Controversies in Practice

References and Further Reading

Ayres I (2007) Super crunchers: why thinking-by-numbers is the new way to be smart. Bantom, New York
De Veaux R, Hand D (2005) How to lie with bad data. Stat Sci 20(3):231–238
De Veaux R, Velleman P, Bock D (2009) Intro Stats, 3rd edn. Pearson
Elder JF IV (2009) Top 10 data mining mistakes. Handbook of statistical analysis and data mining applications, Elsevier, pp 733–754

Freedman D (2005) Statistical models: theory and practice. Cambridge University Press, Cambridge
Freedman DA (2008a) Oasis or mirage? Chance 21(1):59–61
Freedman DA (2008b) Survival analysis: a primer. Am Stat 62(2):110–119
Freedman DA (2008c) Randomization does not justify logistic regression. Stat Sci 23(2):237–249
Freedman DA, Pisani R, Purves R (1978, 1991, 1998, 2007) Statistics. W.W. Norton, USA
Ioannidis J (2005) Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. J Am Med Assoc 294:218–228
Kettenring JR (2006) The Practice of Cluster Analysis. Journal of Classif 23(1):3–30
Kettenring JR (2008) A Perspective on Cluster Analysis. Stat Anal Data Mining 1(1):52–53
Khabaza T (2009) Hard hat area: myths and pitfalls of data mining. An *SPSS* Executive Brief, http://viewer.bitpipe.com/viewer/viewDocument.do?accessId=10318929
Pearl J (2009) Causal inference in statistics: an overview. Stat Surv 3:96–146, http://www.i-journals.org/ss/
Wang C (1993) Sense and nonsense of statistical inference: controversy, misuse, and subtlety. Marcel Dekker, Inc., New York
Wang C, Liu B (2008) Data mining for large datasets and hotspot detection in an urban development project. J Data Sci 6(3): 389–414. http://proj1.sinica.edu.tw/~jds/JDS-501.pdf
Wang C, Zhuravlev M (2009) An analysis of profit and customer satisfaction in consumer finance. Case Stud Bus Indus Govern Stat 2(2):147–156, http://www.bentley.edu/csbigs/documents/Wang.pdf
Young SS, Bang H, Oktay K (2009) Cereal-induced gender selection? Most likely a multiple testing false positive. Proc R Soc B 276:1211–1212

Mixed Membership Models

Elena A. Erosheva[1], Stephen E. Fienberg[2]
[1]Associate Professor
University of Washington, Seattle, WA, USA
[2]Maurice Falk University Professor
Carnegie Mellon University, Pittsburgh, PA, USA

The notion of mixed membership arises naturally in the context of multivariate data analysis (see ▶Multivariate Data Analysis: An Overview) when attributes collected on individuals or objects originate from a mixture of different categories or components. Consider, for example, an individual with both European and Asian ancestry whose mixed origins correspond to a statement of mixed membership: "1/4 European and 3/4 Asian ancestry." This description is conceptually very different from a probability statement of "25% chance of being European and 75% chance of being Asian". The assumption that individuals or objects may combine attributes from several basis categories in a stochastic manner, according to their proportions of membership in each category, is a distinctive feature of mixed membership models. In most applications, the number and the nature of the basis categories, as well as individual membership frequencies, are typically considered latent or unknown. Mixed membership models are closely related to latent class and finite ▶mixture models in general. Variants of these models have recently gained popularity in many fields, from genetics to computer science.

Early Developments

Mixed membership models arose independently in at least three different substantive areas: medical diagnosis and health, genetics, and computer science. Woodbury et al. (1978) proposed one of the earliest mixed membership models in the context of disease classification, known as the *Grade of Membership* or GoM model. The work of Woodbury and colleagues on the GoM model is summarized in the volume *Statistical Applications Using Fuzzy Sets* (Manton et al. 1994).

Pritchard et al. (2000) introduced a variant of the mixed membership model which became known in genetics as the *admixture model* for multilocus genotype data and produced remarkable results in a number of applications. For example, in a study of human population structure, Rosenberg et al. (2002) used admixture models to analyze genotypes from 377 autosomal microsatellite loci in 1,056 individuals from 52 populations. Findings from this analysis indicated a typology structure that was very close to the "traditional" five main racial groups.

Among the first mixed membership models developed in computer science and machine learning for analyzing words in text documents were a multivariate analysis method named Probabilistic Latent Semantic Analysis (Hofmann 2001) and its random effects extension by Blei et al. (2003a, b). The latter model became known as *Latent Dirichlet Allocation* (LDA) due to the imposed Dirichlet distribution assumption for the mixture proportions. Variants of LDA model in computer science are often referred to as *unsupervised generative topic models*. Blei et al. (2003a, b) and Barnard et al. (2003) used LDA to combine different sources of information in the context of analyzing complex documents that included words in main text, photographic images, and image annotations. Erosheva et al. (2004) analyzed words in abstracts and references in bibliographies from a set of research reports published in the *Proceeding of the National Academy of Sciences* (PNAS), exploring

an internal mixed membership structure of articles and comparing it with the formal PNAS disciplinary classifications. Blei and Lafferty (2007) developed another mixed membership model replacing the Dirichlet assumption with a more flexible logistic normal distribution for the mixture proportions. Mixed membership developments in machine learning have spurred a number of applications and further developments of this class of models in psychology and cognitive sciences where they became known as *topic models* for semantic representations (Griffiths et al. 2007).

Basic Structure

The basic structure of a mixed membership model follows from the specification of assumptions at the population, individual, and latent variable levels, and the choice of a sampling scheme for generating individual attributes (Erosheva et al. 2004). Variations in these assumptions can provide us with different mixed membership models, including the GoM, admixture, and generative topic models referred to above.

Assume K basis subpopulations. For each subpopulation $k = 1, \ldots, K$, specify $f(x_j|\theta_{kj})$, a probability distribution for attribute x_j, conditional on a vector of parameters θ_{kj}. Denote individual-level membership score vector by $\lambda = (\lambda_1, \ldots, \lambda_K)$, representing the mixture proportions in each subpopulation. Given λ, the subject-specific conditional distribution for *jth* attribute is

$$Pr(x_j|\lambda) = \sum_k \lambda_k f(x_j|\theta_{kj}).$$

In addition, assume that attributes x_j are independent, conditional on membership scores. Assume membership scores, the latent variables, are random realizations from some underlying distribution D_α, parameterized by α. Finally, specify a sampling scheme by picking the number of observed distinct attributes, J, and the number of independent replications for each attribute, R.

Combining these assumptions, the marginal probability of observed responses $\left\{x_1^{(r)}, \ldots, x_J^{(r)}\right\}_{r=1}^{R}$, given model parameters α and $\boldsymbol{\theta}$, is

$$Pr\left(\left\{x_1^{(r)}, \ldots, x_J^{(r)}\right\}_{r=1}^{R} | \alpha, \boldsymbol{\theta}\right) = \int \left(\prod_{j=1}^{J}\prod_{r=1}^{R}\sum_{k=1}^{K}\lambda_k f\left(x_j^{(r)}|\theta_{kj}\right)\right) dD_\alpha(\lambda). \quad (1)$$

In general, the number of observed attributes need not be the same across subjects, and the number of replications need not be the same across attributes. In addition, instead of placing a probability distribution on membership scores, some mixed membership model variants may treat latent variables as fixed but unknown constants. Finally, other extensions can be developed by specifying further dependence structures among sampled individuals or attributes that may be driven by particular data forms as, e.g., in relational or network data (Airoldi et al. 2008b; Chang and Blei 2010; Xing et al. 2010).

Estimation

A number of estimation methods have been developed for mixed membership models that are, broadly speaking, of two types: those that treat membership scores as fixed and those that treat them as random. The first group includes the numerical methods introduced by Hofmann (2001), and joint maximum likelihood type methods described in Manton et al. (1994) and Cooil and Varki (2003), and related likelihood approaches in Potthoff et al. (2000) and Varki et al. (2000). The statistical properties of the estimators in these approaches, such as consistency, identifiability, and uniqueness of solutions, are yet to be fully understood (Haberman 1995) – empirical evidence suggests that the likelihood function is often multi-modal and can have bothersome ridges. The second group uses Bayesian hierarchical structure for direct computation of the posterior distribution, e.g., with Gibbs sampling based on simplified assumptions (Pritchard et al. 2000; Griffiths and Steyvers 2004) or with fully Bayesian MCMC sampling (Erosheva 2003). Variational methods used by Blei et al. (2003a, b), or expectation-propagation methods developed by Minka and Lafferty (2002), can be used to approximate the posterior distribution. The Bayesian hierarchical methods solve some of the statistical and computational problems, and variational methods in particular scale well for higher dimensions. Many other aspects of working with mixed membership models remain as open challenges, e.g., dimensionality selection (Airoldi et al. 2008a).

Relationship to Other Methods of Multivariate Analysis

It is natural to compare mixed membership models with other latent variable methods, and, in particular, with factor analysis and latent class models (Bartholomew and Knott 1999). For example, the GoM model for binary outcomes can be thought of as a constrained factor analysis model: $E(x|\lambda) = A\lambda$, where x is a column-vector of observed attributes $x = (x_1, \ldots, x_J)'$, $\lambda = (\lambda_1, \ldots, \lambda_K)'$ is a column-vector of factor (i.e., membership) scores, and A is

a $J \times K$ matrix of factor loadings. The respective constraints in this factor model are $\lambda' I_K = 1$ and $AI_K = I_K$, where I_K is a K-dimensional vector of 1s.

Mixed membership models can also address objectives similar to those in ►Correspondence Analysis and Multidimensional Scaling methods for contingency tables. Thus, one could create a low-dimensional map from a contingency table data and graphically examine membership scores (representing table rows or individuals) in the convex space defined by basis or extreme profiles (representing columns or attributes) to address questions such as whether some table rows have similar distribution over the table columns categories.

Finally, there is a special relationship between the sets of mixed membership and latent class models, where each set of models can be thought of as a special case of the other. Manton et al. (1994) and Potthoff et al. (2000) described how GoM model can be thought of as an extension of latent class models. On the other hand, Haberman (1995) first pointed out that GoM model can be viewed as a special case of latent class models. The fundamental representation theorem of equivalence between mixed membership and population-level mixture models clarifies this nonintuitive relationship (Erosheva et al. 2007).

About the Authors

Elena Erosheva is a Core member of the Center for Statistics and the Social Sciences, University of Washington.

For biography of Professor Fienberg *see* the entry ►Data Privacy and Confidentiality.

Acknowledgments

Supported in part by National Institutes of Health grant No. R03 AG030605-01 and by National Science Foundation grant DMS-0631589.

Cross References

►Correspondence Analysis
►Factor Analysis and Latent Variable Modelling
►Multidimensional Scaling
►Multivariate Data Analysis: An Overview

References and Further Reading

Airoldi EM, Blei DM, Fienberg SE, Xing EP (2008a) Mixed-membership stochastic blockmodels. J Mach Learn Res 9:1981–2014

Airoldi EM, Fienberg SE, Joutard C, Love TM (2008b) Discovery of latent patterns with hierarchical Bayesian mixed-membership models and the issue of model choice. In: Poncelet P, Masseglia F, Teisseire M (eds) Data mining patterns: new methods and applications. pp 240–275

Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei DM, Jordan MI (2003) Matching words and pictures. J Mach Learn Res 3: 1107–1135

Bartholomew DJ, Knott M (1999) Latent variable models and factor analysis, 2nd edn. Arnold, London

Blei DM, Lafferty JD (2007) A correlated topic model of Science. Ann Appl Stat 1:17–35

Blei DM, Ng AY, Jordan MI (2003a) Latent Dirichlet allocation. J Mach LearnRes 3:993–1002

Blei DM, Ng AY, Jordan MI (2003b) Modeling annotated data. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp 127–134

Chang J, Blei DM (2010) Hierarchical relational models for document networks. Ann Appl Stat 4, pp 124–150

Cooil B, Varki S (2003) Using the conditional Grade-of-Membership model toassess judgement accuracy. Psychometrika 68:453–471

Erosheva EA (2003) Bayesian estimation of the Grade of Membership Model. In: Bernardo J et al (eds) Bayesian statistics 7. Oxford University Press, Oxford, pp 501–510

Erosheva EA, Fienberg SE (2004) Partial membership models with application to disability survey data. In: Weihs C, Caul W (eds) Classification – the ubiquitous challenge. Springer, Heidelberg, pp 11–26

Erosheva EA, Fienberg SE, Lafferty J (2004) Mixed membership models of scientific publications. Proc Natl Acad Sci 101 (suppl 1):5220–5227

Erosheva EA, Fienberg SE, Joutard C (2007) Describing disability through individual-level mixture models for multivariate binary data. Ann Appl Stat 1:502–537

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101 (suppl 1):5228–5235

Griffiths TL, Steyvers M, Tenenbaum JB (2007) Topics in Semantic Representation. Psychol Rev 114(2):211–244

Haberman SJ (1995) Book review of "Statistical applications using fuzzy sets," by K.G. Manton, M.A. Woodbury and H.D. Tolley. J Am Stat Assoc 90:1131–1133

Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42:177–196

Manton KG, Woodbury MA, Tolley HD (1994) Statistical applications using fuzzy sets. Wiley, New York

Minka TP, Lafferty JD (2002) Expectation-propagation for the generative aspect model. In: Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI–2002), Morgan Kaufmann, San Francisco, pp 352–359

Potthoff RF, Manton KG, Woodburry MA (2000) Dirichlet generalizations of latent-class models. J Classif 17:315–353

Pritchard P, Stephens JK, Donnely P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Varki S, Cooil B, Rust RT (2000) Modeling fuzzy data in qualitative marketing research. J Market Res 37:480–489

Woodbury MA, Clive J, Garson A (1978) Mathematical typology: a grade of membership technique for obtaining disease definition. Comput Biomed Res 11:277–298

Xing E, Fu W, Song L (2010) A state-space mixed membership blockmodel for dynamic network tomography. Ann Appl Stat 4, in press

Mixture Models

Wilfried Seidel
Professor, President of the German Statistical Society
Helmut-Schmidt-Universität, Hamburg, Germany

Introduction

Mixture distributions are convex combinations of "component" distributions. In statistics, these are standard tools for modeling heterogeneity in the sense that different elements of a sample may belong to different components. However, they may also be used simply as flexible instruments for achieving a good fit to data when standard distributions fail. As good software for fitting mixtures is available, these play an increasingly important role in nearly every field of statistics.

It is convenient to explain finite mixtures (i.e., finite convex combinations) as theoretical models for cluster analysis (see ▶Cluster Analysis: An Introduction), but of course the range of applicability is not at all restricted to the clustering context. Suppose that a feature vector X is observed in a heterogeneous population, which consists of k homogeneous subpopulations, the "components." It is assumed that for $i = 1, \ldots, k$, X is distributed in the i-th component according to a (discrete or continuous) density $f(x, \theta_i)$ (the "component density"), and all component densities belong to a common parametric family $\{f(x, \theta), \ \theta \in \Theta\}$, the "component model." The relative proportion of the i-th component in the whole population is p_i, $p_1 + \cdots + p_k = 1$. Now suppose that an item is drawn randomly from the population. Then it belongs to the i-th component with probability p_i, and the conditional probability that X falls in some set A is $\Pr(X \in A \mid \theta_i)$, calculated from the density $f(x, \theta_i)$. Consequently, the marginal probability is

$$\Pr(X \in A \mid P) = p_1 \Pr(X \in A \mid \theta_1) + \cdots + p_k \Pr(X \in A \mid \theta_k)$$

with density

$$f(x, P) = p_1 f(x, \theta_1) + \cdots + p_k f(x, \theta_k), \qquad (1)$$

a "simple finite mixture" with parameter $P = ((p_1, \ldots, p_k), (\theta_1, \ldots, \theta_k))$. The components p_i of P are called "mixing weights," the θ_i "component parameters." For fixed k, let $\mathcal{P}_k$ be the set of all vectors P of this type, with $\theta_i \in \Theta$ and nonnegative mixing weights summing up to one. Then $\mathcal{P}_k$ parameterizes all mixtures with not more than k components. If all mixing weights are positive and component densities are different, then k is the exact number of components. The set of all simple finite mixtures is parameterized by $\mathcal{P}_{\text{fin}}$, the union of all $\mathcal{P}_k$.

This model can be extended in various ways. For example, all component densities may contain additional common parameters (variance parameters, say), they may depend on covariables (mixtures of regression models), and also the mixing weights may depend on covariables. Mixtures of time series models are also considered. Here I shall concentrate on simple mixtures, as all relevant concepts can be explained very easily in this setting. These need not be finite convex combinations; there is an alternative and more general definition of simple mixtures: Observe that the parameter P can be considered as a discrete probability distribution on Θ which assigns probability mass p_i to the parameter θ_i. Then Eq. 1 is an integral with respect to this distribution, and if ξ is an arbitrary probability distribution on Θ, a mixture can be defined by

$$f(x, \xi) = \int_\Theta f(x, \theta)\, d\xi(\theta)\,. \qquad (2)$$

It can be considered as the distribution of a two-stage experiment: First, choose a parameter θ according to the distribution ξ, then choose x according to $f(x, \theta)$. Here, ξ is called a "mixing distribution," and mixture models of this type can be parameterized over every set Ξ of probability distributions on Θ.

In statistical applications of mixture models, a nontrivial key issue is identifiability, meaning that different parameters describe different mixtures. In a trivial sense, models parameterized over vectors P are never identifiable: All vectors that correspond to the same probability distribution on Θ describe the same mixture model. For example, any permutation of the sequence of components leaves the mixing distribution unchanged, or components may be added with zero mixing weights. Therefore identifiability can only mean that parameters that correspond to different mixing distributions describe different mixture models. However, also in this sense identifiability is often violated. For example, the mixture of two uniform distributions with supports $[0, 0.5]$ and $[0.5, 1]$ and equal mixing weights is the uniform distribution with support $[0, 1]$. On the other hand, finite mixtures of many standard families (normal, Poisson, ...) are identifiable, see for example Titterington et al. (1985). Identifiability of mixtures of regression models has been treated among others by Hennig (2000). A standard general reference for finite mixture models is McLachlan and Peel (2000).

Statistical Problems

Consider a mixture model with parameter η (vector or probability measure). In the simplest case, one has i.i.d.

data $x_1, \ldots, x_n$ from $f(x, \eta)$, from which one wants to gain information about η. Typical questions are estimation of (parameters of) η, or mixture diagnostics: Is there strong evidence for a mixture (in contrast to homogeneity in the sense that η is concentrated at some single parameter θ)? What is the (minimum) number of mixture components?

A variety of techniques has been developed. The data provide at least implicitly an estimate of the mixture, and Eqs. 1 and 2 show that mixture and mixing distribution are related by a linear (integral) equation. Approximate solution techniques have been applied for obtaining estimators, and moment estimators have been developed on basis of this structure. Distance estimators exhibit nice properties. Traditionally, mixture diagnostics has been handled by graphical methods. More recent approaches for estimation and diagnostics are based on Bayesian or likelihood techniques; likelihood methods will be addressed below. Although Bayesian methods have some advantages over likelihood methods, they are not straightforward (for example, usually no "natural" conjugate priors are available, therefore posteriors are simulated using MCMC. Choice of "noninformative" priors is not obvious, as improper priors usually lead to improper posteriors. Nonidentifiability of $\mathcal{P}_k$ causes the problem of "label switching"). A nice reference for Bayesian methods is Frühwirth-Schnatter (2006).

Let me close this section with a short discussion of robustness. Robustness with respect to ▶outliers is treated by Hennig (2004). Another problem is that mixture models are extremely nonrobust with respect to misspecification of the component model. Estimating the component model in a fully nonparametric way is of course not possible, but manageable alternatives are for example mixtures of log-concave distributions. Let me point out, however, that issues like nonrobustness and nonidentifiability only cause problems if the task is to interpret the model parameters somehow. If the aim is only to obtain a better data fit, one need not worry about them.

Likelihood Methods

In the above setting, $l(\eta) = \log(f(x_1, \eta)) + \cdots + \log(f(x_n, \eta))$ is the log likelihood function. It may have some undesirable properties: First, the log likelihood is often unbounded. For example, consider mixtures of normals. If the expectation of one component is fixed at some data point and the variance goes to zero, the likelihood goes to infinity. Singularities usually occur at the boundary of the parameter space. Second, the likelihood function is usually not unimodal, although this depends on the parameterization. For example, if the parameter is a probability distribution as in Eq. 2 and if the parameter space Ξ is a convex set (with respect to the usual linear combination of measures), the log likelihood function is concave. If it is bounded, there is a nice theory of "nonparametric likelihood estimation" (Lindsay 1995), and "the" "nonparametric maximum likelihood estimator" is in some sense uniquely defined and can be calculated numerically (Böhning 2000; Schlattmann 2009).

Nonparametric methods, however, work in low dimensional component models, whereas "parametric" estimation techniques like the Expectation-Maximization (EM) method work in nearly any dimensional. The EM is a local maximizer for mixture likelihoods in $\mathcal{P}_k$. Here the mixture likelihood is usually multimodal; moreover, it can be very flat. Analytic expressions for likelihood maxima usually do not exist, they have to be calculated numerically. On the other hand, even for unbounded likelihoods, it is known from asymptotic theory, that the simple heuristics of searching for a large local maximum in the interior of the parameter space may lead to reasonable estimators. However, one must be aware that there exist "spurious" large local maxima that are statistically meaningless. Moreover, except from simple cases, there is no manageable asymptotics for likelihood ratio.

Some of the problems of pure likelihood approaches can be overcome by considering penalized likelihoods. However, here one has the problem of choosing a penalization parameter. Moreover, the EM algorithm is a basic tool for a number of estimation problems, and it has a very simple structure for simple finite mixtures. Therefore it will be outlined in the next section.

EM Algorithm

The EM algorithm is a local maximization technique for the log likelihood in $\mathcal{P}_k$. It starts from the complete-data log-likelihood. Suppose that for observation x_i the (fictive) component membership is known. It is defined by a vector $z_i \in \Re^k$ with $z_{ij} = 1$, if x_i belongs to j-th component, and zero elsewhere. As a random variable Z_i, it has a ▶multinomial distribution with parameters $k, p_1, \ldots, p_k$. Then the complete data likelihood and log likelihood of P, respectively, are $L_c(P) = \prod_{i=1}^n \prod_{j=1}^k (p_j f(x_i, \theta_j))^{z_{ij}}$ and $l_c(P) = \log(L_c(P)) = \sum_{i=1}^n \sum_{j=1}^k z_{ij} \log p_j + \sum_{i=1}^n \sum_{j=1}^k z_{ij} \log f(x_i, \theta_j)$.

The EM needs a starting value P_0, and then proceeds as an iteration between an "E-step" and an "M-step" until "convergence." The first E-step consists in calculating the conditional expectation $E_{P_0}(l_c(P) \mid x_1, \ldots, x_n)$ of $l_c(P)$ for

arbitrary P, given the data, under P_0. As the only randomness is in the z_{ij}, we obtain

$$E_{P_0}(l_c(P)\,|\,x_1,\ldots,x_n) = \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_j(x_i|P_0)\log p_j + \sum_{i=1}^{n}\sum_{j=1}^{k}\tau_j(x_i|P_0)\log f(x_i,\,\theta_j),$$

where

$$\tau_j(x_i|P_0) = \Pr_{P_0}(Z_{ij}=1\,|\,x_i) = \frac{p_j f(x_i,\,\theta_j)}{f(x_i,\,P_0)}$$

is the conditional probability that the i-th observation belongs to component j, given the data, with respect to P_0.

In the following M-step, $E_{P_0}(l_c(P)\,|\,x_1,\ldots,x_n)$ is maximized with respect to P. As it is the sum of terms depending on the mixing weights and on the parameters only, respectively, both parts can be maximized separately. It is easily shown that the maximum in the p_j is achieved for $p_j^{(1)} = (1/n)\sum_{i=1}^{n}\tau_j(x_i|P_0), j = 1,\ldots,n$. For component densities from exponential families, similar simple solutions exist for the θ_j, therefore both the E-step and the M-step can be carried out here analytically. It can be shown that (1) the log-likelihood is not decreasing during the iteration of the EM, and (2) that under some regularity conditions it converges to a stationary point of the likelihood function. However, this may also be a saddle point.

It remains to define the stopping rule and the starting point(s). Both are crucial, and the reader is referred to the literature. There are also techniques that prevent from convergence to singularities or spurious maxima. A final nice issue of the EM is that it yields a simple tool for classification of data points: If $\hat{P}$ is an estimator, then $\tau_j(x_i|\hat{P})$ is the posterior probability that x_i belongs to class j with respect to the "prior" $\hat{P}$. The Bayesian classification rule assigns observation i to the class j that maximizes $\tau_j(x_i|\hat{P})$, and the $\tau_j(x_i|\hat{P})$ measure the plausibility of such a clustering.

Number of Components, Testing and Asymptotics

Even if one has an estimator in each $\mathcal{P}_k$ from the EM, the question is how to assess the number of components (i.e., how to choose k). Usually information criteria like AIC and BIC are recommended. An alternative is to perform a sequence of tests of k against $k+1$ components, for $k = 1, 2\ldots$.

There are several tests for homogeneity, i.e., for the "component model", as for example goodness of fit or dispersion score tests. For testing k_0 against k_1 components, a likelihood ratio test may be performed. However, the usual χ^2-asymptotics fails, so critical values have to be simulated. Moreover, the distribution of the test statistic usually depends on the specific parameter under the null hypothesis. Therefore some sort of bootstrap (see ►Bootstrap Methods) is needed, and as estimators have to be calculated numerically, likelihood ratio tests are computationally intensive.

Let me close with some remarks on asymptotics. Whereas ►asymptotic normality of estimators is guaranteed under some conditions, the usual asymptotics for the likelihood ratio test fails. The reason is that under the null hypothesis, the parameter P_0 is on the boundary of the parameter space, it is not identifiable and the Fisher information matrix in P_0 is singular. There is an asymptotic theory under certain restrictive assumptions, but it is usually hard to calculate critical values from it.

About the Author

Professor Seidel was the Editor of "*AStA – Advances of Statistical Analysis*" (Journal of the German Statistical Society) (2004–2008). He is Dean of the Faculty of Economics and Social Sciences of Helmut-Schmidt-Universität (since January 2009), and has been elected next President of Helmut-Schmidt-University, starting in October 2010.

Cross References

►Bayesian Statistics
►Contagious Distributions
►Identifiability
►Likelihood
►Modeling Count Data
►Multivariate Statistical Distributions
►Nonparametric Estimation
►Optimum Experimental Design

References and Further Reading

Böhning D (2000) Finite mixture models. Chapman and Hall, Boca Raton
Frühwirth-Schnatter S (2006) Finite mixture and Markov switching models. Springer, New York
Hennig C (2000) Identifiability of models for clusterwise linear regression. J Classif 17:273–296
Hennig C (2004) Breakdown points for ML estimators of location-scale mixtures. Ann Stat 32:1313–1340
Lindsay BG (1995) Mixture models: theory, geometry and applications. NSC-CBMS Regional Conference Series in Probability and Statistics, 5
McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
Schlattmann P (2009) Medical applications of finite mixture models. Springer, Berlin
Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions, Wiley, New York

Model Selection

Walter Zucchini[1], Gerda Claeskens[2], Georges Nguefack-Tsague[3]
[1]Professor
Georg-August-Universität, Göttingen, Germany
[2]Professor
Leuven, Belgium
[3]University of Yaoundé I, Yaoundé, Cameroon

Introduction

In applications there are usually several models for describing a population from a given sample of observations and one is thus confronted with the problem of model selection. For example, different distributions can be fitted to a given sample of univariate observations; in polynomial regression one has to decide which degree of the polynomial to use; in multivariate regression one has to select which covariates to include in the model; in fitting an autoregressive model to a stationary time series one must choose which order to use.

When the set of models under consideration is nested, as is the case in polynomial regression, the fit of the model to the sample improves as the complexity of the model (e.g., the number of parameters) increases but, at some stage, its fit to the population deteriorates. That is because the model increasingly moulds itself to the features of the sample rather than to the "true model," namely the one that characterizes the population. The same tendency occurs even if the models are not nested; increasing the complexity eventually leads to deterioration. Thus model selection needs to take both goodness of the fit and the complexity of the competing models into account.

Reference books on model selection include Linhart and Zucchini (1986), Burnham and Anderson (2002), Miller (2002), Claeskens and Hjort (2008). An introductory article is Zucchini (2000).

Information Criteria – Frequentist Approach

The set of models considered for selection can be thought of as approximating models which, in general, will differ from the true model. The answer to the question "Which approximation is best?" depends, of course, on how we decide to measure the quality of the fit. Using the Kullback-Leibler distance for this leads to the popular ▶Akaike Information Criterion (AIC, Akaike 1973):

$$AIC(M) = 2\log(L(\hat{\theta})) - 2p,$$

where M is the model, L the likelihood, and $\hat{\theta}$ the maximum likelihood estimator of the vector of the model's p parameters. The first term of the AIC measures the fit of the model to the *observed sample*; the fit improves as the number of parameters in the model is increased. But improving the fit of the model to the sample does not necessarily improve its fit to the population. The second term is a penalty term that compensates for the complexity of the model. One selects the model that maximizes the AIC. Note, however, that in much of the literature the AIC is defined as minus the above expression, in which case one selects the model that minimizes it.

A *model selection criterion* is a formula that allows one to compare models. As is the case with the AIC, such criteria generally comprise two components: one that quantifies the fit to the data, and one that penalizes complexity. Examples include Mallows' C_p criterion for use in ▶linear regression models, Takeuchi's model-robust information criterion TIC, and refinements of the AIC such as the 'corrected AIC' for selection in linear regression and autoregressive time series models, the network information criterion NIC, which is a version of AIC that can be applied to model selection in ▶neural networks, and the generalized information criterion GIC for use with influence functions. Several of these criteria have versions that are applicable in situations where there are outlying observations, leading to robust model selection criteria; other extensions can deal with missing observations.

Alternative related approaches to model selection that do not take the form of an information criterion are *bootstrap* (see, e.g., Zucchini 2000) and *cross-validation*. For the latter the idea is to partition the sample in two parts: the calibration set, that is used to fit the model, and the validation sample, that is used to assess the fit of the model, or the accuracy of its predictions. The popular "leave-one-out cross-validation" uses only one observation in the validation set, but each observation has a turn at comprising the validation set. In a model selection context, we select the model that gives the best results (smallest estimation or prediction error) averaged over the validation sets. As this approach can be computationally demanding, suggestions have been made to reduce the computational load. In "five-fold cross-validation" the sample is randomly split in five parts of about equal size. One of the five parts is used as validation set and the other four parts as the calibration set. The process is repeated until each of the five sets is used as validation set.

Bayesian Approach

The Bayesian regards the models available for selection as candidate models rather than approximating models; each of them has the potential of being the true model. One begins by assigning to each of them a prior probability, $P(M)$, that it is the true model and then, using ▶Bayes'

theorem, computes the posterior probability of it being so:

$$P(M|\text{ Data}) = \frac{P(\text{ Data}|M)P(M)}{P(\text{ Data})}.$$

The model with the highest posterior probability is selected. The computation of $P(\text{ Data}|M)$ and $P(\text{ Data})$ can be very demanding and usually involves the use of Markov chain Monte Carlo (MCMC) methods (see ►Markov Chain Monte Carlo) because, among other things, one needs to 'integrate out' the distribution of the parameters of M (see e.g., Wasserman 2000).

Under certain assumptions and approximations (in particular the Laplace approximation), and taking all candidate models as a priori equally likely to be true, this leads to the Bayesian Information Criterion (BIC), also known as the Schwarz criterion (Schwarz 1978):

$$\text{BIC}(M) = 2\log(L(\hat{\theta})) - p\log(n),$$

where n is the sample size and p the number of unknown parameters in the model. Note that although the BIC is based on an entirely different approach it differs from the AIC only in the penalty term.

The difference between the frequentist and Bayesian approaches can be summarized as follows. The former addresses the question "Which model is best, in the sense of least wrong?" and the latter the question "Which model is most likely to be true?".

The Deviance Information Criterion (Spiegelhalter et al. 2002) is an alternative Bayesian method for model selection. While explicit formulae are often difficult to obtain, its computation is simple for situations where MCMC simulations are used to generate samples from a posterior distribution.

The principle of minimum description length (MDL) is also related to the BIC. This method tries to measure the complexity of the models and selects the model that is the least complex. The MDL tries to minimize the sum of the description length of the model, plus the description length of the data when fitted to the model. Minimizing the description length of the data corresponds to maximizing the log likelihood of the model. The description length of the model is not uniquely defined but, under certain assumptions, MDL reduces to BIC, though this does not hold in general (Rissanen 1996). Other versions of MDL come closer to approximating the full Bayesian posterior $P(M|)$ Data. See Grünwald (2007) for more details.

Selecting a Selection Criterion

Different selection criteria often lead to different selections. There is no clear-cut answer to the question of which criterion should be used. Some practitioners stick to a single criterion; others take account of the orderings indicated by two or three different criteria (e.g., AIC and BIC) and then select the one that leads to the model which seems most plausible, interpretable or simply convenient in the context of the application.

An alternative approach is to tailor the criterion to the particular objectives of the study, i.e., to construct it in such a way that selection favors the model that best estimates the quantity of interest. The Focussed Information Criterion (FIC, Claeskens and Hjort 2003) is designed to do this; it is based on the premise that a good estimator has a small mean squared error (MSE). The FIC is constructed as an estimator of the MSE of the estimator of the quantity of interest. The model with the smallest value of the FIC is the best.

Issues such as consistency and efficiency can also play a role in the decision regarding which criterion to use. An information criterion is called *consistent* if it is able to select the true model from the candidate models, as the sample size tends to infinity. In a weak version, this holds with probability tending to one; for strong consistency, the selection of the true model is almost surely. It is important to realize that the notion of consistency only makes sense in situations where one can assume that the true model belongs to the set of models available for selection. Thus will not be the case in situations in which researchers "believe that the system they study is infinitely complicated, or there is no way to measure all the important variables" (McQuarrie and Tsai 1998). The BIC is a consistent criterion, as is the Hannan-Quinn criterion that uses $\log\log(n)$ instead of $\log(n)$ in the penalty term.

An information criterion is called *efficient* if the ratio of the expected mean squared error (or expected prediction error) under the selected model and the expected mean squared error (or expected prediction error) under its theoretical minimizer converges to one in probability. For a study of the efficiency of a model selection criterion, we do not need to make the assumption that the true model is one of the models in the search list. The AIC, corrected AIC, and Mallows's C_p are examples of efficient criteria. It can be shown that the BIC and the Hannan-Quinn criterion are not efficient. This is an observation that holds in general: consistency and efficiency cannot occur together.

Model Selection in High Dimensional Models

In some applications, e.g., in radiology and biomedical imaging, the number of unknown parameters in the model is larger than the sample size, and so classical model selection procedures (e.g., AIC, BIC) fail because the parameters cannot be estimated using the method of maximum likelihood. For these so-called high-dimensional models regularized or penalized methods have been suggested in

the literature. The popular Lasso estimator, introduced by Tibshirani (1996), adds an l_1 penalty for the coefficients in the estimation process. This has as a particular advantage that it not only can shrink the coefficients towards zero, but also sets some parameters equal to zero, which corresponds to variable selection. Several extensions to the basic Lasso exist, and theoretical properties include consistency under certain conditions. The Dantzig selector (Candes and Tao 2008) is another type of method for use with high-dimensional models.

Post-model Selection Inference

Estimators that are obtained in a model that has been selected by means of a model selection procedure, are referred to as *estimators-post-selection* or *post-model-selection estimators*. Since the data are used to select the model, the selected model that one works with, is random. This is the main cause of inferences to be wrong when ignoring model selection and pretending that the selected model had been given beforehand. For example, by ignoring the fact that model selection has taken place, the estimated variance of an estimator is likely to be too small, and confidence and prediction intervals are likely to be too narrow. Literature on this topic includes Pötscher (1991), Hjort and Claeskens (2003), Shen et al. (2004), Leeb and Pötscher (2005).

Model selection can be regarded as the special case of model averaging in which the selected model takes on the weight one and all other models have weight zero. However, regarding it as such does not solve the problem because selection depends on the data, and so the weights in the estimator-post-selection are random. This results in non-normal limiting distributions of estimators-post-selection, and requires adjusted inference techniques to take the randomness of the model selection process into account. The problem of correct post-model selection inference has yet to be solved.

About the Authors

Walter Zucchini previously held the Chair of Statistics at the University of Cape Town. He is a Fellow of the Royal Statistical Society and the Royal Society of South Africa. He is Past President of the South African Statistical Association (1992) and Editor of the *South African Statistical Journal* (1986–1989). He was awarded the "Herbert Sichel Medaille" of the South African Statistical Association (2008), and the Shayle Searle Visiting Fellowship in Statistics, Victoria University, New Zealand (2008). Walter Zucchini is the co-author of the text *Model Selection* (with H. Linhart, Wiley 1986).

Gerda Claeskens is Professor at the Faculty of Business and Economics of the K.U. Leuven (Belgium). She is Elected member of the International Statistical Institute and recipient of the Noether Young Scholar Award (2004) "for outstanding achievements and contributions in nonparametric statistics." She is the author of more than 40 papers and of the book *Model selection and model averaging* (with N.L. Hjort, Cambridge University Press, 2008). Currently she is Associate editor of the *Journal of the American Statistical Association*, of *Biometrika*, and of the *Journal of Nonparametric Statistics*.

Georges Nguefack-Tsague is lecturer of Biostatistics in the Department of Public Health at the University of Yaounde I, Cameroon. He is head of the Biostatistics Unit and deputy speaker of the Master Program in Public Health. He was awarded a Lichtenberg Scholarship for his PhD studies, which he completed at the University of Goettingen (Germany). The title of his PhD thesis was *Estimating and Correcting the Effects of Model Selection Uncertainty*. He was teaching assistant (2001–2003) in the Department of Statistics and Econometrics at the University Carlos III of Madrid (Spain). Other awards included a Belgium Ministry of External Affairs (MSc) Scholarship and a Cameroon Ministry of Economy and Finance (MA) Scholarship.

Cross References

►Akaike's Information Criterion
►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bootstrap Methods
►C_p Statistic
►Exponential and Holt-Winters Smoothing
►Kullback-Leibler Divergence
►Marginal Probability: Its Use in Bayesian Statistics as Model Evidence
►Markov Chain Monte Carlo
►Sensitivity Analysis
►Statistical Evidence
►Structural Time Series Models
►Time Series

References and Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov B, Csáki F (eds) Second international symposium on information theory, Akadémiai Kiadó, Budapest, pp 267–281

Burnham PK, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York

Candes E, Tao T (2008) The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat 35:2313–2351
Claeskens G, Hjort NL (2003) The focussed information criterion (with discussion). J Am Stat Assoc 98:900–916
Claeskens G, Hjort NL (2008) Model selection and model averaging. Cambridge University Press, Cambridge
Grünwald P (2007) The minimum description length principle. MIT Press, Boston
Hjort NL, Claeskens G (2003) Frequentist model average estimators (with discussion). J Am Stat Assoc 98:879–899
Leeb H, Pötscher BM (2005) Model selection and inference: fact and fiction. Econometr Theor 21:21–59
Linhart H, Zucchini W (1986) Model selection. Wiley, New York
McQuarrie ADR, Tsai CL (1998) Regression and time series model selection. World Scientific, River Edge
Miller AJ (2002) Subset selection in regression, 2nd edn. Chapman and Hall/CRC, Boca Raton
Pötscher BM (1991) Effects of model selection on inference. Econometr Theor 7:163–185
Rissanen JJ (1996) Fisher information and stochastic complexity. IEEE Trans Inform Theory 42:40–47
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Shen X, Huang HC, Ye J (2004) Inference after model selection. J Am Stat Assoc 99:751–762
Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). J Roy Stat Soc B 64:583–639
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc B 58(1):267–288
Wasserman L (2000) Bayesian model selection and model averaging. J Math Psychol 44:92–107
Zucchini W (2000) An introduction to model selection. J Math Psychol 44:41–61

Model-Based Geostatistics

Hannes Kazianka[1], Jürgen Pilz[2]
[1]University of Technology, Vienna, Austria
[2]Professor, Head
University of Klagenfurt, Klagenfurt, Austria

Stochastic Models for Spatial Data

Diggle and Ribeiro (2007) and Mase (2010) describe geostatistics as a branch of spatial statistics that deals with statistical methods for the analysis of spatially referenced data with the following properties. Firstly, values Y_i, $i = 1, \ldots, n$, are observed at a discrete set of sampling locations $\boldsymbol{x}_i$ within some spatial region $\mathcal{S} \subset \mathbb{R}^d$, $d \geq 2$. Secondly, each observed value Y_i is either a measurement of, or is statistically related to, the value of an underlying continuous spatial phenomenon, $Z(\boldsymbol{x})$, at the corresponding sampling location $\boldsymbol{x}_i$. The term model-based geostatistics refers to geostatistical methods that rely on a stochastic model. The observed phenomenon is viewed as a realization of a continuous stochastic process in space, a so-called random field.

Such a random field $Z(\boldsymbol{x})$ is fully determined by specifying all multivariate distributions, i.e., $P(Z(\boldsymbol{x}_1) \leq z_1, \ldots, Z(\boldsymbol{x}_n) \leq z_n)$ for arbitrary $n \in \mathbb{N}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{S}$. Since a full characterization of a random field is usually hopeless, the mean function $m(\boldsymbol{x}) = E(Z(\boldsymbol{x}))$ and the covariance function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = Cov(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))$ play a prominent role. Thereby, $m(\boldsymbol{x})$ represents the trend while $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ defines the dependence structure of the random field. It is typical that the assumption of weak (second-order) isotropy is made about the random field, i.e., its mean function is constant and its covariance function $K(\boldsymbol{x}_1, \boldsymbol{x}_2)$ depends on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ only through $h = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$, where $\|.\|_2$ denotes the Euclidean distance. In this case K is called an isotropic autocovariance function. The covariance function is directly related to smoothness properties of the random field such as mean square continuity and differentiability. A widely used parametric family of isotropic autocovariance functions is the Matern family

$$K_{\sigma^2,\boldsymbol{\theta}}(h) = \sigma^2 \left((1 - \vartheta_2) + \frac{\vartheta_2}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{2\kappa^{\frac{1}{2}} h}{\vartheta_1} \right)^{\kappa} \mathcal{K}_{\kappa} \left(\frac{2\kappa^{\frac{1}{2}} h}{\vartheta_1} \right) \right),$$

where $\mathcal{K}_{\kappa}$ denotes the modified Bessel function of order $\kappa > 0$, $\vartheta_1 > 0$ is a called the "range parameter" controlling how fast the covariance decays as the distance h gets large, $\vartheta_2 \in [0,1]$ is called the "nugget parameter" and describes a measurement error, σ^2 controls the variance and $\boldsymbol{\theta} = (\vartheta_1, \vartheta_2, \kappa)$ denotes the vector of correlation parameters. The parameter κ controls the smoothness of the corresponding process. A thorough mathematical introduction to the theory of random fields is given in Stein (1999) and Yaglom (1987).

The most important geostatistical model is the linear Gaussian model

$$Y_i = \boldsymbol{f}(\boldsymbol{x}_k)^T \boldsymbol{\beta} + Z(\boldsymbol{x}_i), \quad i = 1, \ldots, n, \tag{1}$$

where $Z(\boldsymbol{x})$ is a weakly isotropic zero-mean Gaussian random field with autocovariance function $K_{\sigma^2,\boldsymbol{\theta}}$, $\boldsymbol{f}$ is a vector of location-dependent explanatory variables and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression parameters. The

likelihood function for the linear Gaussian model is

$$p\left(\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\right) = (2\pi)^{-\frac{n}{2}} \left|\sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{Y} - \boldsymbol{F}\boldsymbol{\beta}\right)\right\},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ denotes the correlation matrix, $\boldsymbol{F}$ is the design matrix and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is the vector of observations. The maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2 in the linear Gaussian model are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{F}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{Y}, \tag{2}$$

$$\hat{\sigma}^2 = \frac{1}{n}\left(\mathbf{Z} - \mathbf{F}\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(\mathbf{Z} - \mathbf{F}\hat{\boldsymbol{\beta}}\right). \tag{3}$$

Plugging these estimates into the log-likelihood, we arrive at the so-called profiled log-likelihood, which just contains the parameters $\boldsymbol{\theta}$

$$\log p\left(\boldsymbol{Y} \mid \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{\theta}\right) = -\frac{n}{2}\left(\log(2\pi) + 1\right) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right| - \frac{n}{2}\log\left(\hat{\sigma}^2\right).$$

To obtain $\hat{\boldsymbol{\theta}}$ we have to maximize the latter equation for $\boldsymbol{\theta}$ numerically. Note that this maximization problem is a lot simpler than the maximization of the complete likelihood where $\boldsymbol{\beta}$ and σ^2 are additional unknowns, especially when p is large. Spatial prediction, which is often the goal in geostatistics, is performed based on the estimated parameters. The plug-in predictive distribution for the value of the random field at an unobserved location $\boldsymbol{x}_0$ is Gaussian

$$Y_0 \mid \boldsymbol{Y}, \hat{\sigma}^2, \hat{\boldsymbol{\theta}} \sim \mathrm{N}\left(\boldsymbol{k}^T \boldsymbol{K}^{-1} \boldsymbol{Y} + \boldsymbol{s}^T \hat{\boldsymbol{\beta}}, \hat{\sigma}^2 - \boldsymbol{k}^T \boldsymbol{K}^{-1} \boldsymbol{k} + \hat{\sigma}^2 \boldsymbol{s}^T \left(\boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{F}\right)^{-1} \boldsymbol{s}\right), \tag{4}$$

where $\boldsymbol{K} = \hat{\sigma}^2 \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}, \boldsymbol{s} = \boldsymbol{f}(\boldsymbol{x}_0) - \boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{k}, \boldsymbol{k} = Cov(\boldsymbol{Z}, Z(\boldsymbol{x}_0))$, $\boldsymbol{Z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))^T$.

Weak isotropy is a rather strong assumption and environmental processes are typically not direction independent but show an anisotropic behavior. A popular extension to isotropic random fields is to consider random fields that become isotropic after a linear transformation of the coordinates (Schabenberger and Gotway 2005). This special variant of anisotropy is called geometric anisotropy. Let $Z_1(\boldsymbol{x})$ be an isotropic random field on $\mathbb{R}^d$ with autocovariance function K_1 and mean μ. For the random field $Z(\boldsymbol{x}) = Z_1(\boldsymbol{T}\boldsymbol{x})$, where $\boldsymbol{T} \in \mathbb{R}^{d \times d}$, we get that $E(Z(\boldsymbol{x})) = \mu$ and the corresponding autocovariance function is $Cov(Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2)) = K_1\left(\|\boldsymbol{T}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\|_2\right)$. When correcting for geometric anisotropy we need to revert the coordinate transformation. $Z\left(\boldsymbol{T}^{-1}\boldsymbol{x}\right)$ has the same mean as $Z(\boldsymbol{x})$ but isotropic autocovariance function K_1. When correcting for stretching and rotation of the coordinates we have

$$\boldsymbol{T}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}.$$

Here, λ and φ are called the anisotropy ratio and anisotropy angle, respectively. All the models that we consider in this chapter can be extended to account for geometric anisotropy by introducing these two parameters.

Bayesian Kriging

The first steps towards Bayesian modeling and prediction in geostatistics were made by Kitanidis (1986) and Omre (1987) who developed a Bayesian version of universal kriging. One of the advantages of the Bayesian approach, besides its ability to deal with the uncertainty about the model parameters, is the possibility to work with only a few measurements. Assume a Gaussian random field model in the form of the form Eq. 1 with known covariance matrix $\boldsymbol{K}$ but unknown parameter vector $\boldsymbol{\beta}$. From Bayesian analysis we know that it is natural to assume a prior of the form $\boldsymbol{\beta} \sim \mathrm{N}\left(\boldsymbol{m}_b, \sigma^2 \boldsymbol{V}_b\right)$ for $\boldsymbol{\beta}$, where $\boldsymbol{V}_b$ is a positive semidefinite matrix. It can be shown that the posterior distribution for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \mid \boldsymbol{Z} \sim \mathrm{N}\left(\tilde{\boldsymbol{\beta}}, \sigma^2 \boldsymbol{V}_{\tilde{\beta}}\right),$$

where $\tilde{\boldsymbol{\beta}} = \boldsymbol{V}_{\tilde{\beta}}\left(\sigma^2 \boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{Z} + \boldsymbol{V}_b^{-1} \boldsymbol{m}_b\right)$ and $\boldsymbol{V}_{\tilde{\beta}} = \left(\sigma^2 \boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{F} + \boldsymbol{V}_b^{-1}\right)^{-1}$. The predictive distribution of $Z(\boldsymbol{x}_0)$ is also Gaussian and given by

$$Z(\boldsymbol{x}_0) \mid \boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{k}^T \boldsymbol{K}^{-1} \boldsymbol{Z} + \boldsymbol{s}^T \tilde{\boldsymbol{\beta}}, \sigma^2 - \boldsymbol{k}^T \boldsymbol{K}^{-1} \boldsymbol{k} + \sigma^2 \boldsymbol{s}^T \boldsymbol{V}_{\tilde{\beta}} \boldsymbol{s}\right),$$

where $\boldsymbol{F}$, $\boldsymbol{s}$ and $\boldsymbol{k}$ are defined as in Section "▶Stochastic Models for Spatial Data". From the above representation of the Bayesian kriging predictor it becomes clear that Bayesian kriging bridges the gap between simple and universal kriging. We get simple kriging in case of complete knowledge of the trend, which corresponds to $\mathbf{V}_b = \mathbf{0}$, whereas we get the universal kriging predictor if we have no knowledge of $\boldsymbol{\beta}$ ($\mathbf{V}_b^{-1} = \mathbf{0}$ in the sense that the smallest eigenvalue of $\mathbf{V}_b$ converges to infinity). Interestingly, the Bayesian universal kriging predictor has a smaller or equal variance than the classical universal kriging predictor (see Eq. 4) since $\left(\boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{F} + \sigma^{-2} \boldsymbol{V}_b^{-1}\right)^{-1} \preceq \left(\boldsymbol{F}^T \boldsymbol{K}^{-1} \boldsymbol{F}\right)^{-1}$, where $\preceq$ denotes the Loewner partial ordering.

Bayesian universal kriging is not fully Bayesian because $\boldsymbol{K}$ is assumed known. Diggle and Ribeiro (2007) summarize the results for a fully Bayesian analysis of Gaussian random field models of the form Eq. 1, where $K_{\sigma^2,\theta} = \sigma^2 \Sigma_{\vartheta_1}$ and ϑ_1 is the range parameter of an isotropic autocorrelation function model.

Transformed Gaussian Kriging

Probably the most simple way to extend the Gaussian random field model is to assume that a differentiable transformation of the original random field, $Z_1(\boldsymbol{x}) = g(Z(\boldsymbol{x}))$, is Gaussian. The mean of the transformed field is unknown and parameterized by $\boldsymbol{\beta}$, $E(Z_1(\boldsymbol{x})) = \boldsymbol{f}(\boldsymbol{x})^T \boldsymbol{\beta}$. If we assume that the transformation function g and the covariance function K of $Y(\boldsymbol{x})$ are known, the optimal predictor for $Z(\boldsymbol{x}_0)$ can be derived using the results from Section "►Stochastic Models for Spatial Data". However, in practice neither K nor g is known and we have to estimate them from the data.

A family of one-parameter transformation functions g_λ that is widely used in statistics is the so-called Box-Cox family

$$g_\lambda(z) = \begin{cases} \frac{z^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(z), & \lambda = 0. \end{cases}$$

The ►Box-Cox transformation is valid for positive-valued random fields and is able to model moderately skewed, unimodal data.

The likelihood of the data $\boldsymbol{Y}$ in the transformed Gaussian model can be written as

$$p(\mathbf{Y} \mid \boldsymbol{\Theta}) = J_\lambda(\mathbf{Y}) (2\pi)^{-\frac{n}{2}} \left|\sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} (g_\lambda(\mathbf{Y}) - \boldsymbol{F}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (g_\lambda(\mathbf{Y}) - \boldsymbol{F}\boldsymbol{\beta})\right],$$

where, $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \lambda)$, $J_\lambda(\mathbf{Y})$ is the determinant of the Jacobian of the transformation, $g_\lambda(\boldsymbol{Y}) = (g_\lambda(Y_1), \ldots, g_\lambda(Y_n))$ and λ is the transformation parameter. De Oliveira et al. (1997) point out that the interpretation of $\boldsymbol{\beta}$ changes with the value of λ, and the same is true for the covariance parameters σ^2 and $\boldsymbol{\theta}$, to a lesser extent though. To estimate the parameters λ and $\boldsymbol{\theta}$, we make use of the profile likelihood approach that we have already encountered in Section "►Stochastic Models for Spatial Data". For fixed values of λ and $\boldsymbol{\theta}$, the maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2 are given by Eqs. 2 and 3 with $\boldsymbol{Y}$ replaced by $g_\lambda(\boldsymbol{Y})$. Again, the estimates for λ and $\boldsymbol{\theta}$ cannot be written in closed form and must be found numerically by plugging $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ in the likelihood for numerical maximization.

The estimated parameters $\hat{\boldsymbol{\Theta}}$ are subsequently used for spatial prediction. To perform a plug-in prediction we make use of the conditional distribution of the Gaussian variable $Y_0 \mid \boldsymbol{Y}, \hat{\boldsymbol{\Theta}}$ and back-transform it to the original scale by g_λ^{-1}. A Bayesian approach to spatial prediction in the transformed Gaussian model is proposed in De Oliveira et al. (1997).

The copula-based geostatistical model (Kazianka and Pilz 2009) also works with transformations of the marginal distributions of the random field and is a generalization of transformed Gaussian kriging. In this approach all multivariate distributions of the random field are described by a copula (Sempi 2010) and a family of univariate marginal distributions. Due to the additional flexibility introduced by the choice of the copula and of the marginal distribution, these models are able to deal with extreme observations and multi-modal data.

Generalized Linear Geostatistical Models

►Generalized linear models (McCullagh and Nelder 1989) provide a unifying framework for regression modeling of both continuous and discrete data. Diggle and Ribeiro (2007) extend the classical generalized linear model to what they call the generalized linear geostatistical model (GLGM). The responses Y_i, $i = 1, \ldots, n$, corresponding to location $\boldsymbol{x}_i$ are assumed to follow a family of univariate distributions indexed by their expectation, μ_i, and to be conditionally independent given $\boldsymbol{Z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))$. The μ_i are specified through

$$h(\mu_i) = \boldsymbol{f}(\boldsymbol{x}_i)^T \boldsymbol{\beta} + Z(\boldsymbol{x}_i),$$

where $Z(\boldsymbol{x})$ is a Gaussian random field with autocovariance function $K_{\boldsymbol{\theta}}$ and h is a pre-defined link function. The two most frequently applied GLGMs are the Poisson log-linear model, where Y_i is assumed to follow a Poisson distribution and the link function is the logarithm, and the binomial logistic-linear model, where Y_i is assumed to follow a Bernoulli distribution with probability $\mu_i = p(\boldsymbol{x}_i)$ and $h(\mu_i) = \log(p(\boldsymbol{x}_i) / (1 - p(\boldsymbol{x}_i)))$. These models are suitable for representing spatially referenced count data and binary data, respectively.

Since maximum likelihood estimation of the parameters is difficult, a Markov chain Monte Carlo (Robert and Casella 2004) approach (see ►Markov Chain Monte Carlo) is proposed to sample from the posteriors of the model parameters as well as from the predictive distributions at unobserved locations $\boldsymbol{x}_0$. The algorithm proceeds by sampling from $\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{\theta}$, from $\boldsymbol{\theta} \mid \boldsymbol{Z}$ and from $\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{Y}$ with the help of Metropolis-Hastings updates. At iteration $t+1$ and

actual sample $\left(\boldsymbol{Z}^t, \boldsymbol{\theta}^t, \boldsymbol{\beta}^t, Z^t(\boldsymbol{x}_0)\right)$, perform the following steps:

- Update $\boldsymbol{Z}$. For $i = 1, \ldots, n$, sample a new proposal $Z'(\boldsymbol{x}_i)$ from the conditional Gaussian distribution $p\left(Z(\boldsymbol{x}_i) \mid \boldsymbol{\theta}^t, \boldsymbol{Z}^t_{-i}\right)$, where $\boldsymbol{Z}^t_{-i}$ denotes $\boldsymbol{Z}^t = (Z^t(\boldsymbol{x}_1), \ldots, Z^t(\boldsymbol{x}_n))$ with its ith element removed. Accept $Z'(\boldsymbol{x}_i)$ with probability $r = \min\left\{1, \frac{p\left(Y_i \mid \boldsymbol{\beta}^t, Z'(\boldsymbol{x}_i)\right)}{p\left(Y_i \mid \boldsymbol{\beta}^t, Z^t(\boldsymbol{x}_i)\right)}\right\}$.
- Update $\boldsymbol{\theta}$. Sample a new proposal $\boldsymbol{\theta}'$ from a proposal distribution $J\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^t\right)$. Accept the new proposal with probability $r = \min\left\{1, \frac{p\left(\boldsymbol{Z}^{t+1} \mid \boldsymbol{\theta}'\right) J\left(\boldsymbol{\theta}^t \mid \boldsymbol{\theta}'\right)}{p\left(\boldsymbol{Z}^{t+1} \mid \boldsymbol{\theta}^t\right) J\left(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^t\right)}\right\}$.
- Update $\boldsymbol{\beta}$. Sample a new proposal $\boldsymbol{\beta}'$ from a proposal distribution $J\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}^t\right)$. Accept the new proposal with probability $r = \min\left\{1, \frac{\prod_{i=1}^n p\left(Y_i \mid Z^{t+1}(\boldsymbol{x}_i), \boldsymbol{\beta}'\right) J\left(\boldsymbol{\beta}^t \mid \boldsymbol{\beta}'\right)}{\prod_{i=1}^n p\left(Y_i \mid Z^{t+1}(\boldsymbol{x}_i), \boldsymbol{\beta}^t\right) J\left(\boldsymbol{\beta}' \mid \boldsymbol{\beta}^t\right)}\right\}$
- Draw a sample $Z^{t+1}(\boldsymbol{x}_0)$ from the conditional Gaussian distribution $Z(\boldsymbol{x}_0) \mid \boldsymbol{Z}^{t+1}, \boldsymbol{\theta}^{t+1}$.

If point predictions for $Z(\boldsymbol{x}_0)$ are needed, the Monte Carlo approximation to the expected value of $Z(\boldsymbol{x}_0) \mid \boldsymbol{Y}$ can be used, i.e., $E\left(\boldsymbol{Z}(\boldsymbol{x}_0) \mid \boldsymbol{Y}\right) \approx \frac{1}{M} \sum_{t=1}^M Z^t(\boldsymbol{x}_0)$, where M is the number of simulations.

About the Author

For the biography *see* the entry ►Statistical Design of Experiments

Cross References

►Analysis of Areal and Spatial Interaction Data
►Box–Cox Transformation
►Gaussian Processes
►Generalized Linear Models
►Geostatistics and Kriging Predictors
►Markov Chain Monte Carlo
►Random Field
►Spatial Statistics

References and Further Reading

De Oliveira V, Kedem B, Short D (1997) Bayesian prediction of transformed Gaussian fields. J Am Stat Assoc 92:1422–1433
Diggle P, Ribeiro P (2007) Model-based geostatistics. Springer, New York
Sempi C (2010) Copulas. (this volume)
Kazianka H, Pilz J (2009) Copula-based geostatistical modeling of continuous and discrete data including covariates. Stoch Env Res Risk Assess, doi: 10.1007/s00477-009-0353-8
Kitanidis P (1986) Parameter uncertainty in estimation of spatial function: Bayesian analysis. Water Resour Res 22: 499–507
Mase S (2010) Geostatistics and kriging predictors. (this volume)
McCullagh P, Nelder J (1989) Generalized linear models. Chapman & Hall/CRC, Boca Raton
Omre H (1987) Bayesian kriging – merging observations and qualified guesses in kriging. Math Geol 19:25–39
Robert C, Casella G (2004) Monte Carlo statistical methods. Springer, New York
Schabenberger O, Gotway C (2005) Statistical methods for spatial data analysis. Chapman & Hall/CRC, Boca Raton
Stein M (1999) Interpolation of spatial data. Springer, New York
Yaglom A (1987) Correlation theory of stationary and related random functions. Springer, New York

Modeling Count Data

Joseph M. Hilbe
Emeritus Professor
University of Hawaii, Honolulu, HI, USA
Adjunct Professor of Statistics
Arizona State University, Tempe, AZ, USA
Solar System Ambassador
California Institute of Technology, Pasadena, CA, USA

Count models are a subset of discrete response regression models. Count data are distributed as non-negative integers, are intrinsically heteroskedastic, right skewed, and have a variance that increases with the mean. Example count data include such situations as length of hospital stay, the number of a certain species of fish per defined area in the ocean, the number of lights displayed by fireflies over specified time periods, or the classic case of the number of deaths among Prussian soldiers resulting from being kicked by a horse during the Crimean War.

►Poisson regression is the basic model from which a variety of count models are based. It is derived from the Poisson probability mass function, which can be expressed as

$$f(y_i; \lambda_i) = \frac{e^{-t_i \lambda_i} (t_i \lambda_i)^{y_i}}{y_i!}, \qquad y = 0, 1, 2, \ldots; \ \mu > 0 \quad (1)$$

with y_i as the count response, λ_i as the predicted count or rate parameter, and t_i the area or time in which counts enter the model. When λ_i is understood as applying to individual counts without consideration of size or time, $t_i = 1$. When $t_i > 1$, it is commonly referred to as an exposure, and is modeled as an offset.

Estimation of the Poisson model is based on the log-likelihood parameterization of the Poisson probability distribution, which is aimed at determining parameter values

making the data most likely. In exponential family form it is given as:

$$L(\mu_i; y_i) = \sum_{i=1}^{n} \{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\}, \quad (2)$$

where μ_i is typically used to symbolize the predicted counts in place of λ_i. Equation 2, or the deviance function based on it, is used when the Poisson model is estimated as a generalized linear model (*GLM*) (*see* ►Generalized Linear Models). When estimation employs a full maximum likelihood algorithm, μ_i is expressed in terms of the linear predictor, $x'\beta$. As such it appears as

$$\mu_i = \exp(x_i\beta). \quad (3)$$

In this form, the Poisson log-likelihood function is expressed as

$$L(\beta; y_i) = \sum_{i=1}^{n} \{y_i(x_i\beta) - \exp(x_i\beta) - \ln(y_i!)\}. \quad (4)$$

A key feature of the Poisson model is the equality of the mean and variance functions. When the variance of a Poisson model exceeds its mean, the model is termed overdispersed. Simulation studies have demonstrated that overdispersion is indicated when the Pearson χ^2 dispersion is greater than 1.0 (Hilbe 2007). The dispersion statistic is defined as the Pearson χ^2 divided by the model residual degrees of freedom. Overdispersion, common to most Poisson models, biases the parameter estimates and fitted values. When Poisson overdispersion is real, and not merely apparent (Hilbe 2007), a count model other than Poisson is required.

Several methods have been used to accommodate Poisson overdispersion. Two common methods are quasi-Poisson and negative binomial regression. Quasi-Poisson models have generally been understood in two distinct manners. The traditional manner has the Poisson variance being multiplied by a constant term. The second, employed in the glm() function that is downloaded by default when installing R software, is to multiply the standard errors by the square root of the Pearson dispersion statistic. This method of adjustment to the variance has traditionally been referred to as scaling. Using R's quasipoisson() function is the same as what is known in standard GLM terminology as the scaling of standard errors.

The traditional negative binomial model is a Poisson-gamma mixture model with a second ancillary or heterogeneity parameter, α. The mixture nature of the variance is reflected in its form, $\mu_i + \alpha\mu_i^2$, or $\mu_i(1 + \alpha\mu_i)$. The Poisson variance is μ_i, and the two parameter gamma variance is μ_i^2/ν. ν is inverted so that $\alpha = 1/\nu$, which allows for a direct relationship between μ_i, and ν. As a Poisson-gamma mixture model, counts are Poisson distributed as they enter into the model. α is the shape (gamma) of the manner counts enter into the model as well as a measure of the amount of Poisson overdispersion in the data.

The negative binomial probability mass function (*see* ►Geometric and Negative Binomial Distributions) may be formulated as

$$f(y_i; \mu_i, \alpha) = \binom{y_i + 1/\alpha - 1}{1/\alpha - 1} (1/(1 + \alpha\mu_i))^{1/\alpha} (\alpha\mu_i/(1 + \alpha\mu_i))_i^y, \quad (5)$$

with a log-likelihood function specified as

$$L(\mu_i; y_i, \alpha) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\tfrac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \left(\tfrac{1}{\alpha}\right) \ln(1 + \alpha\mu_i) + \ln\Gamma\left(y_i + \tfrac{1}{\alpha}\right) - \ln\Gamma(y_i + 1) - \ln\Gamma\left(\tfrac{1}{\alpha}\right) \right\}. \quad (6)$$

In terms of $\mu = \exp(x'\beta)$, the parameterization employed for maximum likelihood estimation, the negative binomial log-likelihood appears as

$$L(\beta; y_i, \alpha) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{\alpha \exp(x_i'\beta)}{1 + \alpha \exp(x_i'\beta)}\right) - \left(\frac{1}{\alpha}\right) \ln\left(1 + \alpha \exp(x_i'\beta)\right) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i + 1) - \ln\Gamma\left(\frac{1}{\alpha}\right) \right\}. \quad (7)$$

This form of negative binomial has been termed *NB2*, due to the quadratic nature of its variance function. It should be noted that the *NB2* model reduces to the Poisson when $\alpha = 0$. When $\alpha = 1$, the model is geometric, taking the shape of the discrete correlate of the continuous negative exponential distribution. Several fit tests exist that evaluate whether data should be modeled as Poisson or *NB2* based on the degree to which α differs from 0.

When exponentiated, Poisson and *NB2* parameter estimates may be interpreted as incidence rate ratios. For example, given a random sample of 1,000 patient observations from the German Health Survey for the year 1984, the following Poisson model output explains the years expected number of doctor visits on the basis of gender and marital status, both recorded as binary (1/0) variables, and the continuous predictor, age.

Docvis	IRR	OIM std. err.	z	P > |z|	[95% Conf. interval]	
Female	1.516855	0.054906	11.51	0.000	1.41297	1.628378
Married	0.8418408	0.0341971	−4.24	0.000	0.7774145	0.9116063
Age	1.018807	0.0016104	11.79	0.000	1.015656	1.021968

The estimates may be interpreted as

- Females are expected to visit the doctor some 50% more times during the year than males, holding marital status and age constant.

  Married patients are expected to visit the doctor some 16% fewer times during the year than unmarried patients, holding gender and age constant.

  For a one year increase in age, the rate of visits to the doctor increases by some 2%, with marital status and gender held constant.

It is important to understand that the canonical form of the negative binomial, when considered as a *GLM*, is not *NB2*. Nor is the canonical negative binomial model, *NB-C*, appropriate to evaluate the amount of Poisson overdispersion in a data situation. The NB-C parameterization of the negative binomial is directly derived from the negative binomial log-likelihood as expressed in Eq. 6. As such, the link function is calculated as $\ln(\alpha\mu/(1+\alpha\mu))$. The inverse link function, or mean, expressed in terms of $x'\beta$, is $1/(\alpha(\exp(-x'\beta)-1))$.

When estimated as a *GLM*, *NB-C* can be amended to *NB2* form by substituting $\ln(\mu)$ and $\exp(x'\beta)$ respectively for the two above expressions. Additional amendments need to be made to have the *GLM*-estimated *NB2* display the same parameter standard errors as are calculated using full maximum likelihood estimation. The *NB-C* log-likelihood, expressed in terms of μ, is identical to that of the *NB2* function. However, when parameterized as $x'\beta$, the two differ, with the *NB-C* appearing as

$$L(\beta;y_i,\alpha)=\sum_{i=1}^{n}\{y_i(x_i\beta)+(1/\alpha)\ln(1-\exp(x_i\beta)) + \ln\Gamma(y_i+1/\alpha)-\ln\Gamma(y_i+1)-\ln\Gamma(1/\alpha)\} \quad (8)$$

The *NB-C* model better fits certain types of count data than *NB2*, or any other variety of count model. However, since its fitted values are not on the log scale, comparisons cannot be made to Poisson or *NB2*.

The *NB2* model, in a similar manner to the Poisson, can also be overdispersed if the model variance exceeds its nominal variance. In such a case one must attempt to determine the source of the extra correlation and model it accordingly.

The extra correlation that can exist in count data, but which cannot be accommodated by simple adjustments to the Poisson and negative binomial algorithms, has stimulated the creation of a number of enhancements to the two base count models. The differences in these enhanced models relates to the attempt of identifying the various sources of overdispersion.

For instance, both the Poisson and negative binomial models assume that there exists the possibility of having zero counts. If a given set of count data excludes that possibility, the resultant Poisson or negative binomial model will likely be overdispersed. Modifying the loglikelihood function of these two models in order to adjust for the non-zero distribution of counts will eliminate the overdispersion, if there are no other sources of extra correlation. Such models are called, respectively, zero-truncated Poisson and zero-truncated negative binomial models.

Likewise, if the data consists of far more zero counts that allowed by the distributional assumptions of the Poisson or negative binomial models, a zero-inflated set of models may need to be designed. Zero-inflated models are ▶mixture models, with one part consisting of a 1/0 binary response model, usually a ▶logistic regression, where the probability of a zero count is estimated in difference to a non-zero-count. A second component is generally comprised of a Poisson or negative binomial model that estimates the full range of count data, adjusting for the overlap in estimated zero counts. The point is to (1) determine the estimates that account for zero counts, and (2) to estimate the adjusted count model data.

Hurdle models are another type mixture model designed for excessive zero counts. However, unlike the zero-inflated models, the hurdle-binary model estimates the probability of being a non-zero count in comparison to a zero count; the hurdle-count component is estimated on the basis of a zero-truncated count model. Zero-truncated, zero-inflated, and hurdle models all address abnormal

Modeling Count Data. Table 1 Models to adjust for violations of Poisson/NB distributional assumptions

Response	Example models
1: no zeros	Zero-truncated models (*ZTP*; *ZTNB*)
2: excessive zeros	Zero-inflated (*ZIP*; *ZINB*; *ZAP*; *ZANB*); hurdle models
3: truncated	Truncated count models
4: censored	Econometric and survival censored count models
5: panel	*GEE*; fixed, random, and mixed effects count models
6: separable	Sample selection, finite mixture models
7: two-responses	Bivariate count models
8: other	Quantile, exact, and Bayesian count models

Modeling Count Data. Table 2 Methods to directly adjust the variance (from Hilbe 2007)

Variance function	Example models
0: μ	Poisson
1: $\mu(\Phi)$	Quasi-Poisson; scaled SE; robust SE
2: $\mu(1+\alpha)$	Linear *NB* (*NB1*)
3: $\mu(1+\mu)$	Geometric
4: $\mu(1+\alpha\mu)$	Standard *NB* (*NB2*); quadratic *NB*
5: $\mu(1+(\alpha v)\mu)$	Heterogeneous *NB* (*NH-H*)
6: $\mu(1+\alpha\mu^{\rho})$	Generalized *NB* (*NB-P*)
7: $V[R]V'$	Generalized estimating equations

zero-count situations, which violate essential Poisson and negative binomial assumptions.

Other violations of the distributional assumptions of Poisson and negative binomial probability distributions exist. Table 1 below summarizes major types of violations that have resulted in the creation of specialized count models.

Alternative count models have also been constructed based on an adjustment to the Poisson variance function, μ. We have previously addressed two of these. Table 2 provides a summary of major types of adjustments.

Three texts specifically devoted to describing the theory and variety of count models are regarded as the standard resources on the subject. Other texts dealing with discrete response models in general, as well as texts on generalized linear models (*see* Generalized Linear Models), also have descriptions of many of the models mentioned in this article.

About the Author

For biography *see* the entry ►Logistic Regression.

Cross References

►Dispersion Models
►Generalized Linear Models
►Geometric and Negative Binomial Distributions
►Poisson Distribution and Its Application in Statistics
►Poisson Regression
►Robust Regression Estimation in Generalized Linear Models
►Statistical Methods in Epidemiology

References and Further Reading

Cameron AC, Trivedi PK (1998) Regression analysis of count data. Cambridge University Press, New York
Hilbe JM (2007) Negative binomial regression. Cambridge University Press, Cambridge, UK
Hilbe JM (2011) Negative binomial regression, 2nd edn. Cambridge University Press, Cambridge, UK
Winkelmann R (2003) Econometric analysis of count data, 4th edn. Springer, Heidelberg

M

Modeling Randomness Using System Dynamics Concepts

Mahender Singh[1], Frank M. Guess[2], Timothy M. Young[2], Lefei Liu[3]
[1]Research Director of Supply Chain 2020
Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Professor
University of Tennessee, Knoxville, TN, USA
[3]University of South Carolina, Columbia, SC, USA

L. J. Savage (1980) and others understood the importance of better computational tools for utilizing Bayesian insights data in real life applications long ago. Such computational tools and software are now available that use subjective (or soft) data as well as quantitative (or hard) data. But

despite the availability of new tools and buildup of massive databases, the increased complexity and integration of economic and other systems involving people poses a significant challenge to a solely statistical driven view of the system. More importantly, evidence suggests that relying solely on standard statistical models is inadequate to represent real life systems effectively for management insights and decisions.

Unpredictability characterizes most real life systems due to non-linear relationships and multiple time-delayed feedback loops between interconnected elements. Senge (1990) describes it as *dynamic complexity* – *"situations* where the cause and effect are subtle, and the effects over time of interventions are not obvious." As a result, such systems are unsuitable for quantitative "only" representations without some subjective expert views. System Dynamics models offer a helpful alternative to modeling randomness that is based on hard data and soft data that models a real world system; see for example Sterman (2000) and his references.

According to , Forrester (1980) three types of data are required to develop the foundation of an effective model: numerical, written and mental data; compare, also, Sterman (2000) discussion on these points. In most cases, however, only a small fraction of the data needed to model a real world system may be available in the form of numerical data. Perhaps, the most important data to build a model, namely the mental data, is difficult to represent only numerically. But due to heavy influence of quantitative bias in model development, some modelers disregard key qualitative information in favor of information that can be estimated statistically. Sterman (2000) considers this reasoning counterintuitive and counterproductive in practice with realistic systems. He states that "omitting structures and variables known to be important because numerical data are unavailable is actually less scientific and less accurate than using your best judgment to estimate their values." This is in line with Forrester's views (1961) asserting that, "to omit such variables is equivalent to saying they have zero effect - probably the only value that is known to be wrong!" A suitable approach in such cases is to iteratively improve the accuracy and reliability of data by leveraging deeper insights into the system and interaction between various variables over time, along with sensitivity analysis of various contingencies.

A key to understanding a dynamic real world system is to identify and study the causal loops (or sub-systems) of the system. An analysis of the structure-behavior relationship in a model can uncover causal loops that are primarily responsible for the observed behavior of the model, i.e., identify the "dominant" loop. The dominant loop is the most influential structure in determining the overall behavior of a system depending on the specific conditions of a system. It is possible for any loop to be the dominant loop at a point in time but then as conditions change the same loop can be displaced by another loop as the dominant loop in a different time frame. Due to the shifting dominance of the loops in determining system performance over time, it is necessary that a system is explored to isolate the interactions between the variables that form various causal loops. Clearly, collecting such information is challenging on many fronts. First, the sheer volume of data required to map a real world system is a challenge; secondly, this kind of information is often qualitative in nature (mental, experiential or judgment) and hence not easy to capture; and thirdly, the information keeps changing over time.

Viewing system performance as a series of connected dominant loop behaviors is a fundamentally different way to study a system. In effect, this point of view suggests that it may not be possible or necessary to find the "one best" single representation to describe the system's performance over time. Instead, we can now treat the system as a composite structure that may be formed by the amalgamation of a number of different sub representations that collectively describe the system performance. This perspective alleviates the unnecessary difficulty that is imposed on a single representation to capture the logic of possibly disconnected patterns. Indeed, this approach has its own challenges in terms of how to superimpose the various patterns to model reality.

Note both Bayesian and System Dynamics have very helpful roles to play in the analysis of real life systems that do not yield easily to purely hard data or classical models. Accordingly, one can consider an integrated approach where a Bayesian model provides specific input to a System Dynamics model to complement the capabilities of the two approaches. A System Dynamics model enhanced by Bayesian inference will allow modelers to iteratively incorporate various data types into a comprehensive model and study the behavior of a system over time. This approach allows for the inclusion of both hard data and soft data into the model. Since the modeling process is iterative, the subjective views can be augmented or replaced with hard data as such information is acquired and improved over time. When appropriate data are available, it can be used as input to the System Dynamics model of various contingencies, such as "fear" curves, "hope" curves, or mixtures of them from a Bayesian perspective. When such data are not available, varied contingencies can still be incorporated as subjective expert views, but with the advantage that sensitivity analyses can be done to measure the impact on the system

performance over time under different assumptions. One can test better which subjective views might lead to more realistic insights using a system dynamic model. Software that helps in such modeling includes Vensim, Powersim, and ithink; compare Sterman (2000).

Cross References

►Bayesian Statistics
►Stochastic Processes

References and Further Reading

Forrester JW (1961) Industrial dynamics. MIT Press, Cambridge, MA
Forrester JW (1980) Information sources for modeling the national economy. J Am Stat Assoc 75(371):555–574
Savage LJ (1980) The writing of Leonard Jimmie savage – a memorial collection. The American Statistical Association and the Institute of Mathematical Statistics
Senge P (1990) The fifth discipline: the art and practice of the learning organization. Doubleday, Boston
Sterman JD (2000) Business dynamics: systems thinking and modeling for a complex world. McGraw-Hill, New York

Modeling Survival Data

Edward L. Melnick
Professor of Statistics
New York University, New York, NY, USA

►Survival Data are measurements in time from a well defined origin until a particular event occurs. The event is usually death (e.g., lifetime from birth to death), but it could also be a change of state (e.g., occurrence of a disease or time to failure of an electrical component).

Of central importance to the study of risk is the probability that a system will perform and maintain its function (remain in a state) during a specified time interval $(0, t)$. Let $F(t) = P(T \le t)$ be the cumulative distribution function for the probability that a system fails before time t and conversely $R(t) = 1 - F(t)$ be the survival function for the system. Data from survival studies are often censored (the system has not failed during the study) so that survival times are larger than censored survival times. For example, if the response variable is the lifetime of an individual (or component), then the censored data are represented as (y_i, δ_i) where the indicator variable δ is equal to 1 if the event occurred during the study, and 0 if the event occurred after the study; i.e., $t_i = y_i$ if $\delta_i = 1$ and $t_i > y_i$ if $\delta_i = 0$. Further, if $f(t)dt$ is the probability of failure in the infinitesimal interval $(t, t + dt)$, then rate of a failure among items that have survived to time t is

$$h(t) = \frac{f(t)}{R(t)} = \frac{-d \ln R(t)}{dt}. \quad (1)$$

The function $h(t)$ is called the hazard function and is the conditional probability of failure, conditioned upon survival up to time t. The log likelihood function of (y_i, δ_i) is

$$\ln L = \delta_i \ln f(y_i) + (1 - \delta_i) \ln R(y_i), \quad (2)$$

and the cumulative hazard rate is

$$H(t) = \int_0^t h(x)dx. \quad (3)$$

The survival rate, $R(t)$, is equivalent to $R(t) = \exp(-H(t))$.

Examining the hazard function, it follows that

1. If $h(t)$ increases with age, $H(t)$ is an increasing failure rate. This would be the case for an object that wears out over time.
2. If $h(t)$ decreases with age, $H(t)$ is a decreasing failure rate. Examples of these phenomena include infant mortality and burn-in periods for engines.
3. If $h(t)$ is constant with age, $H(t)$ is a constant failure rate. In this situation failure time does not depend on age.

Note that $h(t)$ is a conditional probability density function since it is the proportion of items in *service* that fail per unit time. This differs from the probability density function $f(t)$, which is the proportion of the *initial* number of items that fail per unit time.

Distributions for failure times are often determined in terms of their hazard function. The exponential distribution function has a constant hazard function. The lognormal distribution function with standard deviation greater than 1 has a hazard function that increases for small t, and then decreases. The lognormal hazard function for standard deviation less than 1 has maximum at $t = 0$ and is often used to describe length of time for repairs (rather than modeling times to failure).

The ►Weibull distribution is often used to describe failure times. Its hazard function depends on the shape parameter m. The hazard function decreases when $m < 1$, increases when $m > 1$ and is constant when $m = 1$. Applications for this model include structured components in a system that fails when the weakest components fail, and for failure experiences that follow a bathtub curve. A bathtub failure time curve (convex function) has three stages: decreasing (e.g., infant mortality), constant (e.g., useful region), and increasing (e.g., wear out region). This curve is formed by changing m over the three regions. The basic

Modeling Survival Data. Table 1 Basic probability functions used to model survival data

Parametric		
Name	**Cumulative distribution function**	**Hazard function**
Exponential	$F(t) = 1 - \exp(-\lambda t) \quad \lambda > 0$	λ
Weibull	$F(t) = 1 - \exp(-\lambda t^m) \quad \lambda > 0$	$m\lambda$
Gumbel	$F(t) = 1 - \exp(-m(\exp(\lambda t) - 1)) \quad \lambda, m > 0$	$m\lambda \exp(\lambda t)$
Gompertz	$F(t) = 1 - \exp\left(\frac{m}{\lambda}(1 - \exp(\lambda t))\right) \quad \lambda, m > 0$	$m \exp(\lambda t)$
Nonparametric		
[a]Piecewise constant rates of change		$\sum_{i=1}^{n} \lambda_i I\{t_{i-1} < t < t_i\}$
[b]Kaplan–Meier	$\hat{F}(t) = 1 - \prod_{t_i \le t} \left(1 - \frac{d_i}{r_i}\right)$	$\frac{d_i}{r_i(t_{i+1} - t_i)}$
[c]Nelson–Aalen		$\hat{H}(t) = \sum_{t_i \le t} \left(1 - \frac{d_i}{r_i}\right)$

[a]The time axis is split into intervals such that $t_1 < t_2 < \ldots < t_n$ resulting in a non-continuous hazard function with jumps at the interval end points. The notation $I\{A\}$ is 1 if an event occurs in interval A, and is zero otherwise.
[b]The set $t_i \le \ldots \le t_n$ are the ordered event times where r_i are the number of individuals at risk at time t_i and d_i are the total number of individuals either experiencing the event or were censored at time t_i.
[c]The Nelson-Aalen statistic is an estimate of the cumulative hazard rate. It is based on the Poisson distribution.

probability functions used to model ▶survival data are in Table 1. These distributions are left skewed with support on $(0, \infty)$ for continuous distributions and support on the counting numbers $(0, n]$ for discrete distributions.

Nonparametric approaches have also been developed for estimating the survival function. A first approach might be the development of an empirical function such as:

$$\widehat{R}(t) = \frac{\text{Number of individuals with event times} \ge t}{\text{Number of individuals in the data set}}. \quad (4)$$

Unfortunately, this estimate requires that there are no censored observations. For example, an individual whose survival time is censored before time t cannot be used when computing the empirical function at t. This issue is addressed by introducing the ▶Kaplan–Meier estimator [see Kaplan and Meier (1958)]. Further, the variance of the Kaplan–Meier statistic can be estimated and confidence intervals can be constructed based on the normal distribution. Closely related to the Kaplan-Meier estimator is the Nelson–Aalen estimator (Nelson 1972; Aalen 1978) of the cumulative hazard rate function. The estimated variance and confidence interval can also be computed for this function.

Although the models already discussed assume that the occurrences of hazards are independent and identically distributed, often there are known risk factors such as environmental conditions and operating characteristics that affect the quality of a system.

In many problems a researcher is not only interested in the probability of survival, but how a set of explanatory variables affect the survival rate. Cox (1972) proposed the proportional hazard model that allows for the presence of covariates and the partial likelihood estimation procedure for estimating the parameters in the model. The proportional hazard model is of the form:

$$\lambda(t|\underline{Z}) = \lambda_0(t) \exp(\underline{Z}^T \underline{\beta}) \quad (5)$$

where

$\lambda_0(t)$ is the hazard function of unspecified shape (the subscript 0 implies all covariates are zero at time t).

$\underline{Z}$ is a vector of risk factors measured on each individual.

$\underline{\beta}$ is a vector of parameters describing the relative risk associated with the factors.

$\lambda(t|\underline{Z})$ is the hazard function at time t conditioned on the covariates.

The proportional hazard model is semi-parametric because no assumptions are made about the base hazard function but the effect of the risk factors is assumed to be linear on the log of the hazard function; i.e., $\lambda_0(t)$ is an infinite dimensional parameter and $\underline{\beta}$ is finite dimensional.

The proportionality assumption implies that if an individual has a risk of an event twice that of another individual, then the level of risk will remain twice as high for all time. The usual application of the model is to study the effect of the covariates on risk when absolute risk is less important. For example, consider a system where two types of actions can be taken, let

$$Z = \begin{cases} 1 & \text{if the high risk action is taken} \\ 0 & \text{if the low risk action is taken} \end{cases}$$

and let β be the relative risk associated with Z. The relative risk of the two types of actions is computed from the hazard ratio:

$$\frac{\lambda(t|Z=1)}{\lambda(t|Z=0)} = \exp\beta, \tag{6}$$

the instantaneous risk conditioned on survival at time t. In this problem the model describes relative risks and removes the effect of time. In a more general context, the ratio of hazards is the difference of covariates assuming the intercept is independent of time.

In many applications $\lambda_0(t)$ is unknown and cannot be estimated from the data. For example, the proportional hazard model is often used in credit risk modeling for corporate bonds based on interest rates and market conditions. A nonparametric estimation procedure for the conditional proportional hazard function is based on the exponential regression model:

$$\frac{\lambda(t|Z)}{\lambda_0(t)} = \exp(\underline{Z}^T\underline{\beta})$$

where the underlying survival function is estimated with a Kaplan–Meier estimator, a measure of time until failure.

If, however, the absolute risk is also important (usually in prediction problems), then the Nelson–Aalen estimate is preferred over the Kaplan–Meier estimator. The state space time series model [see Commandeur and Koopman (2007)] is useful for predicting risk over time and by using the Kalman Filter, can also include time varying covariates.

The proportional hazard model assumes event times are independent, conditioned on the covariates. The ▶frailty model relaxes this assumption by allowing for the presence of unknown covariates (random effects model). In this model event times are conditionally independent when values are given for the frailty variable. A frailty model that describes unexplained heterogeneity resulting from unobserved risk factors has a hazard function of the form

$$\lambda_{T_{ji}}(t) = w_{ji}\lambda_0(t)\exp\left(\underline{Z}_i^T\underline{\beta}_i\right) \tag{7}$$

where

T_{ji} is the time to failure (event) j for individual i, and

w_{ji} is the frailty variable.

In this model the frailty variable is constant over time, is shared by subjects within a subgroup, and acts multiplicatively on the hazard rates of all members of the subgroup. The two sources of variation for this model are:

1. Individual random variation described by the hazard function.
2. Group variation described by the frailty variable.

The log likelihood function, Eq. 2, for this model can be expressed in simple form if the hazard function has a Gompertz distribution and the frailty variable has a ▶gamma distribution. Other commonly used distributions for the frailty variable are the gamma, compound Poisson, and the lognormal. Estimators for situations where the likelihood function does not have an explicit representation are derived from the penalized partial likelihood function or from algorithms such as EM or Gibbs sampling.

Survival models have also been extended to multivariate conditional frailty survival functions. In the univariate setting, frailty varies from individual to individual whereas in the multivariate setting, frailty is shared with individuals in a subgroup. Consider, for example, the multivariate survival function conditioned on the frailty variable w:

$$s(t_1,\ldots,t_k|w) = \exp\left[-w(\Lambda_1(t_1),\ldots,\Lambda_k(t_k))\right], \tag{8}$$

where $\Lambda_i(t_i)$ is the cumulative hazard rate for group i. By integrating over w, the survival function is:

$$s(t_1,\ldots,t_k) = E\exp\left[-w(\Lambda_1(t_1),\ldots,\Lambda_k(t_k))\right], \tag{9}$$

the Laplace transform of w. Because of the simplicity of computing derivatives from the Laplace transform, this method is often used to derive frailty distributions. The most often assumed distributions are those from the gamma family. See Hougaard (2008) for a complete discussion on modeling multivariate survival data.

Conclusion

This paper presents a discussion for analyzing and modeling time series survival data. The models are then extended to include covariates primarily based upon regression modeling, and finally generalized to include multivariate models. Current research is focused on the development of multivariate time series models for survival data.

About the Author

Edward Melnick is Professor of Statistics and former Chair of the Department of Statistics and Operations Research at

Leonard N. Stern School of Business, New York University. He is an editor (with Brian Everitt) of the four volume *Encyclopedia of Quantitative Risk Analysis and Assessment* (Wiley Blackwell 2008), "valuable reference work ... and a rather beautiful work" (David Hand, *International Statistical Review*, Volume 77, Issue 2, p. 314). The number and impact of his publications were recognized by the American Statistical Association (ASA) when he became Fellow of the ASA. He is also Fellow of the Royal Statistical Society, and Elected Member of the International Statistical Institute. He was Chairman of the Risk Analysis section of the American Statistical Association (2004). Professor Melnick has won 16 teaching awards at NYU including the NYU Distinguished Teaching Award. Currently, he is an Associate Editor of the *Journal of Forecasting*.

Cross References

►Bayesian Semiparametric Regression
►Censoring Methodology
►Degradation Models in Reliability and Survival Analysis
►Demographic Analysis: A Stochastic Approach
►Event History Analysis
►First-Hitting-Time Based Threshold Regression
►Frailty Model
►Generalized Weibull Distributions
►Hazard Ratio Estimator
►Hazard Regression Models
►Kaplan-Meier Estimator
►Life Table
►Logistic Distribution
►Medical Research, Statistics in
►Population Projections
►Statistical Inference in Ecology
►Survival Data
►Time Series Models to Determine the Death Rate of a Given Disease
►Weibull Distribution

References and Further Reading

Aalen OO (1978) Nonparametric inference for a family of counting processes, Ann Stat 6:701–726
Commandeur JJ F, Koopman SJ (2007) An introduction to state space time series analysis. Oxford University Press, Oxford
Cox DR (1972) Regression models and life tables (with discussion). J R Stat Soc B 74:187–220
Hougaard P (2000) Analysis of multivariate survival data. Springer, New York
Jia J, Dyer JS, Butler JC (1999) Measures of perceived risk. Manage Sci 45:519–532
Johnson N, Kotz S, Kemp A (1993) Univariate discrete distributions, 2nd edn. Wiley, New York
Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481
Nelson W (1972) Theory and applications of hazard plotting for censored failure data, Technometrics 14:945–965
Von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton

Models for Z_+-Valued Time Series Based on Thinning

Emad-Eldin A. A. Aly
Professor
Kuwait University, Safat, Kuwait

Introduction

Developing models for integer-valued time series has received increasing attention in the past two decades. Integer-valued time series are useful in modeling dependent count data. They are also useful in the simulation of dependent discrete random variables with specified distribution and correlation structure.

Lawrance and Lewis (1977) and Gaver and Lewis (1980) were the first authors to construct autoregressive processes with non-Gaussian marginals. This has essentially motivated all the research on integer-valued time series. The present review is far from being exhaustive. Our focus is on models for Z_+-valued first-order autoregressive processes $INAR(1)$. We will consider five approaches which are based on "thinning" for developing these models.

First construction

To introduce integer-valued autoregressive moving average processes, McKenzie (1986, 1988) and Al-Osh and Alzaid (1987) used the binomial thinning operator $\odot$ of Steutel and van Harn (1979). The operation $\odot$ is defined as follows: if X is a Z_+−valued random variable (rv) and $\alpha \in (0,1)$, then

$$\alpha \odot X = \sum_{i=1}^{X} Y_i,$$

where $\{Y_i\}$ is a sequence of *i.i.d.* Bernoulli(α) $rv's$ independent of X. A sequence $\{X_n\}$ is said to be an $INAR(1)$ process if for any $n \in Z$,

$$X_n = \alpha \odot X_{n-1} + \varepsilon_n, \tag{1}$$

where $\odot$ is as in (1) and $\{\varepsilon_n\}$ is a sequence of *i.i.d.* Z_+−valued $rv's$ such that ε_n is independent of $\eta \odot X_{n-1}$ and the thinning $\eta \odot X_{n-1}$ is performed independently for each n. McKenzie (1986) constructed stationary Geometric

and Negative Binomial $INAR(1)$ processes and Al-Osh and Alzaid (1987) and independently McKenzie (1988) studied the Poisson $INAR(1)$ process.

Second Construction

Du and Li (1991) generalized the model (1) by introducing the $INAR(p)$ process

$$X_n = \sum_{i=1}^{p} \alpha_i \odot X_{n-i} + \varepsilon_n, \tag{2}$$

where all the thinning processes are independent and for $j < n$,

$$\operatorname{cov}(X_j, \varepsilon_n) = 0.$$

They proved that (2) has a unique stationary Z_+−valued solution $\{X_n\}_{n\in Z}$ if the roots of

$$\lambda^p - \sum_{i=1}^{p} \alpha_i \lambda^{p-i} = 0$$

are inside the unit circle. The main feature of the work of Du and Li (1991) is that it allows for models whose autocorrelation function (ACF) mimics that of the Normal *ARIMA* models.

Latour (1998) generalized Du and Li (1991) model by introducing the general $INAR(p)$ process ($GINAR(p)$),

$$X_n = \sum_{i=1}^{p} \alpha_i \circ X_{n-i} + \varepsilon_n,$$

where

$$\alpha_i \circ X_{n-i} = \sum_{i=1}^{X_{n-i}} Y_i^{(n,i)}$$

$\left\{Y_j^{(n,j)}\right\}$ is a sequence of nonnegative *i.i.d.rv's* independent of the *X's* with finite mean $\alpha_j > 0, j = 1, 2, \ldots, p$ and finite variance β_j and the innovation, ε_n, is assumed to have a finite mean μ_ε and finite variance σ_ε^2. Latour (1998) proved the existence of a stationary $GINAR(p)$ process if $\sum_{j=1}^{p} \alpha_j < 1$. He also showed that a stationary $GINAR(p)$ process, centered around its mean μ_X, admits a standard $AR(p)$ representation with the spectral density

$$f(\lambda) = \frac{\mu_X \sum_{j=1}^{p} \beta_j + \sigma_\varepsilon^2}{2\pi \left|\alpha\left(\exp(-i\lambda)\right)\right|^2}, \lambda \in [-\pi, \pi],$$

where

$$\alpha(t) = 1 - \sum_{j=1}^{p} \alpha_j t^j.$$

Third Construction

In the third approach the $INAR(1)$ stationary time series model takes the form

$$X_n = A_n(X_{n-1}, \eta) + \varepsilon_n, \tag{3}$$

where $\{\varepsilon_n\}$ are *i.i.d.r.v.'s* from the same family as the marginal distribution of $\{X_n\}$ and $A_n(X_{n-1}, \eta)$ is a random contraction operation performed on X_{n-1} which reduces it by the "amount η." Let $G_\theta(\cdot; \lambda_i)$ be the distribution of $Z_i, i = 1, 2$ and assume that Z_1 and Z_2 are independent and $G_\theta(\cdot; \lambda_1) * G_\theta(\cdot; \lambda_2) = G_\theta(\cdot; \lambda_1 + \lambda_2)$, where $*$ is the convolution operator. Let $G(\cdot; x, \lambda_1, \lambda_2)$ be the conditional distribution of Z_1 given $Z_1 + Z_2 = x$. The distribution of the random operator $A(X, \eta)$ given $X = x$, is defined as $G(\cdot; x, \eta\lambda, (1-\eta)\lambda)$. The distribution of $A(X, \eta)$ is $G_\theta(\cdot; \eta\lambda)$ when the distribution of X is $G_\theta(\cdot; \lambda)$. Now, if the distributions of X_0 and ε_1 are respectively $G_\theta(\cdot; \lambda)$ and $G_\theta(\cdot; (1-\eta)\lambda)$, then $\{X_n\}$ of (3) is stationary with marginal distribution $G_\theta(\cdot; \lambda)$. This construction was employed by Al-Osh and Alzaid (1991) for the Binomial marginal and Alzaid and Al-Osh (1993) for the Generalized Poisson marginal. This construction was generalized to the case when X_0 is infinitely divisible by Joe (1996) and to the case when X_0 is in the class of Exponential Dispersion Models by Jørgensen and Song (1998).

Fourth Construction

This construction is based on the expectation thinning operator $K(\eta)\circledast$ of Zhu and Joe (2003). The expectation thinning operator $K(\eta)\circledast$ is defined as follows: if X is a Z_+−valued *rv* and $\eta \in (0,1)$, then

$$K(\eta) \circledast X = \sum_{i=1}^{X} K_i(\eta),$$

where $K_i(\eta)$ are *i.i.d.r.v.'s* and the family $\{K(\alpha) : 0 \le \alpha \le 1\}$ is self-generalized, i.e., $E\{K(\eta) \circledast X | X = x\} = \eta x$ and $K(\eta') \circledast K(\eta) = K(\eta\eta')$. The corresponding $INAR(1)$ stationary time series model takes the form

$$X_n \overset{d}{=} K(\eta) \circledast X_{n-1} + \epsilon(\eta) = \sum_{i=1}^{X_{n-1}} K_i(\eta) + \epsilon(\eta).$$

The marginal distribution of X_n must be generalized discrete self-decomposable with respect to K, that is, $P_{X_n}(z)/P_{X_n}(P_{K(\alpha)}(z))$ must be a proper probability generating function (PGF) for every $\alpha \in [0,1]$. The ACF at lag$k$ is $\rho(k) = \eta^k$. The expectation thinning $K(\eta)\circledast$ governs the serial dependence. Several families of self-generalized *r.v.'s* $\{K(\eta)\}$ are known and the corresponding stationary distributions of $\{X_n\}$ are overdispersed with respect to Poisson (e.g., Generalized Poisson, Negative Binomial, Poisson-Inverse Gaussian). When a marginal distribution is possible for more than one self-generalized family then different $\{K(\eta)\}$ lead to differing amounts of conditional heteroscedasticity.

Fifth Construction

The fifth approach makes use of the thinning operator $\odot_{\mathcal{F}}$ of van Harn et al. (1982) and van Harn and Steutel (1993) which is defined as follows. Let $\mathcal{F} := (F_t, t \geq 0)$ be a continuous composition semigroup of PGF's such that $F_t(0) \neq 1, \delta = \delta(\mathcal{F}) = -\ln F_1'(1) > 0, F_{0+}(z) = z.$ and $F_{\infty-}(z) = 1.$ The infinitesimal generator U of $\mathcal{F}$ is given for $|z| \leq 1$ by

$$U(z) = \lim_{t \to 0+} \frac{F_t(z) - z}{t} = a\{H(z) - z\},$$

where a is a constant and $H(z) = \sum_{n=0}^{\infty} h_n z^n$ is a PGF of a Z_+ valued rv with $h_1 = 0$ and $H'(1) \leq 1$. For a Z_+ valued rv X and $\eta \in (0,1)$

$$\eta \odot_{\mathcal{F}} X = \sum_{i=1}^{X} Y_i,$$

where $\{Y_i\}$ is a sequence of *i.i.d.r.v.'s* independent of X with common PGF $F_{-\ln \eta} \in \mathcal{F}$. The corresponding $\mathcal{F}-$first order integer-valued autoregressive $(\mathcal{F}-INAR(1))$ model takes the form

$$X_n = \eta \odot_{\mathcal{F}} X_{n-1} + \varepsilon_n, \tag{4}$$

where $\{\varepsilon_n\}$ is a sequence of *i.i.d.* Z_+ valued *rv's* such that ε_n is independent of $\eta \odot_{\mathcal{F}} X_{n-1}$ and the thinning $\eta \odot_{\mathcal{F}} X_{n-1}$ is performed independently for each n. Note that $\{X_n\}$ is a Markov chain (see ▶Markov Chains). In terms of PGF's (4) reads

$$P_{X_n}(z) = P_{X_{n-1}}(F_{-\ln \eta}(z)) P_{\varepsilon}(z). \tag{5}$$

A distribution on Z_+ with PGF $P(z)$ is $\mathcal{F}$-self-decomposable (van Harn et al. (1982)) if for any t there exists a PGF $P_t(z)$ such

$$P(z) = P(F_t(z)) P_t(z).$$

Aly and Bouzar (2005) proved that any $\mathcal{F}$-self-decomposable distribution can arise as the marginal distribution of a stationary $\mathcal{F}-INAR(1)$ model. On assuming that the second moments of each of $H(\cdot), \varepsilon$ and X_n are finite for any $n \geq 0$, Aly and Bouzar (2005) proved that (1) the regression of X_n on X_{n-1} is linear, (2) the variance of X_n given X_{n-1} is linear, (3) the ACF at lag k, $\rho(X_{n-k}, X_n) = \eta^{\delta k} \sqrt{V(X_{n-k})/V(X_n)}$. Moreover, if $\{X_n\}$ is stationary, then $\rho(k) = \rho(X_{n-k}, X_n) = \eta^{\delta k}$.

We consider some important stationary time series models based on the composition semigroup

$$F_t^{(\theta)}(z) = 1 - \frac{\overline{\theta} e^{-\overline{\theta} t}(1-z)}{\overline{\theta} + \theta(1 - e^{-\overline{\theta} t})(1-z)}, t \geq 0, |z| \leq 1,$$
$$\overline{\theta} = 1 - \theta, 0 \leq \theta < 1$$

of van Harn et al. (1982). Note that when $\theta = 0$, $F_t^{(0)}(z) = 1 - e^{-t} + e^{-t} z$ and the corresponding thinning is the Binomial thinning of Steutel and van Harn (1979). In this case (4) becomes

$$P_X(z) = P_X(1 - \eta + \eta z) P_{\varepsilon}(z). \tag{6}$$

Particular $INAR(1)$ of (6) are the Poisson (Al-Osh and Alzaid 1987; McKenzie 1988), the Geometric and the Negative Binomial (McKenzie 1986), the Mittag-Leffler (Pillai and Jayakumar 1995) and the discrete Linnik (Aly and Bouzar 2000). Particular $INAR(1)$ time series models when $0 < \theta < 1$ are the Geometric, the Negative Binomial and the Poisson Geometric (Aly and Bouzar 1994) and the Negative Binomial (Al-Osh and Aly 1992).

Remarks

We mention some methods of parameter estimation. The most direct approach is using moment estimation based on the Yule-Walker equations. The conditional least squares method with some modifications, e.g., a two-stage procedure, in order to be able to estimate all the parameters (see, for example, Brännäs and Quoreshi 2004) may be used. Joe and Zhu (2006) used the method of maximum likelihood after using a recursive method to calculate the probability mass function of the innovation. Neal and Subba Rao (2007) used the MCMC approach for parameter estimation. For additional references on parameter estimation we refer to Brännäs (1994), Jung and Tremayne (2006), Silva and Silva (2009) and the references contained therein. Finally, we note that Hall and Scotto (2006) studied the extremes of integer-valued time series.

About the Author

Dr Emad-Eldin A. A. Aly is a Professor since 1994 at the Department of Statistics and Operations Research, Kuwait University, Kuwait. He was the Chair of the Department (2002–2006), and the Vice Dean for Academic Affairs of the Faculty of Graduate Studies, Kuwait University (1996–2002). He was a Faculty member at The University of Alberta, Edmonton, Alberta, Canada (1984–1995) and the Chair of the Department of Statistics and Applied Probability, The University of Alberta (1991–1994). He has authored and co-authored more than 75 papers. He was an Associate Editor of the *Journal of Nonparametric Statistics*. He was awarded (jointly with Professor A. Alzaid of King Saud University) the 1995 Kuwait Prize of the Kuwait Foundation for the Advancement of Sciences for his research in Mathematical Statistics.

Cross References

►Box–Jenkins Time Series Models
►Generalized Quasi-Likelihood (GQL) Inferences
►Time Series

References and Further Reading

Al-Osh MA, Aly E-EAA (1992) First order autoregressive time series with negative binomial and geometric marginals. Commun Statist Theory Meth 21:2483–2492

Al-Osh MA, Alzaid A (1987) First order integer-valued autoregressive (INAR(1)) process. J Time Ser Anal 8:261–275

Al-Osh MA, Alzaid A (1991) Binomial autoregressive moving average models. Commun Statist Stochastic Models 7:261–282

Aly E-EAA, Bouzar N (1994) Explicit stationary distributions for some Galton Watson processes with immigration. Commun Statist Stochastic Models 10:499–517

Aly E-EAA, Bouzar N (2000) On geometric infinite divisibility and stability. Ann Inst Statist Math 52:790–799

Aly E-EAA, Bouzar N (2005) Stationary solutions for integer-valued autoregressive processes. Int J Math Math Sci 1:1–18

Alzaid AA, Al-Osh MA (1993) Some autoregressive moving average processes with generalized Poisson marginal distributions. Ann Inst Statist Math 45:223–232

Brännäs K (1994) Estimation and testing in integer-valued AR(1) models. Umeå Economic Studies No. 335

Brännäs K, Quoreshi AMMS (2004) Integer-valued moving average modeling of the number of transactions in stocks. Umeå Economic Studies No. 637

Du JG, Li Y (1991) The integer-valued autoregressive INAR(p) model. J Time Ser Anal 12:129–142

Gaver DP, Lewis PAW (1980) First-order autoregressive gamma sequences and point processes. Adv Appl Probab 12:724–745

Hall A, Scotto MG (2006) Extremes of periodic integer-valued sequences with exponential type tails Revstat 4:249–273

Joe H (1996) Time series models with univariate margins in the convolution-closed infinitely divisible class. J Appl Probab 33:664–677

Jørgensen B, Song PX-K (1998) Stationary time series models with exponential dispersion model margins. J Appl Probab 35:78–92

Jung RC, Tremayne AR (2006) Binomial thinning models for integer time series. Statist Model 6:81–96

Latour A (1998) Existence and stochastic structure of a non-negative integer-valued autoregressive process. J Time Ser Anal 19:439–455

Lawrance AJ, Lewis PAW (1977) An exponential moving average sequence and point process, EMA(1). J Appl Probab 14:98–113

McKenzie E (1986) Autoregressive-moving average processes with negative binomial and geometric marginal distributions. Adv Appl Probab 18:679–705

McKenzie E (1988) Some ARMA models for dependent sequences of Poisson counts. Adv Appl Probab 20:822–835

Neal P, Subba Rao T (2007) MCMC for integer valued ARMA Models. J Time Ser Anal 28:92–110

Pillai RN, Jayakumar K (1995) Discrete Mittag-Leffler distributions. Statist Probab Lett 23:271–274

Silva I, Silva ME (2009) Parameter estimation for INAR processes based on high-order statistics. Revstat 7:105–117

Steutel FW, van Harn K (1979) Discrete analogues of self-decomposability and stability. Ann Probab 7:893–899

van Harn K, Steutel FW (1993) Stability equations for processes with stationary independent increments using branching processes and Poisson mixtures. Stochastic Process Appl 45:209–230

van Harn K, Steutel FW, Vervaat W (1982) Self-decomposable discrete distributions and branching processes. Z Wahrsch Verw Gebiete 61:97–118

Zhu R, Joe H (2003) A new type of discrete self-decomposability and its application to continuous-time Markov processes for modelling count data time series. Stochastic Models 19:235–254

Zhu R, Joe H (2006) Modelling count data time series with Markov processes based on binomial thinning. J Time Ser Anal 27: 725–738

Moderate Deviations

Jayaram Sethuraman
Robert O. Lawton Distinguished Professor, Professor Emeritus
Florida State University, Tallahassee, FL, USA

M

Moderate Deviations

Consider the familiar simple set up for the central limit theorem (CLT, see ►Central Limit Theorems). Let $X_1, X_2, \ldots$ be independently and identically distributed real random variables with common distribution function $F(x)$. Let $Y_n = \frac{1}{n}(X_1 + \cdots + X_n)$, $n = 1, 2, \ldots$. Suppose that

$$\int xF(dx) = 0, \int x^2 F(dx) = l \tag{1}$$

Then the central limit theorem states that

$$P\left(|Y_n| > \frac{a}{\sqrt{n}}\right) \to 2[1 - \Phi(a)] \tag{2}$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-t^2/2)dt$ and $a > 0$.

In other words, the CLT gives an approximation to the two-sided deviation of size $\frac{a}{\sqrt{n}}$ of Y_n and the approximation is a number in $(1/2, 1)$. Deviations of the this type are called *ordinary deviations*.

However, one needs to study deviations larger than ordinary deviations to understand finer properties of the distributions of Y_n and to approximate expectations of other functions of Y_n. Thus a deviation of magnitude λ_n will be called a *excessive deviation* if $n\lambda_n^2 \to \infty$. In the particular case of $\lambda_n = \lambda$ where λ is a constant, it is called a *large deviation* (see also ►Large Deviations and Applications).

The following, due to Cramér (1938), Chernoff (1952), Bahadur and Rao (1960), etc., is a classical result on large deviations. Let

$$\int \exp(tx)F(dx) < \infty \text{ for } t \text{ in some neighborhood of } 0. \tag{3}$$

Then

$$\frac{1}{n}\log P(|Y_n| > \lambda) \to -I(\lambda) \tag{4}$$

where

$$I(\lambda) = \sup_t(t\lambda - \log\phi(t)) \tag{5}$$

and $0 < I(\lambda) \leq \infty$. This result is usually read as "the probability of large deviations tends to zero exponentially." For sequences of random variables $\{Y_n\}$ distributed in more general spaces like $R^k, C([0,1]), D([0,1])$, etc. (i.e., ▶stochastic processes), there is no preferred direction for deviations. The appropriate generalization of the large deviation result (4) is the *large deviation principle*, which states that for all Borel sets A

$$-I(A^0) \leq \underline{\lim}_n \frac{1}{n}\log P(Y_n \in A) \leq -I(\overline{A}) \tag{6}$$

where $A^0, \overline{A}$ denote the interior and closure of A, and

$$I(A) = \inf_{\lambda \in A} I(\lambda) \tag{7}$$

for some function $I(\lambda)$ whose level sets $\{\lambda : I(\lambda) \leq K\}$ are compact for $K < \infty$. The function $I(x)$ is called the *large deviation rate function*.

When the moment generating function condition (3) holds, Cramér (1938) has further shown that

$$P(|Y_n| > \lambda_n) \sim \frac{2}{\sqrt{2\pi n\lambda_n^2}}\exp\left(\frac{-n\lambda_n^2}{2}\right) \tag{8}$$

when $n\lambda_n^3 \to 0$ and $n\lambda_n^2 \to \infty$. This excludes large deviations ($\lambda_n = \lambda$), but it gives a rate for the probability (and not just the logarithm of the probability) of a class of excessive deviations and is therefore called a *strong excessive deviation result*.

Rubin and Sethuraman (1965a) called deviations λ_n with $\lambda_n = c\sqrt{\frac{\log n}{n}}$ where c is a constant as *moderate deviations*. Moderate deviations found their first applications in Bayes risk efficiency which was introduced in Rubin and Sethuraman (1965b). Cramér's result in (8) reduces to

$$P(|Y_n| > c\sqrt{\frac{\log n}{n}}) \sim \frac{2}{c\sqrt{2\pi\log n}}n^{-c^2/2} \tag{9}$$

and holds under the moment generating function condition (3). Rubin and Sethuraman (1965a) showed that the moderate deviation result (9) holds under the weaker condition

$$E(|X_1|^{c^2+2+\delta}) < \infty \text{ for some } \delta > 0. \tag{10}$$

They also showed that when (9) holds we have

$$E(|X_1|^{c^2+2-\delta}) < \infty \text{ for all } \delta > 0. \tag{11}$$

Slastnikov (1978) showed that the strong moderate deviation result (9) if and only if

$$\lim_{t\to\infty} t^{2+c}(\log(t))^{-(1+c)/2}P(|X_1| > t) = 0. \tag{12}$$

Since (8) was called a strong excessive deviation result, we should call (9) as a *strong moderate deviation result*. Analogous to the logarithmic large deviation result (4) is the *logarithmic moderate deviation result* which states that

$$\frac{1}{\log(n)}\log P(|Y_n| \geq c\sqrt{\frac{\log(n)}{n}}) \sim n^{-c^2/2} \tag{13}$$

which may be the only possible result for more complicated random variables $\{Y_n\}$ than are not means of i.i.d. random variables,

For random variables $\{Y_n\}$ which take values in $R^k, C([0,1]), D([0,1])$ etc., we can, under some conditions, establish the *moderate deviation principle* which states

$$-J(A^0) \leq \underline{\lim}_n \frac{1}{\log(n)}P(\sqrt{\frac{n}{\log(n)}}Y_n \in A) \leq -J(\overline{A}) \tag{14}$$

where $J(A) = \inf_{x\in A} J(x)$ for some function $J(x)$ whose level sets are compact. The function $J(x)$ is then called the *moderate deviation rate function*. This is analogous to the large deviation principle (6).

Following the paper of Rubin and Sethuraman (1965a), there is a vast literature on moderate deviations for a large class of random variables $\{Y_n\}$ that arise in a multitude of contexts. The asymptotic distribution of $\{Y_n\}$ can be more general than Gaussian. We will give just a brief summary below.

We stated the definition of two-sided moderate deviations and quoted Slastnikov's necessary and sufficient condition. One can also consider one-sided moderate deviations results and the necessary and sufficient conditions are slightly different and these are given in Slastnikov (1978). Without assuming á priori that the mean and variance of the i.i.d. random variables $X_1, X_2 \ldots$ are 0 and 1 respectively, one can ask for necessary and sufficient conditions for moderate deviations. This problem has been completely addressed in Amosova (1979). Another variant of moderate deviations has been studied in Davis (1968).

The case where $\{Y_n\}$ is the sum of triangular arrays of independent random variables or a U-statistic were begun in Rubin and Sethuraman (1965). Ghosh (1974) studied moderate deviations for sums of m-dependent random variables. Michel (1974) gave results on rates of convergence in the strong moderate deviation result (9). Gut (1980) considered moderate deviations for random variables with multiple indices. Dembo (1996) considered moderate deviations for ▶martingales.

Moderate deviations in general topological spaces with applications in Statistical Physics and other areas can be found in Borovkov and Mogulskii (1978), (1980), Deo and Babu (1981), De Acosta (1992), Liming (1995), Djellout and Guillin (2001).

About the Author

Professor Jayaram Sethuraman earned a Ph.D. in statistics from the Indian Statistical Institute in 1962. Professor Sethuraman has received many recognitions for his contributions to the discipline of statistics: the U.S. Army S. S. Wilks Award (1994), the Teaching Incentive Program Award, FSU (1995), the Professorial Excellence Award, FSU (1996), an ASA Service Award (2001), the President's Continuing Education Award, FSU (2002), and the Bhargavi and C. R. Rao Prize, Pennsylvania State University (2005).

"Sethuraman has been a superior researcher throughout his career, making important contributions in many areas including asymptotic distribution theory, large deviations theory, moderate deviations theory for which he was the pioneer, limit theory, nonparametric statistics, Dirichlet processes and Bayesian nonparametrics, stopping times for sequential estimation and testing, order statistics, stochastic majorization, Bahadur and Pitman efficiency, Markov chain Monte Carlo, reliability theory, survival analysis and image analysis." (Myles Hollander (2008). A Conversation with Jayaram Sethuraman, *Statistical Science* 23, 2, 272–285).

Cross References

▶Central Limit Theorems
▶Estimation: An Overview
▶Large Deviations and Applications
▶Prior Bayes: Rubin's View of Statistics
▶Statistics on Ranked Lists

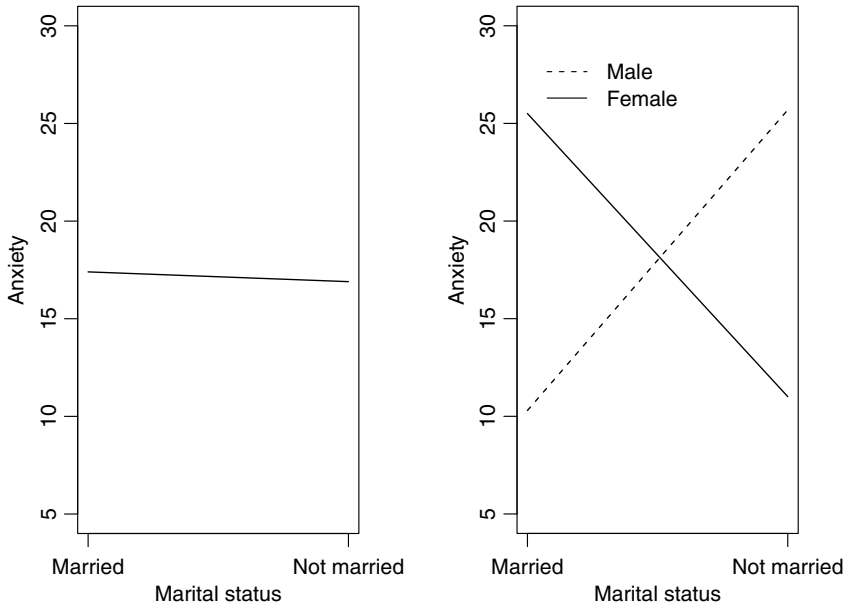
References and Further Reading

Borovkov AA, Mogulskii AA (1978) Probabilities of large deviations in topological vector space I. Siberian Math J 19:697–709
Borovkov AA, Mogulskii AA (1980) Probabilities of large deviations in topological vector space II. Siberian Math J 21:12–26
Cramér H (1938) Sur un nouveau théorème limite de la probabilités. Actualites Sci Indust 736:5–23
Davis AD (1968) Convergence rates for probabilities of moderate deviations. Ann Math Statist 39:2016–2028
De Acosta A (1992) Moderate deviations and associated Laplace approximations for sums of independent random vectors. Trans Am Math Soc 329:357–375
Dembo A (1996) Moderate deviations for martingales with bounded jumps. Elec Comm Probab 1:11–17
Deo CM, Babu JG (1981) Probabilities of moderate deviations in a Banach space. Proc Am Math Soc 24:392–397
Djellout H, Guillin A (2001) Moderate deviations for Markov chains with atom. Stoch Proc Appl 95:203–217
Gao FQ (2003) Moderate deviations and large deviations for kernel density estimators. J Theo Probab 16:401–418
Ghosh M (1974) Probabilities of moderate deviations under m-dependence. Canad J Statist 2:157–168
Gut A (1980) Convergence rates for probabilities of moderate deviations for sums of random variables with multidimensional indices. Ann Probab 8:298–313
Liming W (1995) Moderate deviations of dependent random variables related to CLT. Ann Probab 23:420–445
Michel R (1974) Results on probabilities of moderate deviations. Ann Probab 2:349–353
Rubin H, Sethuraman J (1965a) Probabilities of moderate deviations. Sankhya Ser A 27:325–346
Rubin H, Sethuraman J (1965b) Bayes risk efficiency. Sankhya Ser A 27:347–356
Slastnikov AD (1978) Limit theorems for moderate deviation probabilities. Theory Probab Appl 23:322–340

Moderating and Mediating Variables in Psychological Research

Petar Milin[1], Olga Hadžić[2]
[1]Associate Professor
University of Novi Sad, Novi Sad, Serbia
[2]Professor
University of Novi Sad, Novi Sad, Serbia

Moderating and mediating variables, or simply *moderators* and *mediators*, are related but distinct concepts in both general statistics and its application in psychology. A moderating variable is a variable that affects the relationship between two other variables. This effect is usually referred to as an *interaction*. The simplest case of an interaction can occur in ▶analysis of variance (ANOVA).



Moderating and Mediating Variables in Psychological Research. Fig. 1 The main effect of one categorical variable on a continuous dependent variable (*left-hand panel*), and how it is moderated by the third categorical variable (*right-hand panel*)

For example, we tested whether there is a significant difference in the *level of anxiety* (as measured with an appropriate standardized psychological test) between married and unmarried participants (i.e., variable *marital status*). The effect was not statistically significant. However, when we enter the third variable – *gender* (female/male) – it appears that, on average, unmarried males are significantly more anxious than married males, while for females the effect is the reverse. Figure 1 represents the results from two models described above. In the left-hand panel, we can see that, on average, there are no differences between married and unmarried participants in the level of anxiety. From the right-hand panel, we can conclude that gender moderates the effect of marital status on the level of anxiety: married males and unmarried females are significantly less anxious than the other two groups (unmarried males and married females).

We can generalize the previous example to more complex models, with two independent variables having more than just two levels for comparison, or even with more than two independent variables. If all variables in the model are continuous variables, we would apply multiple regression analysis, but the phenomenon of a moderating effect would remain the same, in essence. For example, we confirmed a positive relationship between the *hours of learning* and the *result in an assessment test*. Yet, *music loudness* during learning can moderate test results. We can imagine this as if a hand on the volume knob of an amplifier rotates clockwise and turns the volume up, students get all the worse results the longer they learn. Depending on the music volume level, the relationship between the hours of learning and the knowledge assessment changes continuously. This outcome is presented in Fig. 2. On the left-hand side, we can observe a positive influence of the hours of learning on the results in the assessment test, while on the right-hand side, we can see how music loudness moderates this relationship.

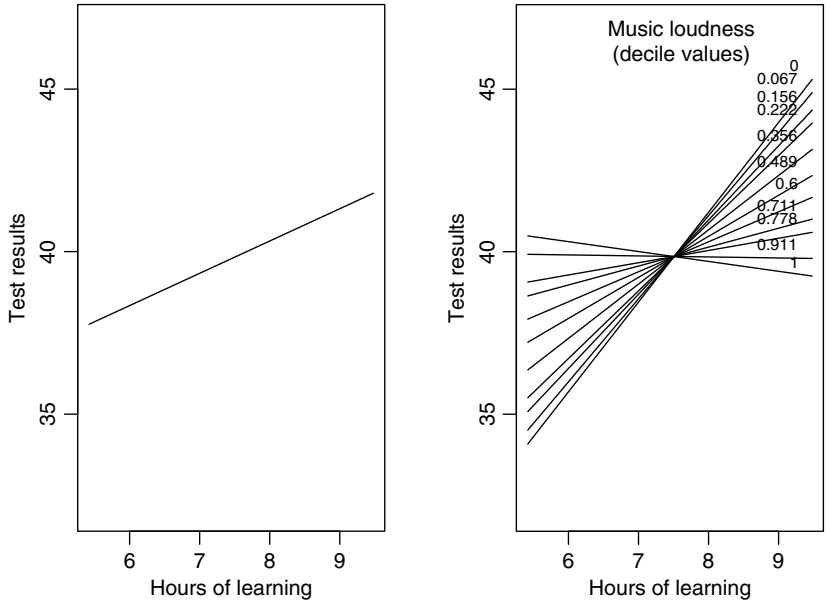
The general linear form with one dependent, one independent, and one moderating variable is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \varepsilon,$$

where β_3 evaluates the interaction between X_1 and X_2.

Mediating variables typically emerge in multiple regression analysis, where the influence of some independent variable (*predictor*) on the dependent variable (*criterion*) is not direct, but mediated through the third variable. For example, the correlation between *ageing* and the *number of work accidents* in the car industry appears to be strong and negative. Nevertheless, the missing link in this picture is *work experience*: it affects injury rate, and is itself affected by the age of worker.

In regression modeling, one can distinguish between *complete mediation* and *incomplete mediation*. In practice, if the effects of ageing on the number of work injuries



Moderating and Mediating Variables in Psychological Research. Fig. 2 The main effect of one continuous variable on another (*left-hand panel*), and how it is moderated by a third continuous variable (*right-hand panel*). Lines on the right panel represent decile values for the moderator variable

would not differ statistically from zero when work experience is included in the model, then mediation is complete. Otherwise, if this effect still exists (in the statistical sense), then mediation is incomplete. Complete and incomplete mediation are presented in Fig. 3.

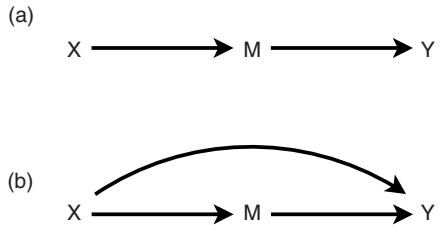
In principle, a mediating variable flattens the effect of an independent variable on the dependent variable. The opposite phenomenon would occur if the mediator variable would increase the effect. This is called *suppression*. It is a controversial concept in statistical theory and practice, but contemporary applied approaches take a more neutral position, and consider that suppression may provide better insights into the relationships between relevant variables.

The simplest case of linear regression with one dependent, one independent, and one mediating variable is defined by the following equations:

$$Y = \beta_0 + \beta_1 X + \varepsilon_1$$
$$M = \gamma_0 + \gamma_1 X + \varepsilon_2$$
$$Y = \beta_0' + \beta_1' X + \beta_2 M + \varepsilon_3,$$

where of particular interest are β_1, which is called the *total effect*, and β_1', named the *direct effect*. If suppression does not take place, which would occur if $\beta_1' > \beta_1$, then we can continue the analysis with a standard regression model. First, we ascertain whether mediation is complete or incomplete, depending on whether the direct effect drops to zero ($\beta_1' \approx 0$). The most important step in the analysis is the inference about the *indirect effect*, or the *amount of mediation*. It is defined as the reduction in the effect of the initial variable on the model outcome ($\beta_1 - \beta_1'$). In simple hierarchical regression models, the difference of the coefficients is exactly the same as the product of the effect of the independent variable on the mediating variable multiplied by the effect of the mediating variable on the dependent variable. In the general case, this equality only approximately holds.

Mediation and moderation can co-occur in statistical models. This is often the case in psychology. *Mediated moderation* takes place when the independent variable is actually an interaction ($X = X_A \times X_B$). Thus, the mediator acts between interacting variables (X_A and X_B) and the dependent variable (Y). For example, the effect of interacting variable *hours of learning* and *music loudness* on the dependent variable *result in an assessment test* can be mediated by the *importance of the test*, as rated by the participants. Conversely, *moderated mediation* is realized in two forms: (a) the effect of the independent variable on the mediator is affected by a moderator (γ_1 varies; as if the effect of *ageing* on *work experience* is moderated by a particular personality trait, like H. J. Eysenck's *Neuroticism*), or (b) a moderator may interact with the mediating variable (β_2 varies; as if the *work experience* and the *level of anxiety* would interact and mediate between *ageing* and *number of*



Moderating and Mediating Variables in Psychological Research. Fig. 3 Schematic representation of a complete mediation effect (panel a, *upper*), and an incomplete mediation effect (panel b, *lower*)

work accidents). If moderated mediation exists, inference about its type must be given.

Finally, special attention is required in moderation and mediation analyses since both can be influenced by ►*multicollinearity*, which makes estimates of regression coefficients unstable. In addition, in an analysis with a moderating term – i.e., an interaction effect – the product of the variables can be strongly related to either the independent or the moderating variable, or both of them. If two variables are collinear, one of them can be centred to its mean. In this way, half of its value will become negative, and consequently, collinearity will decrease. Another possibility is to regress the independent variable with a moderator or mediator, and then to use the *residuals* or unexplained values, of the independent variable in the main analysis. Thus, the independent variable will be orthogonal to the moderating or mediating variable, with zero correlation, which will bring collinearity under control. However, in applying the previous two remedies, and others that are available, one must choose a conservative approach. The risk of emphasizing, or even inventing, what is not present in the data ought to be as little as possible. In any circumstances, the ultimate way of securing more reliable estimates is simply to obtain enough data.

Acknowledgment

We would like to thank Professor David Kenny for reading a draft of this article, and providing us with comments and suggestions which resulted in many improvements.

About the Author

Dr. Olga Hadzic is Professor, Department of Mathematics and Informatics, University of Novi Sad, Serbia. She is an Elected Member of the Serbian Academy of Sciences and Arts (since 1984). Her research interests are in fixed point theory, functional analysis, probability theory, and organizational psychology. She has (co-)authored about 180 scientific papers, 5 monographs, and 4 textbooks, including, *Fixed Point Theory in Probabilistic Metric Spaces* (with Endre Pap, Kluwer Academic Publishers, Dordrecht 2001). Professor Hadzic was Rector (Chancellor) of the University of Novi Sad (1996–1998). She was an external adviser for two Ph.D. theses defended abroad.

Cross References

►Analysis of Variance
►Interaction
►Linear Regression Models
►Multilevel Analysis
►Psychology, Statistics in
►Variables

References and Further Reading

Baron R, Kenny D (1986) The moderator-mediator variable distinction in social psychological research – conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51(6):1173–1182

Eysenck H (2006) The biological basis of personality. Transaction Publishers, London

Friedman L, Wall M (2005) Graphical views of suppression and multicollinearity in multiple linear regression. Am Stat 59(2): 127–136

Hayes A, Matthes J (2009) Computational procedures for probing interactions in ols and logistic regression: SPSS and SAS implementations. Behav Res Meth 41(3):924–936

Judd C, Kenny D, McClelland G (2001) Estimating and testing mediation and moderation in within-participant designs. Psychol Meth 6(2):115–134

Muller D, Judd C, Yzerbyt V (2005) When moderation is mediated and mediation is moderated. J Pers Soc Psychol 89(6):852–863

Shrout P, Bolger N (2002) Mediation in experimental and nonexperimental studies: new procedures and recommendations. Psychol Meth 7(4):422–445

Moment Generating Function

Jan Beran[1], Sucharita Ghosh[2]
[1]Professor
University of Konstanz, Konstanz, Germany
[2]Scientific Staff Member
Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

The moment generating function (mgf) of a real valued random variable X with distribution $F(x) = P(X \leq x)$ is defined by

$$M_X(t) = E\left[e^{tX}\right] = \int e^{tx} dF(x). \qquad (1)$$

For distributions with a density function $f = F'$, M_X can also be interpreted as a (two-sided) Laplace transform of f. In order that M_X exists and is finite for $t \in (-a, a)$ and some $a > 0$, all moments $\mu_j = E\left[X^j\right]$ must be finite and such that $\sum \mu_j t^j/j!$ is a convergent series. We then have

$$M_X(t) = \sum_{j=0}^{\infty} \frac{\mu_j}{j!} t^j \tag{2}$$

so that

$$\mu_j = M_X^{(j)}(0) = \frac{d^j}{dt^j} M_X(t) \mid_{t=0} \tag{3}$$

which explains the name moment generating function. A counter example where M_X does not exist in any open neighborhood of the origin is the Cauchy distribution, since there even μ_1 is not defined. The lognormal distribution is an example where all μ_j are finite but the series in (2) does not converge. In cases where $X > 0$ and $M_X(t) = \infty$ for $t \neq 0$, the mgf of $-X$ may be used (see e.g., Severini (2005) for further results). Related to M_X are the characteristic function $\phi_X(t) = M_X(it)$ and the probability generating function $H_X(z) = E(z^X)$ for which $M_X(t) = H_X(e^t)$. Note however that, in contrast to M_X, $\phi_X(t) = E\left[\exp(itX)\right]$ always exists. A furhter important function is the cumulant generating function $K_X(t) = \log M_X(t)$ which can be written as power series

$$K_X(t) = \sum_{j=1}^{\infty} \frac{\kappa_j}{j!} t^j \tag{4}$$

where κ_j are cumulants. The first two cumulants are $\kappa_1 = \mu = E(X)$ and $\kappa_2 = \sigma^2 = var(X)$. In contrast to the raw moments μ_j, higher order cumulants κ_j ($j \geq 3$) do not depend on the location μ and scale σ^2. For vector valued random variables $X = (X_1, ..., X_k)' \in \mathbb{R}^k$, M_X is defined in an analogous manner by $M_X(t) = E\left[\exp\left(t'X\right)\right] = E\left[\exp\left(\sum_{j=1}^k t_j X_j\right)\right]$. This implies

$$\frac{\partial^{j_1+j_2+\cdots+j_k}}{\partial t_1^{j_1} \partial t_1^{j_2} \cdots \partial t_1^{j_k}} M_X(0) = E\left[X_1^{j_1} X_2^{j_2} \cdots X_k^{j_k}\right] \tag{5}$$

and corresponding expressions for joint cumulants as derivatives of K_X. In particular,

$$\frac{\partial^2}{\partial t_i \partial t_j} K_X(0) = cov(X_i, X_j). \tag{6}$$

An important property is uniqueness: if $M_X(t)$ exists and is finite in an open interval around the origin, then there is exactly one distribution function with this moment generating function. For instance, if $\kappa_j = 0$ for $j \geq 3$, then $X \in \mathbb{R}$ is normally distributed with expected value $\mu = \kappa_1$ and variance $\sigma^2 = \kappa_2$. The moment generating function is very practical when handling sums of independent random variables. If X and Y are independent with existing moment generating function, then $M_{X+Y}(t) = M_X(t) M_Y(t)$ (and vice versa). For the cumulant generating function this means $K_{X+Y}(t) = K_X(t) + K_Y(t)$. For limit theorems, the following result is useful: Let X_n be a sequence of random variables with moment generating functions $M_{X_n}(t)$ which converge to the moment generating function $M_X(t)$ of a random variable X. Then X_n converges to X in distribution. This together with the additivity property of the cumulant generating function can be used for a simple proof of the central limit theorem (see ►Central Limit Theorems).

Moment Generating Function. Table 1 $M_X(t)$ for some important distributions

Distribution	$M_X(t)$
Binomial with n trials, success probability $p = 1 - q$	$[q + pe^t]^n$
Geometric distribution with success probability $p = 1 - q$	$pe^t (1 - qe^t)^{-1}$
Poisson with expected value λ	$\exp[\lambda (e^t - 1)]$
Uniform on $[a, b]$	$t^{-1}(b-a)^{-1}\left(e^{tb} - e^{ta}\right)$
Normal $N(\mu, \sigma^2)$	$\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$
Multivariate Normal $N(\mu, \Sigma)$	$\exp\left(\mu' t + \frac{1}{2} t' \Sigma t\right)$
Chi-square χ_k^2	$(1 - 2t)^{-\frac{k}{2}}$
Exponential with expected value λ^{-1}	$\left(1 - t\lambda^{-1}\right)^{-1}$
Cauchy distribution	not defined

The empirical counterparts of M_X, K_X and ϕ_X, defined by

$$m_n(t) = n^{-1} \sum_{i=1}^{n} \exp(tX_i), \tag{7}$$

$k_n(t) = \log m_n(t)$ and $\varphi_n(t) = \log m_n(it)$, are often useful for statistical inference. For instance, testing the null hypothesis that X and Y are independent can be done by testing $M_{X+Y} \equiv M_X M_Y$ or $\varphi_{X+Y} \equiv \varphi_X \varphi_Y$ (see e.g., Csörgő 1985; Feuerverger 1987). Testing normality of a random sample $X_1, ..., X_n$ is the same as testing $H_o : \partial^3/\partial t^3 K_X(t) \equiv 0$ (see Ghosh 1996; Fang et al. 1998). For further applications of empirical moment and cumulant generating functions see e.g., Csörgő (1982, 1986), Epps et al. (1982),

Feuerverger (1989), Feuerverger and McDunnough (1984), Knight and Satchell (1997), Ghosh and Beran (2000, 2006).

Cross References

►Bivariate Distributions
►Financial Return Distributions
►Random Variable
►Statistical Distributions: An Overview
►Univariate Discrete Distributions: An Overview

References and Further Reading

Csörgő S (1982) The empirical moment generating function. In: Gnedenko BV, Puri ML, Vincze I (eds) Nonparametric statistical inference: Coll Math Soc J Bolyai, 32, Amsterdam, North-Holland, pp 139–150
Csörgő S (1985) Testing for independence by the empirical characteristic function. J Multivariate Anal 16(3):290–299
Csörgő S (1986) Testing for normality in arbitrary dimension. Ann Stat 14:708–723
Epps TW, Singleton KJ, Pulley LB (1982) A test of separate families of distributions based on the empirical moment generating function. Biometrika 69:391–399
Fang K-T, Li R-Z, Liang J-J (1998) A multivariate version of Ghosh's T3-plot to detect non-multinormality. Comput Stat Data Anal 28:371–386
Feuerverger A (1987) On some ECF procedures for testing independence. In: MacNeill IB, Umphrey GJ, Festschrift J (eds) Time series and econometric modeling, Reidel, New York, pp 189–206
Feuerverger A (1989) On the empirical saddlepoint approximation. Biometrika 76(3):457–464
Feuerverger A, McDunnough P (1984) On statistical transform methods and their efficiency. Can J Stat 12:303–317
Ghosh S (1996) A new graphical tool to detect non-normality. J Roy Stat Soc B 58:691–702
Ghosh S, Beran J (2000) The two-sample T3 test – a graphical method for comparing two distributions. J Comput Graph Stat 9(1):167–179
Ghosh S, Beran J (2006) On estimating the cumulant generating function of linear processes. Ann Inst Stat Math 58:53–71
Knight JL, Satchell SE (1997) The cumulant generating function estimation method: implementation and asymptotic efficiency. Economet Theor 13(2):170–184
Severini TA (2005) Elements of distribution theory. Cambridge University Press, Cambridge

Monte Carlo Methods in Statistics

Christian Robert
Professor of Statistics
Université Paris-Dauphine, CEREMADE, Paris, France

Monte Carlo methods are now an essential part of the statistician's toolbox, to the point of being more familiar to graduate students than the measure theoretic notions upon which they are based! We recall in this note some of the advances made in the design of Monte Carlo techniques towards their use in Statistics, referring to Robert and Casella (2004, 2010) for an in-depth coverage.

The Basic Monte Carlo Principle and Its Extensions

The most appealing feature of Monte Carlo methods [for a statistician] is that they rely on sampling and on probability notions, which are the bread and butter of our profession. Indeed, the foundation of Monte Carlo approximations is identical to the validation of empirical moment estimators in that the average

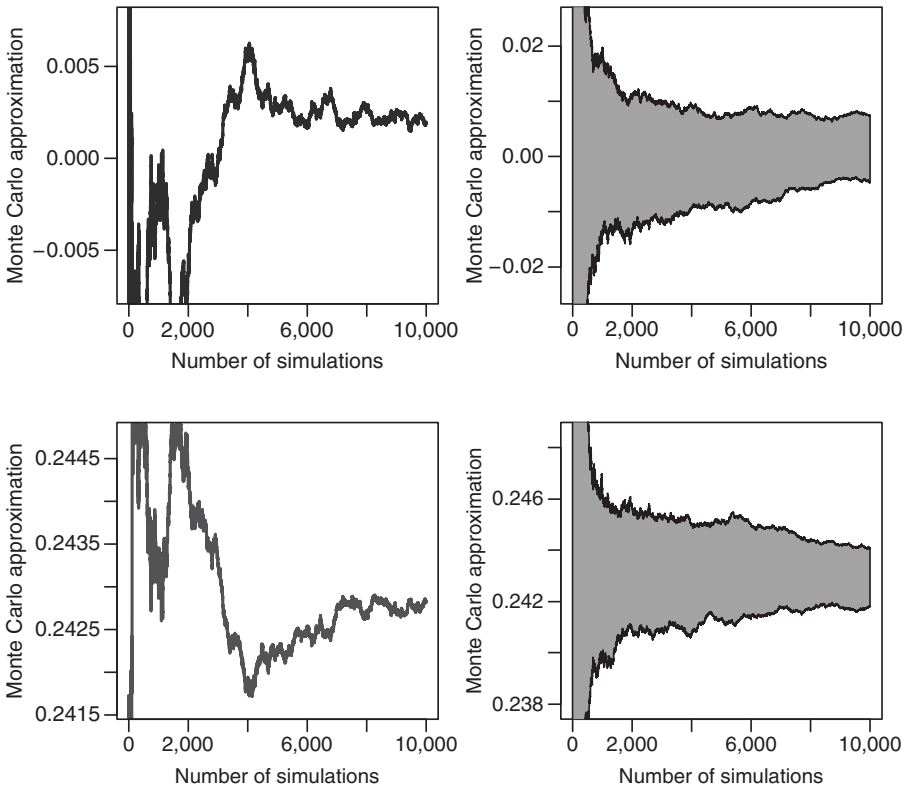
$$\frac{1}{T}\sum_{t=1}^{T} h(x_t), \quad x_t \sim f(x), \tag{1}$$

is converging to the expectation $\mathbb{E}_f[h(X)]$ when T goes to infinity. Furthermore, the precision of this approximation is exactly of the same kind as the precision of a statistical estimate, in that it usually evolves as $O(\sqrt{T})$. Therefore, once a sample $x_1, \ldots, x_T$ is produced according to a distribution density f, all standard statistical tools, including bootstrap (see ►Bootstrap Methods), apply to this sample (with the further appeal that more data points can be produced if deemed necessary). As illustrated by Fig. 1, the variability due to a single Monte Carlo experiment must be accounted for, when drawing conclusions about its output and evaluations of the overall variability of the sequence of approximations are provided in Kendall et al. (2007). But the ease with which such methods are analyzed and the systematic resort to statistical intuition explain in part why Monte Carlo methods are privileged over numerical methods.

The representation of integrals as expectations $\mathbb{E}_f[h(X)]$ is far from unique and there exist therefore many possible approaches to the above approximation. This range of choices corresponds to the importance sampling strategies (Rubinstein 1981) in Monte Carlo, based on the obvious identity

$$\mathbb{E}_f[h(X)] = \mathbb{E}_g[h(X)f(X)/g(X)]$$

provided the support of the density g includes the support of f. Some choices of g may however lead to appallingly poor performances of the resulting Monte Carlo estimates, in that the variance of the resulting empirical average may be infinite, a danger worth highlighting since often neglected while having a major impact on the quality of the approximations. From a statistical perspective, there exist some natural choices for the importance function



Monte Carlo Methods in Statistics. Fig. 1 Monte Carlo evaluation (1) of the expectation $\mathbb{E}[X^3/(1+X^2+X^4)]$ as a function of the number of simulation when $X \sim \mathcal{N}(\mu,1)$ using (*left*) one simulation run and (*right*) 100 independent runs for (*top*) $\mu = 0$ and (*bottom*) $\mu = 2.5$

g, based on Fisher information and analytical approximations to the likelihood function like the Laplace approximation (Rue et al. 2008), even though it is more robust to replace the normal distribution in the Laplace approximation with a t distribution. The special case of Bayes factors (Andrieu et al. 2005) (Andrieu et al. 2005)

$$B_{01}(x) = \int_\Theta f(x|\theta)\pi_0(\theta)\mathrm{d}\theta \Big/ \int_\Theta f(x|\theta)\pi_1(\theta)\mathrm{d}\theta,$$

which drive Bayesian testing and model choice, and of their approximation has led to a specific class of importance sampling techniques known as *bridge sampling* (Chen et al. 2000) where the optimal importance function is made of a mixture of the posterior distributions corresponding to both models (assuming both parameter spaces can be mapped into the same Θ). We want to stress here that an alternative approximation of marginal likelihoods relying on the use of *harmonic means* (Gelfand and Dey 1994; Newton and Raftery 1994) and of direct simulations from a posterior density has repeatedly been used in the literature, despite often suffering from infinite variance (and thus numerical instability). Another potentially very efficient approximation of Bayes factors is provided by Chib's (1995) representation, based on parametric estimates to the posterior distribution.

MCMC Methods

Markov chain Monte Carlo (MCMC) methods (see ►Markov Chain Monte Carlo) have been proposed many years (Metropolis et al. 1953) before their impact in Statistics was truly felt. However, once Gelfand and Smith (1990) stressed the ultimate feasibility of producing a Markov chain (see ►Markov Chains) with a given stationary distribution f, either via a Gibbs sampler that simulates each conditional distribution of f in its turn, or via a Metropolis–Hastings algorithm based on a proposal $q(y|x)$ with acceptance probability [for a move from x to y]

$$\min\left\{1, f(y)q(x|y)\big/f(x)q(y|x)\right\},$$

then the spectrum of manageable models grew immensely and almost instantaneously.

Due to parallel developments at the time on graphical and hierarchical Bayesian models, like generalized linear mixed models (Zeger and Karim 1991), the wealth of multivariate models with available conditional distributions (and hence the potential of implementing the Gibbs sampler) was far from negligible, especially when the availability of latent variables became quasi universal due to the slice sampling representations (Damien et al. 1999; Neal 2003). (Although the adoption of Gibbs samplers has primarily taken place within ►Bayesian statistics, there is nothing that prevents an artificial augmentation of the data through such techniques.)

For instance, if the density $f(x) \propto \exp(-x^2/2)/(1 + x^2 + x^4)$ is known up to a normalizing constant, f is the marginal (in x) of the joint distribution $g(x,u) \propto \exp(-x^2/2)\mathbb{I}(u(1 + x^2 + x^4) \leq 1)$, when u is restricted to $(0,1)$. The corresponding slice sampler then consists in simulating

$$U|X = x \sim \mathcal{U}(0, 1/(1 + x^2 + x^4))$$

and

$$X|U = u \sim \mathcal{N}(0,1)\mathbb{I}(1 + x^2 + x^4 \leq 1/u),$$

the later being a truncated normal distribution. As shown by Fig. 2, the outcome of the resulting Gibbs sampler perfectly fits the target density, while the convergence of the expectation of X^3 under f has a behavior quite comparable with the iid setting.

While the Gibbs sampler first appears as *the* natural solution to solve a simulation problem in complex models if only because it stems from the true target f, as exhibited by the widespread use of BUGS (Lunn et al. 2000), which mostly focus on this approach, the infinite variations offered by the Metropolis–Hastings schemes offer much more efficient solutions when the proposal $q(y|x)$ is appropriately chosen. The basic choice of a random walk proposal (see ►Random Walk) $q(y|x)$ being then a normal density centered in x) can be improved by exploiting some features of the target as in Langevin algorithms (see Andrieu et al. 2005 Sect. 7.8.5) and Hamiltonian or hybrid alternatives (Duane et al. 1987; Neal 1999) that build upon gradients. More recent proposals include particle learning about the target and sequential improvement of the proposal (Douc et al. 2007; Rosenthal 2007; Andrieu et al. 2010). Fig. 3 reproduces Fig. 2 for a random walk Metropolis–Hastings algorithm whose scale is calibrated towards an acceptance rate of 0.5. The range of the convergence paths is clearly wider than for the Gibbs sampler, but the fact that this is a generic algorithm applying to any target (instead of a specialized version as for the Gibbs sampler) must be borne in mind.

Another major improvement generated by a statistical imperative is the development of variable dimension generators that stemmed from Bayesian model choice requirements, the most important example being the reversible jump algorithm in Green (1995) which had a significant impact on the study of graphical models (Brooks et al. 2003).

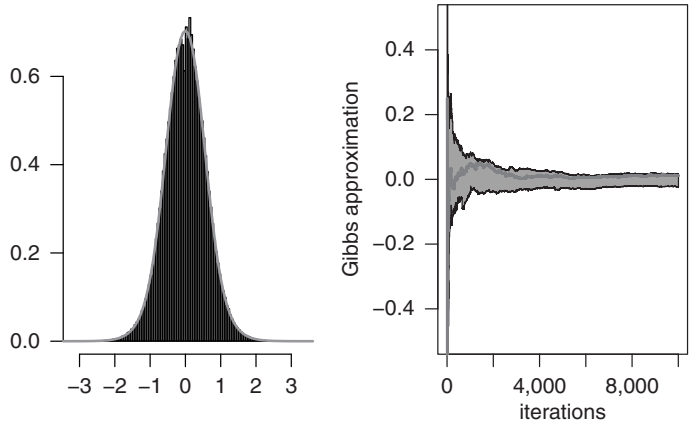
Some Uses of Monte Carlo in Statistics

The impact of Monte Carlo methods on Statistics has not been truly felt until the early 1980s, with the publication of Rubinstein (1981) and Ripley (1987), but Monte Carlo methods have now become invaluable in Statistics because they allow to address optimization, integration and exploration problems that would otherwise be unreachable. For instance, the calibration of many tests and the derivation of their acceptance regions can only be achieved by simulation techniques. While integration issues are often linked with the Bayesian approach – since Bayes estimates are posterior expectations like

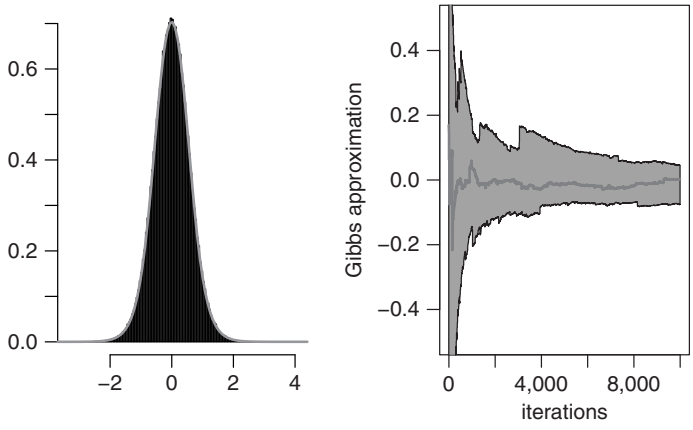
$$\int h(\theta)\pi(\theta|x)\,\mathrm{d}\theta$$

and Bayes tests also involve integration, as mentioned earlier with the Bayes factors, and optimization difficulties with the likelihood perspective, this classification is by no way tight – as for instance when likelihoods involve unmanageable integrals – and all fields of Statistics, from design to econometrics, from genomics to psychometry and environmics, have now to rely on Monte Carlo approximations. A whole new range of statistical methodologies have entirely integrated the simulation aspects. Examples include the bootstrap methodology (Efron 1982), where multilevel resampling is not conceivable without a computer, indirect inference (Gouriéroux et al. 1993), which construct a pseudo-likelihood from simulations, MCEM (Cappé and Moulines 2009), where the E-step of the EM algorithm is replaced with a Monte Carlo approximation, or the more recent approximated Bayesian computation (ABC) used in population genetics (Beaumont et al. 2002), where the likelihood is not manageable but the underlying model can be simulated from.

In the past fifteen years, the collection of real problems that Statistics can [afford to] handle has truly undergone a quantum leap. Monte Carlo methods and in particular MCMC techniques have forever changed the emphasis from "closed form" solutions to algorithmic ones, expanded our impact to solving "real" applied problems while convincing scientists from other fields that statistical solutions were indeed available, and led us into a world



Monte Carlo Methods in Statistics. Fig. 2 (*left*) Gibbs sampling approximation to the distribution $f(x) \propto \exp(-x^2/2)/(1+x^2+x^4)$ against the true density; (*right*) range of convergence of the approximation to $\mathbb{E}_f[X^3] = 0$ against the number of iterations using 100 independent runs of the Gibbs sampler, along with a single Gibbs run



Monte Carlo Methods in Statistics. Fig. 3 (*left*) Random walk Metropolis–Hastings sampling approximation to the distribution $f(x) \propto \exp(-x^2/2)/(1+x^2+x^4)$ against the true density for a scale of 1.2 corresponding to an acceptance rate of 0.5; (*right*) range of convergence of the approximation to $\mathbb{E}_f[X^3] = 0$ against the number of iterations using 100 independent runs of the Metropolis–Hastings sampler, along with a single Metropolis–Hastings run

where "exact" may mean "simulated." The size of the data sets and of the models currently handled thanks to those tools, for example in genomics or in climatology, is something that could not have been conceived 60 years ago, when Ulam and von Neumann invented the Monte Carlo method.

Acknowledgments

Supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75775 Paris) through the 2009–2012 project ANR-08-BLAN-0218 Big'MC. The author is grateful to Jean-Michel Marin for helpful comments.

About the Author

Dr. Christian P. Robert is Professor of Statistics in the Department of Mathematics, Université Paris-Dauphine, and Head of the Statistics Laboratory, Centre de Recherche en Economie et Statistique, Institut National de la Statistique et des Études Économiques (INSEE), Paris, France. He has authored and co-authored more than 130 papers and 9 books, including *The Bayesian Choice* (Springer Verlag, 2001), which received the DeGroot Prize in 2004, *Monte Carlo Statistical Methods* with George Casella (Springer Verlag, 2004), *Bayesian Core* with Jean-Michel Marin (Springer Verlag, 2007), and *Introducing Monte Carlo Methods with R* with George Casella (Springer Verlag, 2009). He was President of the International

Society for Bayesian Analysis (ISBA) in 2008. He is an IMS Fellow (1996) and an Elected member of the Royal Statistical Society (1998). Professor Robert has been the Editor of the *Journal of the Royal Statistical Society Series* (2005–2009) and an Associate Editor for *Annals of Statistics* (1998–2006), the *Journal of the American Statistical Society* (1996–1999 and 2005–2008), *Annals of the Institute of Statistical Mathematics* (2003–2005), *Statistical Science* (2000–2004), *Bayesian Analysis* (2003–2005), *TEST* (1994–1997 and 2000–2003), and *Sankhya* (1999–2002 and 2010).

Cross References

►Bootstrap Methods
►Computational Statistics
►Copulas: Distribution Functions and Simulation
►Entropy and Cross Entropy as Diversity and Distance Measures
►Frequentist Hypothesis Testing: A Defense
►Markov Chain Monte Carlo
►Multivariate Statistical Simulation
►Non-Uniform Random Variate Generations
►Numerical Integration
►Sensitivity Analysis
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Modeling of Financial Markets
►Uniform Distribution in Statistics
►Uniform Random Number Generators

References and Further Reading

Andrieu C, Doucet A, Holenstein R (2010) Particle Markov chain Monte Carlo (with discussion). J Roy Stat Soc B 72:269–342
Beaumont M, Zhang W, Balding D (2002) Approximate Bayesian computation in population genetics. Genetics 162:2025–2035
Brooks S, Giudici P, Roberts G (2003) Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). J Roy Stat Soc B 65:3–55
Cappé O, Moulines E (2009) On-line expectation-maximization algorithm for latent data models. J Roy Stat Soc B, 71(3):593–613
Chen M, Shao Q, Ibrahim J (2000) Monte Carlo methods in Bayesian computation. Springer, New York
Chib S (1995) Marginal likelihood from the Gibbs output. J Am Stat Assoc 90:1313–1321
Damien P, Wakefield J, Walker S (1999) Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. J Roy Stat Soc B 61:331–344
Douc R, Guillin A, Marin J-M, Robert C (2007) Convergence of adaptive mixtures of importance sampling schemes. Ann Stat 35(1):420–448
Duane S, Kennedy AD, Pendleton BJ, Roweth D (1987) Hybrid Monte Carlo. Phys Lett B 195:216–222
Efron B (1982) The Jacknife, the Bootstrap and other resampling plans, vol 38. SIAM, Philadelphia
Gelfand A, Dey D (1994) Bayesian model choice: asymptotics and exact calculations. J Roy Stat Soc B 56:501–514
Gelfand A, Smith A (1990) Sampling based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409
Gouriéroux C, Monfort A, Renault E (1993) Indirect inference. J Appl Econom 8:85–118
Green P (1995) Reversible jump MCMC computation and Bayesian model determination. Biometrika 82:711–732
Kendall W, Marin J-M, Robert C (2007) Confidence bands for Brownian motion and applications to Monte Carlo simulations. Stat Comput 17:1–10
Lunn D, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325–337
Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1092
Neal R (1999) Bayesian learning for neural networks, vol 118. Springer, New York
Neal R (2003) Slice sampling (with discussion). Ann Statist 31: 705–767
Newton M, Raftery A (1994) Approximate Bayesian inference by the weighted likelihood boostrap (with discussion). J Roy Stat Soc B 56:1–48
Ripley B (1987) Stochastic simulation. Wiley, New York
Rober C, Casella G (2004) Monte Carlo statistical methods. 2nd ed. Springer-Verlag, New York
Robert C, Casella G (2010) Introducing Monte Carlo methods with *R*. Springer, New York
Rosenthal J (2007) AMCM: an *R* interface for adaptive MCMC. Comput Stat Data Anal 51:5467–5470
Rubinstein R (1981) Simulation and the Monte Carlo method. Wiley, New York
Rue H, Martino S, Chopin N (2008) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). J Roy Stat Soc B 71(2):319–392
Zeger S, Karim R (1991) Generalized linear models with random effects; a Gibbs sampling approach. J Am Stat Assoc 86:79–86

Monty Hall Problem : Solution

Richard D. Gill
Professor, Faculty of Science, President of the Dutch Society for Statistics and Operations Research
Leiden University, Leiden, Netherlands

Introduction

The *Three Doors Problem*, or *Monty Hall Problem*, is familiar to statisticians as a paradox in elementary probability theory often found in elementary probability texts (especially in their exercises sections). In that context it is usually meant to be solved by careful (and elementary) application of ►Bayes' theorem. However, in different forms, it is much discussed and argued about and written

about by psychologists, game-theorists and mathematical economists, educationalists, journalists, lay persons, blog-writers, wikipedia editors.

In this article I will briefly survey the history of the problem and some of the approaches to it which have been proposed. My take-home message to you, dear reader, is that one should distinguish two levels to the problem.

There is an informally stated problem which you could pose to a friend at a party; and there are many concrete *versions* or *realizations* of the problem, which are actually the result of mathematical or probabilistic or statistical *modeling*. This modeling often involves adding supplementary assumptions chosen to make the problem well posed in the terms of the modeler. The modeler finds those assumptions perfectly natural. His or her students are supposed to guess those assumptions from various key words (like: "indistinguishable," "unknown") strategically placed in the problem re-statement. Teaching statistics is often about teaching the students to read the teacher's mind. Mathematical (probabilistic, statistical) modeling is, unfortunately, often solution driven rather than problem driven.

The very same criticism can, and should, be leveled at this very article! By cunningly presenting the history of *The Three Doors Problem* from my rather special point of view, I have engineered complex reality so as to convert the *Three Doors Problem* into an illustration of my personal Philosophy of Science, my Philosophy of Statistics.

This means that I have re-engineered the *Three Doors Problem* into an example of the point of view that Applied Statisticians should always be wary of the lure of *Solution-driven Science*. Applied Statisticians are trained to know Applied Statistics, and are trained to know how to convert real world problems into statistics problems. That is fine. But the best Applied Statisticians know that Applied Statistics is not the only game in town. Applied Statisticians are merely some particular kind of Scientists. They know lots about modeling uncertainty, and about learning from more or less random data, but probably not much about anything else. The Real Scientist knows that there is not a universal *disciplinary* approach to every problem. The *Real Statistical Scientist* modestly and persuasively and realistically offers what his or her discipline has to offer in synergy with others.

To summarize, we must distinguish between:

(0) the *Three-Doors-Problem Problem* [sic], which is to make sense of some real world question of a real person.
(1) a large number of solutions to this *meta*-problem, i.e., the many Three-Doors-Problem *Problems*, which are competing mathematizations of the meta-problem (0).

Each of the solutions at level (1) can well have a number of different solutions: nice ones and ugly ones; correct ones and incorrect ones. In this article, I will discuss three level (1) solutions, i.e., three different Monty Hall problems; and try to give three short correct and attractive solutions.

Now read on. Be critical, use your intellect, don't believe anything on authority, and certainly not on mine. Especially, don't forget the problem at meta-level (−1), not listed above.

C'est la vie.

Starting Point

I shall start not with the historical roots of the problem, but with the question which made the Three Doors Problem famous, even reaching the front page of the *New York Times*.

Marilyn vos Savant (a woman allegedly with the highest IQ in the world) posed the *Three Door Problem* or *Monty Hall Problem* in her "Ask Marilyn" column in *Parade* magazine (September 1990:16), as posed to her by a correspondent, a Mr. Craig Whitaker. It was, quoting vos Savant literally, the following:

> *Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?*

Apparently, the problem refers to a real American TV quiz-show, with a real presenter, called Monty Hall.

The literature on the Monty Hall Problem is enormous. At the end of this article I shall simply list two references which for me have been especially valuable: a paper by Jeff Rosenthal (2008) and a book by Jason Rosenhouse (2009). The latter has a huge reference list and discusses the pre- and post-history of vos Savant's problem.

Briefly regarding the pre-history, one may trace the problem back through a 1975 letter to the editor in the journal *The American Statistician* by biostatistician Steve Selkin, to a problem called *The Three Prisoners Problem* posed by Stephen Gardner in his Mathematical Games column in *Scientific American* in 1959, and from there back to *Bertrand's Box Problem* in his 1889 text on Probability Theory. The internet encyclopedia `wikipedia.org` discussion pages (in many languages) are a fabulous though every-changing resource. Almost everything that I write here was learnt from those pages.

Despite making homage here to the two cited authors Rosenthal (2008) and Rosenhouse (2009) for their wonderful work, I emphasize that I strongly disagree with

both Rosenhouse ("the canonical problem") and Rosenthal ("the original problem") on what the essential Monty Hall problem is. I am more angry with certain other authors, who will remain nameless but for the sake of argument I'll just call Morgan et al. for unilaterally declaring in *The American Statistician* in 1981 *their* Monty Hall problem to be the only possible sensible problem, for calling everyone who solved different problems stupid, and for getting an incorrect theorem (I refer to their result about the situation when we do not know the quiz-master's probability of opening a particular door when he has a choice, and put a uniform prior on this probability.) published in the peer-reviewed literature.

Deciding unilaterally (Rosenhouse 2009) that a certain formulation is *canonical* is asking for a schism and for excommunication. Calling a particular version *original* (Rosenthal 2008) is asking for a historical contradiction. In view of the pre-history of the problem, the notion is not well defined. Monty Hall is part of folk-culture, culture is alive, the Monty Hall problem is not owned by a particular kind of mathematician who looks at such a problem from a particular point of view, and who adds for them "natural" extra assumptions which merely have the role of allowing their solution to work. Presenting any "canonical" or "original" Monty Hall problem together with a solution, is an example of *solution driven science* – you have learnt a clever trick and want to show that it solves lots of problems.

Three Monty Hall Problems

I will concentrate on three different particular Monty Hall problems. One of them (Q-0) is simply to answer the question literally posed by Marilyn vos Savant, "would you switch?". The other two (Q-1, Q-2) are popular mathematizations, particularly popular among experts or teachers of elementary probability theory: one asks for the unconditional probability that "always switching" would gets the car, the other asks for the conditional probability given the choices made so far. Here they are:

Q-0: Marilyn vos Savant's (or Craig Whitaker's) question "*Is it to your advantage to switch?*"

Q-1: A mathematician's question "*What is the unconditional probability that switching gives the car?*"

Q-2: A mathematician's question "*What is the conditional probability that switching gives the car, given everything so far?*"

The free, and freely editable, internet encyclopedia Wikipedia is the scene of a furious debate as to which mathematization Q-1 or Q-2 is the right starting point for answering the verbal question Q-0 (to be honest, many of the actors claim another "original" question as *the* original question). Alongside that, there is a furious debate as to which supplementary conditions are obviously implicitly being made. For each protagonist in the debate, those are the assumptions which ensure that his or her question has a unique and nice answer. My own humble opinion is "neither Q-1 nor Q-2, though the unconditional approach comes closer." I prefer Q-0, and I prefer to see it as a question of *game theory* for which, to my mind, [almost] no supplementary conditions need to be made.

Here I admit that I will suppose that the player knows game-theory and came to the quiz-show prepared. I will also suppose that the player wants to get the Cadillac while Monty Hall, the quizmaster, wants to keep it.

My analysis below of both problems Q-1 and Q-2 yields the good answer "2/3" under minimal assumptions, and almost without computation or algebraic manipulation. I will use Israeli (formerly Soviet Union) mathematician Boris Tsirelson's proposal on Wikipedia talk pages to use symmetry to deduce the conditional probability from the unconditional one. (Boris graciously gave me permission to cite him here, but this should not be interpreted to mean that anything written here also has his approval).

You, the reader, may well prefer a calculation using Bayes' theorem, or a calculation using the definition of conditional probability; I think this is a matter of taste.

I finally use a game-theoretic point of view, and von Neumann's minimax theorem, to answer the question Q-0 posed by Marilyn vos Savant, on the assumptions just stated.

Let the three doors be numbered in advance 1, 2, and 3. I add the universally agreed (and historically correct) additional assumptions: Monty Hall knows in advance where the car is hidden, Monty Hall always opens a door revealing a goat.

Introduce four random variables taking values in the set of door-numbers $\{1, 2, 3\}$:

C: the quiz-team hides the Car (a Cadillac) behind door C,
P: the Player chooses door P,
Q: the Quizmaster (Monty Hall) opens door Q,
S: Monty Hall asks the player if she'ld like to Switch to door S.

Because of the standard story of the Monty Hall show, we certainly have:

$Q \neq P$, the quizmaster *always* opens a door different to the player's first choice,
$Q \neq C$, opening that door *always* reveals a goat,
$S \neq P$, the player is *always* invited to switch to another door,
$S \neq Q$, *no* player wants to go home with a goat.

It does not matter for the subsequent mathematical analysis whether probabilities are subjective (Bayesian) or objective (frequentist); nor does it matter whose probabilities they are supposed to be, at what stage of the game. Some writers think of the player's initial choice as fixed. For them, P is degenerate.

I simply merely down some mathematical assumptions and deduce mathematical consequences of them.

Solution to Q-1: Unconditional Chance That Switching Wins

By the rules of the game and the definition of S, if $P \neq C$ then $S = C$, and vice-versa. A "switcher" would win the car if and only if a "stayer" would lose it. Therefore:

If $\Pr(P = C) = 1/3$ *then* $\Pr(S = C) = 2/3$, *since the two events are complementary.*

Solution to Q-2: Probability Car is Behind Door 2 Given You Chose Door 1, Monty Hall Opened 3

First of all, suppose that P and C are uniform and independent, and that given (P, C), suppose that Q is uniform on its possible values (unequal to those of P and of C). Let S be defined as before, as the third door-number different from P and Q. The joint law of C, P, Q, S is by this definition invariant under renumberings of the three doors. Hence $\Pr(S = C|P = x, Q = y)$ is the same for all $x \neq y$. By the law of total probability, $\Pr(S = C)$ (which is equal to 2/3 by our solution to Q-1) is equal to the weighted average of all $\Pr(S = C|P = x, Q = y), x \neq y \in \{1, 2, 3\}$. Since the latter are all equal, all these six conditional probabilities are equal to their average 2/3.

Conditioning on $P = x$, say, and letting y and y' denote the remaining two door numbers, we find the following corollary:

Now take the door chosen by the player as fixed, $P \equiv 1$, say. We are to compute $\Pr(S = C|Q = 3)$. *Assume that all doors are equally likely to hide the car and assume that the quizmaster chooses completely at random when he has a choice.* Without loss of generality we may as well pretend that P was chosen in advance completely at random. Now we have embedded our problem into the situation just solved, where P and C are uniform and independent.

▶ *If* $P \equiv 1$ *is fixed,* C *is uniform, and* Q *is symmetric, then "switching gives car" is independent of quizmaster's choice, hence*

$$\Pr(S = C|Q = 3) = \Pr(S = C|Q = 2') = \Pr(S = C) = 2/3.$$

Some readers may prefer a direct calculation. Using Bayes' theorem in the form "posterior odds equal prior odds times likelihoods" is a particularly efficient way to do this. The probabilities and conditional probabilities below are all conditional on $P = 1$, or if your prefer with $P \equiv 1$.

We have uniform prior odds

$$\Pr(C = 1) : \Pr(C = 2) : \Pr(C = 3) = 1 : 1 : 1.$$

The likelihood for C, the location of the car, given data $Q = 3$, is (proportional to) the discrete density function of Q given C (and P)

$$\Pr(Q = 3|C = 1) : \Pr(Q = 3|C = 2) : \Pr(Q = 3|C = 3) = \frac{1}{2} : 1 : 0.$$

The posterior odds are therefore proportional to the likelihood. It follows that the posterior probabilities are

$$\Pr(Q = 3|C = 1) = \frac{1}{3}, \quad \Pr(Q = 3|C = 2) = \frac{2}{3},$$
$$\Pr(Q = 3|C = 3) = 0.$$

Answer to Marilyn Vos Savant's Q-0: Should *You* Switch Doors?

Yes. Recall, *You only know that Monty Hall always opens a door revealing a goat.* You didn't know what strategy the quiz-team and quizmaster were going to use for their choices of the distribution of C and the distribution of Q given P and C, so naturally (since you know elementary Game Theory) you had picked your door uniformly at random. Your strategy of choosing C uniformly at random guarantees that $\Pr(C = P) = 1/3$ and hence that $\Pr(S = C) = 2/3$.

It was easy for you to find out that this combined strategy, which I'll call "symmetrize and switch," is your so-called minimax strategy.

On the one hand, "symmetrize and switch" guarantees you a 2/3 (unconditional) chance of winning the car, whatever strategy used by the quizmaster and his team.

On the other hand, if the quizmaster and his team use their "symmetric" strategy "hide the car uniformly at random and toss a fair coin to open a door if there is choice", then you cannot win the car with a *better* probability than 2/3.

The fact that your "symmetrize and switch" strategy gives you "at least" 2/3, while the quizmaster's "symmetry" strategy prevents you from doing better, proves that these are the respective minimax strategies, and 2/3 is the game-theoretic value of this two-party zero-sum game. (Minimax strategies and the accompanying "value" of the game exist by virtue of John von Neumann's (1929) minimax theorem for finite two-party zero-sum games).

There is not much point for you in worrying about your conditional probability of winning conditional on

your specific initial choice and the specific door opened by the quizmaster, say doors 1 and 3 respectively. You don't know this conditional probability anyway, since you don't know the strategy used by quiz-team and the quizmaster. (Even though you know probability theory and game theory, they maybe don't). However, it is maybe comforting to learn, by easy calculation, that if the car is hidden uniformly at random, then your conditional probability cannot be *smaller* than 1/2. So in that case at least, it certainly never *hurts* to switch door.

Discussion

Above I tried to give short clear mathematical solutions to three mathematical problems. Two of them were problems of elementary probability theory, the third is a problem of elementary game theory. As such, it involves not much more than elementary probability theory and the beautiful minimax theorem of John von Neumann (1928). That a finite two-party zero-sum game has a saddle-point, or in other words, that the two parties in such a game have matching minimax strategies (if ►randomization is allowed), is not obvious. It seems to me that probabilists ought to know more about game theory, since every ordinary non-mathematician who hears about the problem starts to wonder whether the quiz-master is trying to cheat the player, leading to an infinite regress: if I know that he knows that I know that....

I am told that the literature of mathematical economics and of game theory is full of Monty Hall examples, but no-one can give me a nice reference to a nice game-theoretic solution of the problem. Probably game-theorists like to keep their clever ideas to themselves, so as to make money from playing the game. Only losers write books explaining how the reader could make money from game theory.

It would certainly be interesting to investigate more complex game-theoretic versions of the problem. If we take Monty Hall as a separate player to the TV station, and note that TV ratings are probably helped if nice players win while annoying players lose, we leave elementary game theory and must learn the theory of Nash equilibria.

Then there is a sociological or historical question: who "owns" the Monty Hall problem? I think the answer is obvious: no-one. A beautiful mathematical paradox, once launched into the real world, lives it own life, it evolves, it is re-evaluated by generation after generation. This point of view actually makes me believe that Question 0: *would you switch* is the right question, and no further information should be given beyond the fact that you know that the quizmaster knows where the car is hidden, and always opens a door exhibiting a goat. Question 0 is a question you can ask a non-mathematician at a party, and if they have not heard of the problem before, they'll give the wrong answer (or rather, one of the two wrong answers: *no* because nothing is changed, or *it doesn't matter* because it's now 50–50). My mother, who was one of Turing's computers at Bletchley Park during the war, but who had had almost no schooling and in particular never learnt any mathematics, is the only person I know who immediately said: *switch*, by immediate intuitive consideration of the 100-door variant of the problem. The problem is a *paradox* since you can next immediately convince anyone (except lawyers, as was shown by an experiment in Nijmegen), that their initial answer is wrong.

The mathematizations Questions 1 and 2 are not (in my humble opinion!) *the* Monty Hall problem; they are questions which probabilists might ask, anxious to show off Bayes' theorem or whatever. Some people intuitively try to answer Question 0 via Questions 1 and 2; that is natural, I do admit. And sometimes people become very confused when they realize that the answer to Question 2 can only be given its pretty answer "2/3" under further conditions. It is interesting how in the pedagogical mathematical literature, the further conditions are as it were held under your nose, e.g., by saying "three *identical* doors," or replacing Marilyn's "say, door 1" by the more emphatic "door 1."

It seems to me that adding into the question explicitly the remarks that the three doors are equally likely to hide the car, and that when the quizmaster has a choice he secretly tosses a fair coin to decide, convert this beautiful paradox into a probability puzzle with little appeal any more to non experts.

It also converts the problem into one version of the three prisoner's paradox. The three prisoners problem is isomorphic to the conditional probabilistic three doors problem. I always found it a bit silly and not very interesting, but possibly that problem too should be approached from a sophisticated game theoretic point of view.

By the way, Marilyn vos Savant's original question is semantically ambiguous, though this might not be noticed by a non-native English speaker. Are the mentioned door numbers, huge painted numbers on the front of the doors *a priori*, or are we just for convenience *naming* the doors by the choices of the actors in our game *a posteriori*. Marilyn stated in a later column in *Parade* that she had originally been thinking of the latter. However, her own offered solutions are not consistent with a single unambiguous formulation. Probably she did not find the difference very interesting.

This little article contains nothing new, and only almost trivial mathematics. It is a plea for future generations to preserve the life of *The True Monty Hall paradox*, and not

let themselves be misled by probability purists who say "you *must* compute a conditional probability."

About the Author

Professor Gill has been selected as the 2010–2011 Distinguished Lorentz Fellow by the Netherlands Institute for Advanced Study in Humanities and Social Sciences. He is a member of the Royal Netherlands Academy of Arts and Sciences.

Cross References

►Bayes' Theorem
►Conditional Expectation and Probability
►Statistics and Gambling

References and Further Reading

Gill RD (2010) The one and only true Monty Hall problem. Submitted to *Statistica Neerlandica*. arXiv.org:1002.0651 [math.HO]

Rosenhouse J (2009) The Monty Hall problem. Oxford University Press, Oxford

Rosenthal JS (2008) Monty Hall, Monty Fall, Monty Crawl. Math Horizons September 2008:5–7. Reprint: http://probability.ca/jeff/writing/montyfall.pdf

Mood Test

Justice I. Odiase[1], Sunday M. Ogbonmwan[2]
[1]University of Benin, Benin City, Nigeria
[2]Professor and Dean, The Faculty of Physical Sciences University of Benin, Benin City, Nigeria

In 1954, A.M. Mood developed the square rank test for dispersion known as Mood test. It is based on the sum of squared deviations of the ranks of one sample from the mean rank of the combined samples. The null hypothesis is that there is no difference in spread against the alternative hypothesis that there is some difference. The Mood test assumes that location remains the same. It is assumed that differences in scale do not cause a difference in location. The samples are assumed to be drawn from continuous distributions.

In two-sample scale tests, the population distributions are usually assumed to have the same location with different spreads. However, Neave and Worthington (1988) cautioned that tests for difference in scale could be severely impaired if there is a difference in location as well.

In a two-sample problem composed of $X = \{x_1, x_2, \ldots, x_m\}$ with distribution $F(X)$ and $Y = \{y_1, y_2, \ldots, y_n\}$ with distribution $G(Y)$, arrange the combined samples in ascending order of magnitude and rank all the $N = m + n$ observations from 1 (smallest) to N (largest). Let W be the sum of squares of the deviations of one of the samples' (say X) ranks from the mean rank of the combined samples,

$$W = \sum_{i=1}^{m} \left(r_i - \frac{m+n+1}{2} \right)^2,$$

where r_i is the rank of the i^{th} X observation. The table of exact critical values can be found in Odiase and Ogbonmwan (2008).

Under the null hypothesis ($F = G$), the layout of the ranks of the combined samples is composed of N independent and identically distributed random variables, and hence conditioned on the observed data set, the mean and variance of W are $m(N^2-1)/12$ and $mn(N+1)(N^2-4)/180$, respectively. The large sample Normal approximation of W is

$$\frac{W - \frac{m(N^2-1)}{12}}{\sqrt{\frac{mn(N+1)(N^2-4)}{180}}}.$$

The efficiency of the two-sample Mood test against the normal alternative to the null hypothesis is $\frac{15}{2\pi^2} \cong 76\%$.

A Monte Carlo study of several nonparametric test statistics to obtain the minimum sample size requirement for large sample approximation was carried out by Fahoome (2002). Adopting Bradley's (1978) liberal criterion of robustness, Fahoome (2002) recommends the asymptotic approximation of the Mood test when $\min(m, n) = 5$ for the level of significance $\alpha = 0.05$ and $\min(m, n) = 23$ for $\alpha = 0.01$. However, Odiase and Ogbonmwan (2008) generated the exact distribution of the Mood test statistics by the permutation method and therefore provided the table of exact critical values at different levels of significance.

The idea of a general method of obtaining an exact test of significance originated with Fisher (1935). The essential feature of the method is that all the distinct permutations of the observations are considered, with the property that each permutation is equally likely under the hypothesis to be tested.

About the Authors

Dr. Justice Ighodaro Odiase is a Senior Lecturer, Department of Mathematics, University of Benin, Nigeria. He is the Scientific Secretary of the Statistics Research Group (SRG), Department of Mathematics, University of Benin. He is a member of the Nigerian Statistical Association (NSA), International Association for Statistical Computing (IASC), and The Society for Imprecise Probability:

Theories and Applications (SIPTA). He has authored and coauthored more than 30 papers.

Sunday Martins Ogbonmwan is a Professor of Statistics, Department of Mathematics, University of Benin, Benin City, Nigeria. He is the President of the Statistics Research Group (SRG), Department of Mathematics, University of Benin. He was the Head of Department of Mathematics, University of Benin (2006–2009). He is currently the Dean of the Faculty of Physical Sciences, University of Benin. He is a member of the Institute of Mathematical Statistics (IMS). He is also a member of the Nigerian Statistical Association (NSA). He has authored and coauthored more than 50 papers. He was the Editor-in-Chief of the *Journal of the Nigerian Statistical Association* (JNSA (1990–1995)). Professor Ogbonmwan was an award winner in a competition organized by the International Statistical Institute for young statisticians in developing countries (Madrid, Spain, 1983).

Cross References

►Asymptotic Normality
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Tests for Homogeneity of Variance

References and Further Reading

Bradley JV (1978) Robustness? Br J Math Stat Psychol 31:144–152
Fahoome G (2002) Twenty nonparametric statistics and their large sample approximations. J Mod Appl Stat Meth 1:248–268
Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh
Mood AM (1954) On the asymptotic efficiency of certain nonparametric two-sample tests. Ann Math Stat 25:514–522
Neave HR, Worthington PL (1988) Distribution-free tests. Unwin Hyman, London
Odiase JI, Ogbonmwan SM (2008) Critical values for the Mood test of equality of dispersion. Missouri J Math Sci 20(1):40–52

Most Powerful Test

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

This notion plays a key role in testing statistical hypotheses. Testing is a two-decision statistical problem.

Case Study

A producer of hydraulic pumps applies plastic gaskets purchased from a deliverer. The gaskets are supplied in batches of 5,000. Since the cost of repairing a pump found to be faulty is far higher than the cost of the gasket itself, each batch is subject to testing. Not only the testing is costly but also any gasket used in the process is practically damaged. Thus the producer decides to verify 50 gaskets taken randomly from each batch.

Assume the deliverer promised that the fraction of defective gaskets would not exceed 5%. Suppose 4 defective gaskets were disclosed in a sample of size 50. Is this enough to reject the batch? The situation is illustrated by the following table

Batch\decision	Accept	Reject
Good	+	Type I Error
Bad	Type II Error	+

Since the decision is taken on the basis of a random variable (the number of defective gaskets), the quality of test may be expressed in terms of the probabilities of these two errors. We would like to minimize these probabilities simultaneously. However, any decrease of one of these probabilities causes increase of the second one. Consequences of these two errors should also be taken into consideration. Similarly as in law, one presumes that the tested hypothesis is true. Thus the probability of the error of the first type should be under control. Theory of testing statistical hypotheses, regarding these postulates, was formalized in 1933 by Neyman and Pearson.

Neyman-Pearson Theory

Let X be a random variable (or: random vector) taking values in a sample space $(\mathcal{X}, \mathcal{A})$ with a distribution P belonging to a class $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and let Θ_0 be a proper subset of Θ. We are interested in deciding, on the basis of observation X, whether $\theta \in \Theta_0$ (decision d_0) or not (decision d_2).

Any statement of the form $H : \theta \in \Theta_0$ is called a statistical hypothesis. We consider also the alternative hypothesis $K : \theta \notin \Theta_0$, i.e., $\theta \in \Theta \setminus \Theta_0$. A criterion of rejecting H (called a test) may be assigned by a *critical region* $S \subseteq \mathcal{X}$, according to the rule: reject H if $X \in S$ and accept otherwise.

When performing a test one may arrive at the correct decision, or one may commit one of two errors: rejecting H when it is true or accepting when it is false. The upper bound of the probability $P_\theta(d_0(X))$ for all $\theta \in \Theta_0$ is called

the *size* while the function $\beta(\theta) = P_\theta(d_0)$ for $\theta \in \Theta \setminus \Theta_0$ is called the power function of the test.

The general principle in Neyman-Pearson theory is to find such a procedure that maximizes $\beta(\theta)$ for all $\theta \in \Theta \backslash \Theta_0$ under assumption that $P_\theta(d_0(X)) \leq \alpha$ (significance level) for all $\theta \in \Theta_0$. Any such test (if exists) is called to be *uniformly most powerful* (UMP). The well known Neyman-Pearson fundamental lemma (see ▶Neyman-Pearson Lemma) states that for any two-element family of densities or mass probabilities $\{f_0, f_1\}$ such test always exists and it can be expressed by the likelihood ratio $r(x) = \frac{f_1(x)}{f_0(x)}$. In this case the power function β reduces to a scalar and the word *uniformly* is redundant.

It is worth to add that in the continuous case the size of the UMP test coincides with its significance level. However, it may not be true in the discrete case. The desired equality can be reached by considering the *randomized* decision rules represented by functions $\phi = \phi(x)$, taking values in the interval $[0,1]$ and interpreted as follows:

"If $X = x$ then reject H with probability $\phi(x)$
and accept it with probability $1 - \phi(x)$"

The size of the MP randomized test coincides with its significance level and its power may be greater than for the nonrandomized one. According to the Neyman-Pearson lemma, the randomized MP test has the form

$$\phi(x) = \begin{cases} 1, & \text{if} \quad p_1(x) > kp_0(x) \\ \gamma, & \text{if} \quad p_1(x) = kp_0(x \\ 0, & \text{if} \quad p_1(x) < kp_0(x \end{cases}$$

for some k induced by the significance level. If $\gamma = 0$ then it is non-randomized.

One-Side Hypothesis and Monotone Likelihood Ratio

In practical situations distribution of the observation vector depends on one or more parameters and we make use of composite hypotheses $\theta \in \Theta_0$ against $\theta \in \Theta \setminus \Theta_0$. Perhaps one of the simple situations of this type is testing one-side hypothesis $\theta \leq \theta_0$ or $\theta \geq \theta_0$ in a scalar parameter family of distributions.

We say that a family of densities $\{f_\theta : \theta \in \Theta\}$ has *monotone likelihood ratio* if there exists a statistic $T = t(X)$ such that for any $\theta < \theta'$ the ratio $\frac{f_{\theta'}(x)}{f_\theta(x)}$ is a monotone function of T. It appears that for testing a hypothesis $H : \theta \leq \theta_0$ against $K : \theta > \theta_0$ in such a family of densities there exists a UMP test of the form

$$\phi(x) = \begin{cases} 1 & \text{when } T(x) > C \\ \gamma & \text{when } T(x) = C \\ 0 & \text{when } T(x) < C. \end{cases}$$

An important class of families with monotone likelihood ratio are one-parameter exponential families with densities of type $f_\theta(x) = C(\theta)e^{Q(\theta)T(x)}h(x)$. In a discrete case with integer parameter instead the mononotonity condition it suffices to verify that the ratio $\frac{P_{k+1}(x)}{P_k(x)}$ is a monotone function of T for all k.

Example 1 (Testing expectation in a simple sample from normal distribution with known variance). Let $X_1, ..., X_n$ be independent and identically distributed. Random variables with distribution $N(\mu, \sigma^2)$, where σ^2 is known. Consider the hypothesis $H : \mu \leq \mu_0$ under the alternative $K : \mu > \mu_0$. The family of distributions has a monotone likelihood ratio with respect to the statistic $T = \sum_{i=1}^n X_i$. Therefore there exists a UMP test which rejects H if $\sum_{i=1}^n X_i$. is too large.

Example 2 (Statistical control theory). From a great number (N) of elements with an unknown number D of defective ones we draw without replacement a sample of size n. Then the potential number X of defective elements in the sample has the hypergeometric distribution

$$P_D(X = x) = \begin{cases} \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}, & \text{if } \max(0, n+D-N) < \\ & x < \min(n, D) \\ 0, & \text{otherwise.} \end{cases}.$$

One can verify that

$$\frac{P_{D+1}(x)}{P_D(x)} = \begin{cases} 0, & \text{if } x = n + D - N \\ \frac{D+1}{N-D}\frac{N-D-n+x}{D+1-x}, & \text{if } n + D + 1 - N \leq x \leq D \\ \infty & \text{if } x = D + 1 \end{cases}$$

is a monotone function of x. Therefore there exists a UMP test for the hypothesis $H : D \leq D_0$ against $K : D > D_0$, which rejects H if x is too large.

Invariant and Unbiased Tests

If distribution of the observation vector depends on several parameters, some of them may be out of our interest and play the role of nuisance parameters. Such a situation occurs, for instance, in testing linear hypotheses. In this case the class of all unbiased estimators is usually too large for handle. Then we may seek for a test with maximum power in a class of tests which are invariant with respect to some transformations of observations or their powers do not depend on the nuisance parameters. This is called the

most powerful invariant test. The class of tests under consideration may be also reduced by unbiasedness condition. A member of this class with maximum power is then called the most powerful unbiased test. The standard tests for linear hypotheses in a linear normal model are most powerful in each of these classes.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Asymptotic Relative Efficiency in Testing
►Frequentist Hypothesis Testing: A Defense
►Neyman-Pearson Lemma
►Power Analysis
►Significance Testing: An Overview
►Significance Tests, History and Logic of
►Statistical Evidence
►Statistical Inference
►Statistics: An Overview
►Testing Variance Components in Mixed Linear Models

References and Further Reading

Lehmann EL, Romano JP (2005) Testing statistical hypotheses 3rd edn. Springer, New York
Neyman J, Pearson E (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans Roy Stat Soc London 231:289–337
Pfanzagl J (1994) Parametric statistical theory. Gruyter, Berlin
Zacks S (1981) Parametric statistical inference. Pergamon, Oxford

Moving Averages

Rob J. Hyndman
Professor of Statistics
Monash University, Melbourne, VIC, Australia

A moving average is a time series constructed by taking averages of several sequential values of another time series. It is a type of mathematical convolution. If we represent the original time series by $y_1, \ldots, y_n$, then a *two-sided moving average* of the time series is given by

$$z_t = \frac{1}{2k+1} \sum_{j=-k}^{k} y_{t+j}, \qquad t = k+1, k+2, \ldots, n-k.$$

Thus $z_{k+1}, \ldots, z_{n-k}$ forms a new time series which is based on averages of the original time series, $\{y_t\}$. Similarly, a *one-sided moving average* of $\{y_t\}$ is given by

$$z_t = \frac{1}{k+1} \sum_{j=0}^{k} y_{t-j}, \qquad t = k+1, k+2, \ldots, n.$$

More generally, weighted averages may also be used. Moving averages are also called running means or rolling averages. They are a special case of "filtering", which is a general process that takes one time series and transforms it into another time series.

The term "moving average" is used to describe this procedure because each average is computed by dropping the oldest observation and including the next observation. The averaging "moves" through the time series until z_t is computed at each observation for which all elements of the average are available.

Note that in the above examples, the number of data points in each average remains constant. Variations on moving averages allow the number of points in each average to change. For example, in a cumulative average, each value of the new series is equal to the sum of all previous values.

Moving averages are used in two main ways: Two-sided (weighted) moving averages are used to "smooth" a time series in order to estimate or highlight the underlying trend; one-sided (weighted) moving averages are used as simple forecasting methods for time series. While moving averages are very simple methods, they are often building blocks for more complicated methods of time series smoothing, decomposition and forecasting.

Smoothing Using Two-Sided Moving Averages

It is common for a time series to consist of a smooth underlying trend observed with error:

$$y_t = f(t) + \varepsilon_t,$$

where $f(t)$ is a smooth and continuous function of t and $\{\varepsilon_t\}$ is a zero-mean error series. The estimation of $f(t)$ is known as smoothing, and a two-sided moving average is one way of doing so:

$$\hat{f}(t) = \frac{1}{2k+1} \sum_{j=-k}^{k} y_{t+j}, \qquad t = k+1, k+2, \ldots, n-k.$$

The idea behind using moving averages for smoothing is that observations which are nearby in time are also likely to be close in value. So taking an average of the points near an observation will provide a reasonable estimate of the trend at that observation. The average eliminates some of the randomness in the data, leaving a smooth trend component.

Moving averages do not allow estimates of $f(t)$ near the ends of the time series (in the first k and last k periods). This can cause difficulties when the trend estimate is used for forecasting or analyzing the most recent data.

Each average consists of $2k+1$ observations. Sometimes this is known as a $(2k+1)$ MA smoother. The larger the value of k, the flatter and smoother the estimate of $f(t)$ will be. A smooth estimate is usually desirable, but a flat estimate is biased, especially near the peaks and troughs in $f(t)$. When ε_t is a white noise series (i.e., independent and identically distributed with zero mean and variance σ^2), the bias is given by $\mathrm{E}[\hat{f}(x)] - f(x) \approx \frac{1}{6} f''(x)k(k+1)$ and the variance by $\mathrm{V}[\hat{f}(x)] \approx \sigma^2/(2k+1)$. So there is a trade-off between increasing bias (with large k) and increasing variance (with small k).

Centered Moving Averages

The simple moving average described above requires an odd number of observations to be included in each average. This ensures that the average is centered at the middle of the data values being averaged. But suppose we wish to calculate a moving average with an even number of observations. For example, to calculate a 4-term moving average, the trend at time t could be calculated as

$$\hat{f}(t-0.5) = (y_{t-2} + y_{t-1} + y_t + y_{t+1})/4$$

or

$$\hat{f}(t+0.5) = (y_{t-1} + y_t + y_{t+1} + y_{t+2})/4$$

That is, we could include two terms on the left and one on the right of the observation, or one term on the left and two terms on the right, and neither of these is centered on t. If we now take the average of these two moving averages, we obtain something centered at time t.

$$\begin{aligned}\hat{f}(t) &= \frac{1}{2}\left[(y_{t-2} + y_{t-1} + y_t + y_{t+1})/4\right] \\ &\quad + \frac{1}{2}\left[(y_{t-1} + y_t + y_{t+1} + y_{t+2})/4\right] \\ &= \frac{1}{8}y_{t-2} + \frac{1}{4}y_{t-1} + \frac{1}{4}y_t + \frac{1}{4}y_{t+1}\frac{1}{8}y_{t+2}\end{aligned}$$

So a 4 MA followed by a 2 MA gives a *centered moving average*, sometimes written as 2×4 MA. This is also a weighted moving average of order 5, where the weights for each period are unequal. In general, a $2 \times m$ MA smoother is equivalent to a weighted MA of order $m+1$ with weights $1/m$ for all observations except for the first and last observations in the average, which have weights $1/(2m)$.

Centered moving averages are examples of how a moving average can itself be smoothed by another moving average. Together, the smoother is known as a *double moving average*. In fact, any combination of moving averages can be used together to form a double moving average. For example, a 3×3 moving average is a 3 MA of a 3 MA.

Moving Averages. Table 1 Weight functions a_j for some common weighted moving averages

Name	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
3 MA	.333	.333										
5 MA	.200	.200	.200									
2 × 12 MA	.083	.083	.083	.083	.083	.083	.042					
3 × 3 MA	.333	.222	.111									
3 × 5 MA	.200	.200	.133	.067								
S15 MA	.231	.209	.144	.066	.009	−.016	−.019	−.009				
S21 MA	.171	.163	.134	.037	.051	.017	−.006	−.014	−.014	−.009	−.003	
H5 MA	.558	.294	−.073									
H9 MA	.330	.267	.119	−.010	−.041							
H13 MA	.240	.214	.147	.066	.000	−.028	−.019					
H23 MA	.148	.138	.122	.097	.068	.039	.013	−.005	−.015	−.016	−.011	−.004

S, Spencer's weighted moving average.
H, Henderson's weighted moving average.

Moving Averages with Seasonal Data

If the centered 4 MA was used with quarterly data, each quarter would be given equal weight. The weight for the quarter at the ends of the moving average is split between the two years. It is this property that makes 2×4 MA very useful for estimating a trend in the presence of quarterly seasonality. The seasonal variation will be averaged out exactly when the moving average is computed. A slightly longer or a slightly shorter moving average will still retain some seasonal variation. An alternative to a 2×4 MA for quarterly data is a 2×8 or 2×12 which will also give equal weights to all quarters and produce a smoother fit than the 2×4 MA. Other moving averages tend to be contaminated by the seasonal variation.

More generally, a $2 \times (km)$ MA can be used with data with seasonality of length m where k is a small positive integer (usually 1 or 2). For example, a 2×24 MA may be used for estimating a trend in monthly seasonal data (where $m = 12$).

Weighted Moving Averages

A weighted k-point moving average can be written as

$$\hat{f}(t) = \sum_{j=-k}^{k} a_j y_{t+j}.$$

For the weighted moving average to work properly, it is important that the weights sum to one and that they are symmetric, that is $a_j = a_{-j}$. However, we do not require that the weights are between 0 and 1. The advantage of weighted averages is that the resulting trend estimate is much smoother. Instead of observations entering and leaving the average abruptly, they can be slowly downweighted. There are many schemes for selecting appropriate weights. Kendall et al. (1983, Chap. 46) give details.

Some sets of weights are widely used and have been named after their proposers. For example, Spencer (1904) proposed a $5 \times 4 \times 4$ MA followed by a weighted 5-term moving average with weights $a_0 = 1$, $a_1 = a_{-1} = 3/4$, and $a_2 = a_{-2} = -3/4$. These values are not chosen arbitrarily, but because the resulting combination of moving averages can be shown to have desirable mathematical properties. In this case, any cubic polynomial will be undistorted by the averaging process. It can be shown that Spencer's MA is equivalent to the 15-point weighted moving average whose weights are −.009, −.019, −.016, .009, .066, .144, .209, .231, .209, .144, .066, .009, −.016, −.019, and −.009. Another Spencer's MA that is commonly used is the 21-point weighted moving average. Henderson's weighted moving averages are also widely used, especially as part of seasonal adjustment methods (Ladiray and Quenneville 2001). The set of weights is known as the *weight function*. Table 1 shows some common weight functions. These are all symmetric, so $a_{-j} = a_j$.

Weighted moving averages are equivalent to kernel regression when the weights are obtained from a kernel function. For example, we may choose weights using the quartic function

$$Q(j,k) = \begin{cases} \left\{1 - [j/(k+1)]^2\right\}^2 & \text{for } -k \le j \le k; \\ 0 & \text{otherwise.} \end{cases}$$

Then a_j is set to $Q(j,k)$ and scaled so the weights sum to one. That is,

$$a_j = \frac{Q(j,k)}{\sum_{i=-k}^{k} Q(i,k)}. \quad (1)$$

Forecasting Using One-Sided Moving Averages

A simple forecasting method is to average the last few observed values of a time series. Thus

$$\hat{y}_{t+h|t} = \frac{1}{k+1} \sum_{j=0}^{k} y_{t-j}$$

provides a forecast of y_{t+h} given the data up to time t.

As with smoothing, the more observations included in the moving average, the greater the smoothing effect. A forecaster must choose the number of periods $(k+1)$ in a moving average. When $k = 0$, the forecast is simply equal to the value of the last observation. This is sometimes known as a "naïve" forecast.

An extremely common variation on the one-sided moving average is the exponentially weighted moving average. This is a weighted average, where the weights decrease exponentially. It can be written as

$$\hat{y}_{t+h|t} = \sum_{j=0}^{t-1} a_j y_{t-j}$$

where $a_j = \lambda(1-\lambda)^j$. Then, for large t, the weights will approximately sum to one. An exponentially weighted moving average is the basis of simple exponential smoothing. It is also used in some process control methods.

Moving Average Processes

A related idea is the moving average process, which is a time series model that can be written as

$$y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q},$$

where $\{e_t\}$ is a white noise series. Thus, the observed series y_t, is a weighted moving average of the unobserved e_t

series. This is a special case of an Autoregressive Moving Average (or ARMA) model and is discussed in more detail in the entry ►Box-Jenkins Time Series Models. An important difference between this moving average and those considered previously is that here the moving average series is directly observed, and the coefficients $\theta_1, \ldots, \theta_q$ must be estimated from the data.

Cross References

►Box–Jenkins Time Series Models
►Forecasting with ARIMA Processes
►Forecasting: An Overview
►Median Filters and Extensions
►Seasonality
►Smoothing Techniques
►Statistical Quality Control: Recent Advances
►Time Series
►Trend Estimation

References and Further Reading

Kendall MG, Stuart A, Ord JK (1983) Kendall's advanced theory of statistics. vol 3. Hodder Arnold, London

Ladiray D, Quenneville B (2001) Seasonal adjustment with the X-11 method, vol 158, of Lecture notes in statistics. Springer, Berlin

Makridakis S, Wheelwright SC, Hyndman RJ (1998) Forecasting: methods and applications, 3rd edn. Wiley, New York

Spencer J (1904) On the graduation of the rates of sickness and mortality presented by the experience of the Manchester Unity of Oddfellows during the period 1893–1897. J Inst Actuaries 38:334–343

Multicollinearity

Vlasta Bahovec
Professor, Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

One of the assumptions of the standard regression model $y = X\beta + \varepsilon$ is that there is no exact linear relationship among the explanatory variables, or equivalently, that the matrix X of explanatory variables has a full rank. The problem of multicollinearity occurs if two or more explanatory variables are linearly dependent, or near linearly dependent (including the variable $x_0' = [1, 1, \cdots, 1]$, which generates a constant term). There are two types of multicollinearity: perfect and near multicollinearity.

Perfect multicollinearity occurs if at least two explanatory variables are linearly dependent. In that case, the determinant of matrix $X'X$ equals zero (the $X'X$ matrix is singular), and therefore ordinary least squares (OLS) estimates of regression parameters $\beta' = (\beta_0, \beta_1, \cdots, \beta_k)$

$$\hat{\beta} = (X'X)^{-1}X'y = \frac{adj(X'X)}{\det(X'X)} \cdot X'y$$

are not unique. This type of multicollinearity is rare, but may occur if the regression model includes qualitative explanatory variables, whose effect is taken into account by ►dummy variables. Perfect multicollinearity occurs in a regression model with an intercept, if the number of dummy variables for each qualitative variable is not less than the number of groups of this variable. Perfect multicollinearity can easily be revealed. A more difficult problem is near or imperfect multicollinearity. This problem arises if at least two regressors are highly intercorrelated. In that case, $\det(X'X) \approx 0$, the matrix $X'X$ is ill conditioned, and therefore the estimated parameters are numerically imprecise. Furthermore, since the covariance matrix of estimated parameters is calculated by the formula $Cov(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$, the variances and covariances of the estimated parameters will be large. Large standard errors $SE(\hat{\beta}_j) = \hat{\sigma}\sqrt{(X'X)^{-1}_{jj}}$ imply that empirical t-ratios $\left(t_j = \hat{\beta}_j / SE\left(\hat{\beta}_j\right)\right)$ could be insignificant, which may lead to an incorrect conclusion that some explanatory variables have to be omitted from the regression model. Also, large standard errors make interval parameter estimates imprecise.

Imperfect multicollinearity often arises in the time series regression model (see ►Time Series Regression), especially in data involving economic time series, while variables over time tend to move in the same direction.

The simplest way to detect serious multicollinearity problems is to analyze variances of estimated parameters, which are calculated with the following formula:

$$var(\hat{\beta}_j) = \sigma^2(X'X)^{-1}_{jj} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \cdot \left(1 - R_j^2\right)},$$

where R_j^2 is the coefficient of determination in the regression, variable x_j is the dependent, and the remaining x's are explanatory variables. If variable x_j is highly correlated with other regressors, R_j^2 will be large (near to 1), and therefore the variance of $\hat{\beta}_j$ will be large. There are some measures of multicollinearity included in standard statistical software: the variance inflation factor (VIF), tolerance (TOL), condition number (CN), and condition indices (CI). VIF and TOL are calculated with the following formulas:

$$VIF_j = \frac{1}{1 - R_j^2} \quad j = 1, 2, \cdots, k \quad TOL_j = \frac{1}{VIF_j} = 1 - R_j^2.$$

The multicollinearity problem is serious if $R_j^2 > 0.8$, consequently if $VIF_j > 5$, or equivalently if $TOL_j < 0.2$.

More sophisticated measures of multicollinearity are condition number, CN, and condition indices, CI_i, based on the use of eigenvalues of the $X'X$ matrix. CN is the square root of the ratio of the largest eigenvalue to the smallest eigenvalue, and CI_i, $i = 1, 2, \ldots, k$, are square roots of the ratio of the largest eigenvalue to each individual eigenvalue. These measures, which are calculated with the formulas

$$CN = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \qquad CI_i = \sqrt{\frac{\lambda_{\max}}{\lambda_i}} \quad i = 1, 2, \cdots, k,$$

are measures of sensitivity of parameter estimates to small changes in data. Some authors, such as Belsley et al. (1980), suggested that a condition index of 30–100 indicates moderate to strong multicollinearity.

Several solutions have been suggested to rectify the multicollinearity problem. Some are the following: (1) increasing the sample size to reduce multicollinearity, as multicollinearity is a problem of the sample, and not the population; (2) dropping one or more variables suspected of causing multicollinearity; (3) transforming data as the first differences $\Delta X_t = X_t - X_{t-1}$ or ratios X_t/X_{t-1} $t = 2, 3, \ldots, n$ to eliminate linear or exponential trends; (4) ridge regression (see ►Ridge and Surrogate Ridge Regressions); and (5) principal component regression.

The problem of multicollinearity is approached differently by econometricians depending on their research goal. If the goal is to forecast future values of the dependent variable, based on the determined regression model, the problem of multicollinearity is neglected. In all other cases, this problem is approached more rigorously.

Cross References

►Dummy Variables
►Heteroscedasticity
►Linear Regression Models
►Multivariate Statistical Analysis
►Partial Least Squares Regression Versus Other Methods
►Ridge and Surrogate Ridge Regressions

References and Further Reading

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: Identifying: Influential data and sources of collinearity. Wiley, New York

Green WH (2002) Econometric analysis, 5th edn. Prentice Hall, New Jersey

Gujarati DN (2002) Basic econometrics, 4th edn. McGraw-Hill/Irwin, New York

Maddala GS (2002) Introduction to econometrics, 3rd edn. Wiley, Chichester

Multicriteria Clustering

Anuška Ferligoj
Professor, Head of the Center of Informatics and Methodology, Faculty of Social Sciences
University of Ljubljana, Ljubljana, Slovenia

Some clustering problems cannot be appropriately solved with classical clustering algorithms because they require optimization over more than one criterion. In general, solutions optimal according to each particular criterion are not identical. Thus, the problem arises of how to find the best solution satisfying as much as possible all criteria considered. In this sense the set of Pareto efficient clusterings was defined: a clustering is Pareto efficient if it cannot be improved on any criterion without sacrificing some other criterion.

A multicriteria clustering problem can be approached in different ways:

- By reduction to a clustering problem with a single criterion obtained as a combination of the given criteria;
- By constrained clustering algorithms where a selected criterion is considered as the clustering criterion and all others determine the constraints;
- By direct algorithms: Hanani (1979) proposed an algorithm based on the dynamic clusters method using the concept of the kernel, as a representation of any given criterion. Ferligoj and Batagelj (1992) proposed modified relocation algorithms and modified agglomerative hierarchical algorithms.

Usual Clustering Problems

Cluster analysis (known also as classification and taxonomy) deals mainly with the following general problem: given a set of units, $\mathcal{U}$, determine subsets, called clusters, C, which are homogeneous and/or well separated according to the measured variables (e.g., Sneath and Sokal 1973; Hartigan 1975; Gordon 1981). The set of clusters forms a clustering. This problem can be formulated as an optimization problem:

Determine the clustering $\mathbf{C}^*$ for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi} P(\mathbf{C})$$

where $\mathbf{C}$ is a clustering of a given *set of units*, $\mathcal{U}$, Φ is the set of all feasible clusterings and $P : \Phi \to \mathbb{R}$ a *criterion function*.

As the set of feasible clusterings is finite a solution of the clustering problem always exists. Since this set is usually large it is not easy to find an optimal solution.

A Multicriteria Clustering Problem

In a *multicriteria clustering problem* $(\Phi, P_1, P_2, \ldots, P_k)$ we have several criterion functions $P_t, t = 1, \ldots, k$ over the same set of feasible clusterings Φ, and our aim is to determine the clustering $\mathbf{C} \in \Phi$ in such a way that

$$P_t(\mathbf{C}) \to \min, \quad t = 1, \ldots, k.$$

In the ideal case, we are searching for the dominant set of clusterings. The solution $\mathbf{C}_0$ is the *dominant* solution if for each solution $\mathbf{C} \in \Phi$ and for each criterion P_t, it holds that

$$P_t(\mathbf{C}_0) \le P_t(\mathbf{C}), \quad t = 1, \ldots, k.$$

Usually the set of dominant solutions is empty. Therefore, the problem arises of finding a solution to the problem that is as good as is possible according to each of the given criteria. Formally, the *Pareto-efficient* solution is defined as follows:

For $\mathbf{C}_1, \mathbf{C}_2 \in \Phi$, solution $\mathbf{C}_1$ *dominates* solution $\mathbf{C}_2$ if and only if

$$P_t(\mathbf{C}_1) \le P_t(\mathbf{C}_2), \quad t = 1, \ldots, k,$$

and for at least one $i \in 1, \ldots, k$ the strict inequality $P_i(\mathbf{C}_1) < P_i(\mathbf{C}_2)$ holds. We denote the dominance relation by $\prec$. $\prec$ is a strict partial order. The set of Pareto-efficient solutions, Π, is the set of minimal elements for the dominance relation:

$$\Pi = \{\mathbf{C} \in \Phi : \neg\exists \mathbf{C}' \in \Phi : \mathbf{C}' \prec \mathbf{C}\}$$

In other words, the solution $\mathbf{C}^* \in \Phi$ is *Pareto-efficient* if there exists no other solution $\mathbf{C} \in \Phi$ such that

$$P_t(\mathbf{C}) \le P_t(\mathbf{C}^*), \quad t = 1, \ldots, k,$$

with strict inequality for at least one criterion. A *Pareto-clustering* is a Pareto-efficient solution of the multicriteria clustering problem (Ferligoj and Batagelj 1992).

Since the optimal clusterings for each criterion are Pareto-efficient solutions the set Π is not empty. If the set of dominant solutions is not empty then it is equal to the set of Pareto-efficient solutions.

Solving Discrete Multicriteria Optimization Problems

Multicriteria clustering problems can be approached as a multicriteria optimization problem, that has been treated by several authors (e.g., Chankong and Haimes 1983; Ferligoj and Batagelj 1992). In the clustering case, we are dealing with discrete multicriteria optimization (the set of feasible solutions is finite), which means that many very useful theorems in the field of multicriteria optimization do not hold, especially those which require convexity. It was proven that if, for each of the given criteria, there is a unique solution, then the minimal number of Pareto-efficient solutions to the given multicriteria optimization problem equals the number of different minimal solutions of the single criterion problems.

Although several strategies haven been proposed for solving multicriteria optimization problems explicitly, the most common is the conversion of the multicriteria optimization problem to a single criterion problem.

Direct Multicriteria Clustering Algorithms

The multicriteria clustering problem can be approached efficiently by using direct algorithms. Two types of direct algorithms are known: a version of the relocation algorithm, and the modified agglomerative (hierarchical) algorithms (Ferligoj and Batagelj 1992).

Modified Relocation Algorithm

The idea of the *modified relocation* algorithm for solving the multicriteria clustering problem follows from the definition of a Pareto-efficient clustering. The solutions obtained by the proposed procedure can be only *local Pareto clusterings*. Therefore, the basic procedure should be repeated *many* times (at least hundreds of times) and the obtained solutions should be reviewed. An efficient review of the obtained solutions can be systematically done with an appropriate *metaprocedure* with which the true set of Pareto clusterings can be obtained.

Modified Agglomerative Hierarchical Approach

Agglomerative hierarchical clustering algorithms usually assume that all relevant information on the relationships between the n units from the set $\mathcal{U}$ is summarized by a symmetric pairwise dissimilarity matrix $D = [d_{ij}]$. In the case of multicriteria clustering we assume we have k dissimilarity matrices $D^t, t = 1, \ldots, k$, each summarizing all relevant information obtained, for example, in the k different situations. The problem is to find the best hierarchical solution which satisfies as much as is possible all k dissimilarity matrices.

One approach to solving the multicriteria clustering problem combines the given dissimilarity matrices (at each step) into a composed matrix. This matrix $D = [d_{ij}]$ can, for example, be defined as follows:

$$d_{ij} = \max\left(d_{ij}^t; t = 1, \ldots, k\right)$$
$$d_{ij} = \min\left(d_{ij}^t; t = 1, \ldots, k\right)$$
$$d_{ij} = \sum_{t=1}^{k} \alpha_t d_{ij}^t \ , \ \sum_{t=1}^{k} \alpha_t = 1$$

Following this approach, one of several *decision rules* (e.g., pessimistic, optimistic, Hurwicz, Laplace) for making decisions under uncertainty (Chankong and Haimes 1983) can be used at the composition and selection step of the agglomerative procedure.

Conclusion

The multicriteria clustering problem can be treated with the proposed approaches quite well if only a few hundreds units are analysed. New algorithms have to be proposed for large datasets.

About the Author

Anuška Ferligoj is Professor at the Faculty of Social Sciences at University of Ljubljana, head of the graduate program on Statistics at the University of Ljubljana and head of the Center of Methodology and Informatics at the Institute of Social Sciences. She is editor of the journal Advances in Methodology and Statistics (since 2004). She was awarded the title of Ambassador of Science of the Republic of Slovenia in 1997. Dr Ferligoj is a Fellow of the European Academy of Sociology. For the monograph *Generalized Blockmodeling* she was awarded the Harrison White Outstanding Book Award for 2007, the Mathematical Sociology Section of the American Sociological Association. In 2010 she received Doctor et Professor Honoris Causa at ELTE University in Budapest.

Cross References

►Cluster Analysis: An Introduction
►Data Analysis
►Distance Measures
►Fuzzy Logic in Statistical Data Analysis
►Hierarchical Clustering
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Random Permutations and Partition Models

References and Further Reading

Chankong V, Haimes YY (1983) Multiobjective decision making. North-Holland, New York
Ferligoj A, Batagelj V (1992) Direct multicriteria clustering algorithms. J Clas. 9:43–61
Gordon AD (1981) Classification. Chapman & Hall, London
Hanani U (1979) Multicriteria dynamic clustering, Rapport de Recherche No. 358, IRIA, Rocquencourt
Hartigan JA (1975) Clustering algorithms. Wiley, New York
Sneath PHA, Sokal RR (1973) Numerical taxonomy Freeman, San Francisco

Multicriteria Decision Analysis

Theodor J. Stewart
Emeritus Professor
University of Cape Town, Rondebosch, South Africa
University of Manchester, Manchester, UK

Basic Definitions

The field variously described as *multicriteria decision making (MCDM)* or *multicriteria decision analysis* or *aid (MCDA)* is that branch of operational research/management science (OR/MS) that deals with the explicit modeling of multiple conflicting goals or objectives in management decision making. Standard texts in OR/MS typically do include identification of objectives (often stated as plural) as a key step in the decision-making process, but the ensuing discussion appears to assume that such objectives are easily aggregated into a single measure of achievement which can formally be optimized. The field of MCDA, however, arose from a recognition that systematic and coherent treatment of multiple objectives requires structured decision support to ensure that all interests are kept in mind and that an informed balance is achieved. See, for example, the discussions and associated references in Chap. 2 of Belton and Stewart (2002) and Chap. 1 of Figueira et al. (2005).

The starting point of MCDA is the identification of the critical *criteria* according to which potential courses of action (choices, policies, strategies) may be compared and evaluated. In this sense, each *criterion* is a particular point of view or consideration according to which preference orders on action outcomes can (more-or-less) unambiguously be specified. Examples of such criteria may include issues such as investment costs, job creation, levels of river pollution etc., as well as more subjective criteria such as aesthetic appeal. With careful selection of the criteria, preference ordering according to each could be essentially self-evident apart from some fuzziness around the concept equality of performance.

Selection of criteria is a profound topic in its own right, but is perhaps beyond the scope of the present article. Some discussion may be found in Keeney and Raiffa (1976); Keeney (1992); Belton and Stewart (2010). In essence, the analyst needs to ensure that values and aspirations of the decision maker(s) have been fully captured by the chosen criteria, while still retaining a manageably small number of criteria (typically, one strives for not much more than 15 or 25 criteria in most applications). Care needs to be taken not

to double-count issues, and that preference orders can be understood on each criterion independently of the others.

Suppose then that say m criteria have been defined as above. For any specified course of action, say $a \in \mathscr{A}$ (the set of all possible actions), we define $z_i(a)$ to be a measure of performance of a according to the perspective of criterion i, for $i = 1, \ldots, m$. The scaling at this stage is not important, the only requirement being that action a is preferred to action b in terms of criterion i ($a \succ_i b$) if and only if $z_i(a) > z_i(b) + \epsilon_i$ for some tolerance parameter ϵ_i. Apart from the brief comments in the final section, we assume that these measures of performance are non-stochastic.

The primary aim of MCDA is to support the decision maker in aggregating the single-criterion preferences into an overall preference structure, in order to make a final selection which best satisfies all criteria, or to select a reduced subset of $\mathscr{A}$ for further discussion and evaluation. It is important to recognize that this aggregation phase contains fundamentally subjective elements, namely the value judgments and tradeoffs provided by the decision maker. We shall briefly review some of the support processes which are used. A comprehensive overview of these approaches may be found in Figueira et al. (2005).

Methods of Multicriteria Analysis

It is important to recognize that two distinct situations may arise in the context described above, and that these may lead to broadly different forms of analysis:

- *Discrete choice problems:* In this case, $\mathscr{A}$ consists of a discrete set of options, e.g., alternative locations for a power station. The discrete case arises typically at the level of high level strategic choices, within which many of the criteria may require subjective evaluation of alternatives.
- *Multiobjective optimization problems:* These problems are often defined in mathematical programming terms, i.e., an option will be defined in terms of a vector of *decision variables*, say $\mathbf{x} \in \mathbb{X} \subset \mathbb{R}^n$. The measures of performance for each criterion typically need to be defined quantitatively in terms of functions $f_i(\mathbf{x})$ mapping $\mathbb{R}^n \to \mathbb{R}$ for each i.

The methods adopted can be characterized in two ways:

- By the underlying paradigm for modeling human preferences (*preference modeling*);
- By the stage of the analysis at which the decision makers' judgments are brought into play (*timing of preference statements*).

We deal with each of these in turn.

Preference Modeling

At least four different paradigms can be identified.

1. **Value scoring or utility methods:** The approach is first to re-scale the performance measures $z_i(a)$ so as to be commensurate in some way, typically by means of transformation through a *partial value function*, say $v_i(z_i)$. This rescaling needs to ensure that equal-sized intervals in the transformed scale represent the same importance to the decision maker (in terms of trade-offs with other criteria) irrespective of where they occur along the scale. Relatively mild assumptions (under conditions of deterministic performance measures) imply that an overall value of a can be modeled additively, i.e., as $V(a) = \sum_{i=1}^{m} w_i v_i(z_i(a))$. The assessment of the partial values and weights (w_i) may be carried out by direct assessment (e.g., Dyer 2005), indirectly such as by the analytic hierarchy process approach (Saaty 2005), or by learning from previous choices (Siskos et al. 2005).
2. **Metric methods:** In this approach, some form of goal or aspiration is specified (by the decision maker) for each criterion, say G_i for each i. A search (discrete or by mathematical optimization) is then conducted to find the option for which the performance levels $z_1(a), z_2(a), \ldots, z_m(a)$ approach the goal levels $G_1, G_2, \ldots, G_m$ as closely as possible. Typically, L_1, L_2, or L_∞ metrics are used to define closeness, with provision for differential weighting of criteria. Differences do also arise in terms of whether over-achievement of goals adds additional benefits or not. Such approaches are termed (generalized) goal programming, and are reviewed in Lee and Olson; Wierzbicki (1999; 1999). Goal programming is primarily applied in the context of the multiobjective optimization class of model.
3. **Outranking methods:** These methods consider action alternatives pairwise in terms of their performance levels on all criteria, in order to extract the level of evidence in the data provided by the performance measures which either support (are concordant with) or oppose (are discordant with) a conclusion that the one action is better than the other. These considerations generate partial rankings of the actions, or at least a classification of the actions into ordered preference classes. Descriptions of different outranking approaches may be found in Part III of Figueira et al. (2005).
4. **Artificial intelligence:** Greco et al. (2005) describe how observed choices by the decision maker(s) can

be used to extract decision rules for future multicriteria decisions, without explicit or formal preference modeling along the lines described above.

Timing of Preference Statements

Three possible stages of elicitation of values and preferences from the decision maker may be recognized as described below (although in practice no one of these is used completely in isolation).

1. **Elicitation prior to analysis of options:** In this approach, a complete model of the decision maker preferences is constructed from a sequence of responses to questions about values, trade-offs, relative importance, etc. The resulting model is then applied to the elements of $\mathscr{A}$ in order to select the best alternative or a shortlist of alternatives. This approach is perhaps most often used with value scoring methods, in which a simple and transparent preference model (e.g., the additive value function) is easily constructed and applied.
2. **Interactive methods:** Here a tentative preference model, incomplete in many ways, is used to generate a small number of possible choices which are presented to the decision maker, who may either express strong preferences for some or dislike of others. On the basis of these stated preferences, models are refined and a new set of choices generated. Even in the prior elicitation approach, some degree of interaction of this nature will occur, where in the application of value scoring or outranking approaches to discrete choice problems, results will inevitably be fed back to decision makers for reflection on the value judgements previously specified. However, it is especially with continuous multiobjective optimization problems that the interaction becomes firmly designed and structured into the process. See Chap. 5 of Miettinen (1999) for a comprehensive coverage of such structured interaction.
3. **Posterior value judgements:** If each performance measure is to be maximized, then an action a is said to *dominate* action b if $z_i(a) \geq z_i(b)$ for all criteria, with strict inequality for at least one criterion. With discrete choice problems, the removal of dominated actions from $\mathscr{A}$ may at times reduce the set of options to such a small number that no more analysis is necessary – decision makers can make a holistic choice. In some approaches to multiobjective optimization (see also Miettinen 1999), a similar attempt is made to compute the "efficient frontier," i.e., the image in criterion space of all non-dominated options, which can be displayed to the decision maker for a holistic choice. In practice, however, this approach is restricted to problems with two or three criteria only which can be displayed graphically (although there have been attempts at graphical displays for slightly higher dimensionality problems).

Stochastic MCDA

As indicated at the start, we have focused on deterministic problems, i.e., in which a fixed (even if slightly "fuzzy") performance measure $z_i(a)$ can be associated with each action-criterion combination. However, there do of course exist situations in which each $z_i(a)$ will be a *random variable*.

The introduction of stochastic elements into the multicriteria decision making problem introduces further complications. Attempts have been made to adapt value scoring methods to be consistent with the von Neumann/Morgenstern axioms of expected utility theory, to link multicriteria decision analysis with scenario planning, and to treat probabilities of achieving various critical outcomes as separate "criteria." Discussion of these extensions is beyond the scope of space available for this short article, but a review is available in Stewart (2005).

About the Author

Professor Stewart is Past-President of both the Operations Research Society of South Africa (1978) and the South African Statistical Association (1989). He was Vice President of IFORS (International Federation of Operational Research Societies) for the period 2004–2006, and President of the International Society on Multiple Criteria Decision Making for the period 2004–2008. He is currently Editor-in-Chief of the *Journal of Multi-Criteria Decision Analysis*, and African Editor of *International Transactions in Operations*. He is a Member of the Academy of Science of South Africa. In 2008 Professor Stewart was awarded the Gold medal of the International Society on Multiple Criteria Decision Making (for marked contributions to theory, methodology and practice in the field), and has been awarded ORSSA's Tom Roszwadowski Medal (for written contributions to OR) on five occasions.

Cross References

►Decision Theory: An Introduction
►Decision Theory: An Overview

References and Further Reading

Belton V, Stewart TJ (2002) Multiple criteria decision analysis: an integrated approach. Kluwer, Boston

Belton V Stewart TJ (2010) Problem structuring and MCDA. In: Ehrgott M, Figueira JR, Greco S (eds) Trends in multiple criteria decision analysis, chapter 8. Springer, Berlin, pp 237–271

Dyer JS (2005) MAUT – multiattribute utility theory. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 7. Springer, New York, pp 265–295

Figueira J, Greco S, Ehrgott M (eds) (2005) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76. Springer, New York

Gal T, Stewart TJ, Hanne T (eds) (1999) Multicriteria decision making: advances in MCDM models, algorithms, theory, and applications. Kluwer, Boston

Greco S, Matarazzo B, Słowiński R (2005) Decision rule approach. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 13. Springer, New York, pp 507–561

Keeney RL (1992) Value-focused thinking: a path to creative decision making. Harvard University Press, Cambridge

Keeney RL, Raiffa H (1976) Decisions with multiple objectives. Wiley, New York

Lee SM, Olson DL (1999) Goal programming. In: Gal T, Stewart TJ, Hanne T (eds) Multicriteria decision making: advances in MCDM models, algorithms, theory, and applications, chapter 8. Kluwer, Boston

Miettinen K (1999) Nonlinear multiobjective optimization, International series in operations research and management science, vol 12. Kluwer, Dordrecht

Saaty TL (2005) The analytic hierarchy and analytic network processes for the measurement of intangible criteria and for decision-making. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 9. Springer, New York, pp 345–407

Siskos Y, Grigoroudis E, Matsatsinis N (2005) MAUT – multiattribute utility theory. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 8. Springer, New York, pp 299–343

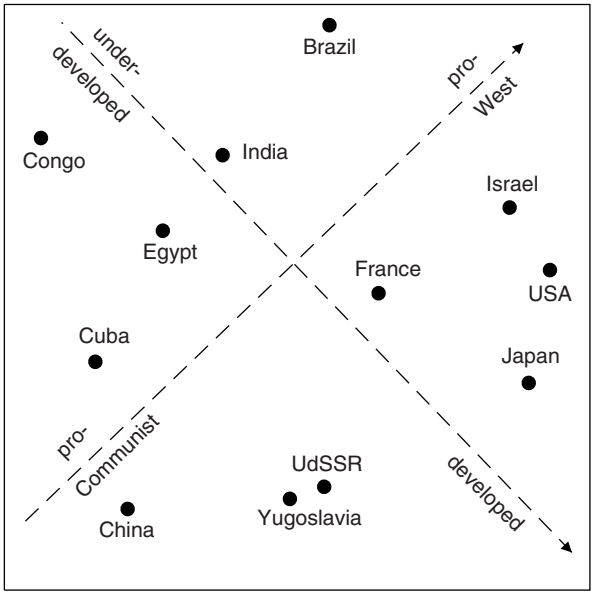
Stewart TJ (2005) Dealing with uncertainties in MCDA. In: Figueira J, Greco S, Ehrgott M (eds) Multiple criteria decision analysis – state of the art annotated surveys. International series in operations research and management science, vol 76, chapter 11. Springer, New York, pp 445–470

Wierzbicki AP (1999) Reference point approaches. In: Gal T, Stewart TJ, Hanne T (eds) Multicriteria decision making: advances in MCDM models, algorithms, theory, and applications, chapter 9, Kluwer, Boston

Multidimensional Scaling

Ingwer Borg
Professor of Applied Psychological Methods
University of Giessen, Giessen, Germany
Scientific Director
GESIS, Mannheim, Germany

►Multidimensional scaling (MDS) is a family of methods that optimally map *proximity indices* of objects into distances between points of a multidimensional space with a given dimensionality (usually two or three dimensions). The main purpose for doing this is to visualize the data so that the user can test structural hypotheses or discover patterns "hidden" in the data.



Multidimensional Scaling. Fig. 1 MDS configuration for country similarity data

Historically, MDS began as a psychological model for judgments of (dis)similarity. A typical example of this early era is the following. Wish (1971) was interested to find out how persons generate overall judgments on the similarity of countries. He asked a sample of subjects to assess each pair of twelve countries with respect to their global similarity. For example, he asked "How similar are Japan and China?", offering a 9-point answer scale from "very dissimilar" to "very similar" for the answer. On purpose, "there were no instructions concerning the characteristics on which these similarity judgments were to be made; this was information to discover rather than to impose" (Kruskal and Wish 1978:30). The resulting numerical ratings were averaged over subjects, and then mapped via MDS into the distances among 12 points of a Euclidean plane. The resulting MDS configuration (Fig. 1) was interpreted to show that the ratings were essentially generated from two underlying dimensions.

As an MDS model, Wish (1971) used *ordinal MDS*, the most popular MDS model. It maps the proximities of the n objects (δ_{ij}) into distances d_{ij} of the $n \times m$ configuration $\mathbf{X}$ such that their ranks are optimally preserved. Hence, assuming that the δ_{ij}'s are dissimilarities, the function $f : \delta_{ij} \rightarrow d_{ij}(\mathbf{X})$ is monotone so that $f : \delta_{ij} < \delta_{kl} \rightarrow d_{ij}(\mathbf{X}) \leq d_{kl}(\mathbf{X})$, for all pairs (i,j) and (k,l) for which

data are given. Missing data impose no constraints onto the distances.

Another popular MDS model is *interval MDS*, where $f : \delta_{ij} \rightarrow a + b \cdot \delta_{ij} = d_{ij}(\mathbf{X})$. This model assumes that the data are given on an interval scale. Hence, both a and $b(\neq 0)$ can be chosen arbitrarily. In particular, they can be chosen such that the re-scaled proximities are equal to the distances of a given MDS configuration $\mathbf{X}$.

A second facet of an MDS model is the distance function that it uses. In psychology, the family of *Minkowski distances* has been studied extensively as a model of judgment. Minkowski distances can be expressed by the formula

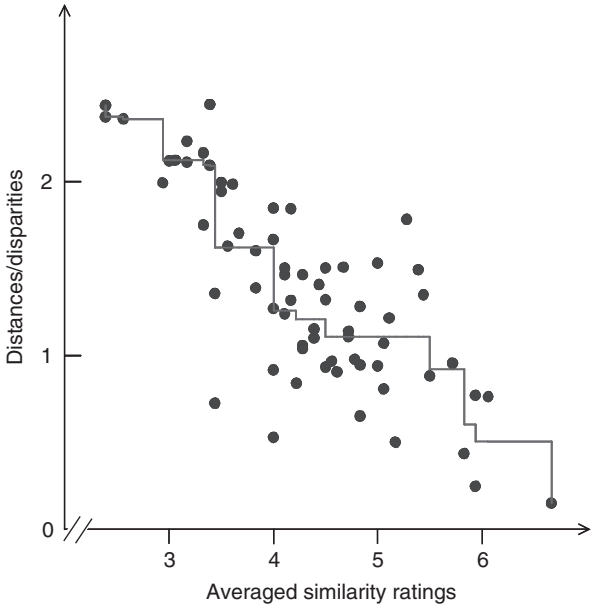
$$d_{ij}^{(p)}(\mathbf{X}) = \left(\sum_{a=1}^{m} |x_{ia} - x_{ja}|^p \right)^{1/p}, p \geq 1. \quad (1)$$

Setting $p = 1$ results in the *city-block metric*, setting $p = 2$ in the *Euclidean distance*. If p grows, d_{ij} is quickly dominated by its largest intra-dimensional difference (out of the $a = 1, \ldots, m$ dimensions). Such metrics supposedly explain fast and frugal (dis)similarity judgments. The city-block metric, in contrast, models careful judgments with important consequences for the individual. When MDS is used for exploratory purposes, however, only $p = 2$ should be used, because all other choices imply geometries with non-intuitive properties.

The fit of the MDS representation to the data can be seen from its *Shepard diagram*. For our country-similarity example, this is shown in Fig. 2. The plot exhibits how the data are related to the distances. It also shows the monotone regression line. The vertical scatter of the points about this regression line corresponds to the model's loss or misfit. It is measured as $\sum_{i<j} e_{ij}^2 = \sum_{i<j} (d_{ij}(\mathbf{X}) - f(\delta_{ij}))^2$, for all points i und j. The $f(\delta_{ij})$'s here are *disparities*, i.e., proximities that are re-scaled using all admissible transformations of the chosen scale level to optimally approximate the corresponding distances of the MDS configuration $\mathbf{X}$. The optimization is done by ordinal or linear regression (or, generally, by regression of type f) so that $f(\delta_{ij}) = \widehat{d}_{ij}(\mathbf{X})$. In order to obtain an interpretable measure of model misfit, the error sum is normed to yield the standard MDS loss function

$$Stress = \sqrt{\sum_{i<j} (d_{ij}(\mathbf{X}) - \widehat{d}_{ij})^2 / \sum_{i<j} d_{ij}^2(\mathbf{X})}. \quad (2)$$

A perfect MDS solution has a Stress of zero. In this case, the distances of the MDS solution correspond perfectly to the disparities. For the above example, we get Stress = 0.19. Evaluating if this is an acceptably low value is complex. A minimum criterion is that the observed Stress value should be clearly smaller than the Stress that results for random data. Other criteria (such as the number of points (n), the number of missing data, the restrictiveness of the MDS model, or the dimensionality of the MDS space (m)), but also the interpretability of the solution have to be taken into account. Indeed, it may be true that Stress is high but the configuration is nevertheless stable over replications of the data. This case can result if the data have a large random error component. MDS, then, acts as a *data smoother* that irons out the error in the distance representation.



Multidimensional Scaling. Fig. 2 Shepard diagram of MDS solution in Fig. 1

MDS methods allow one to utilize many different proximity measures. One example is direct judgments of similarity or dissimilarity as in the example given above. Another example are intercorrelations of test items over a sample of persons. A third example are co-occurrence coefficients that assess how often an event X is observed together with another event Y.

MDS is also robust against randomly distributed missing data. Computer simulations show that some 80% of the proximities may be missing, provided the data contain little error and the number of points (n) is high relative to the dimensionality of the MDS space (m). The data can also be quite coarse and even dichotomous.

A popular variety of MDS is *Individual Differences Scaling* or INDSCAL (Carroll and Chang 1970). Here, we have N different proximity matrices, one for each of N persons. The idea of the model is that these proximities can

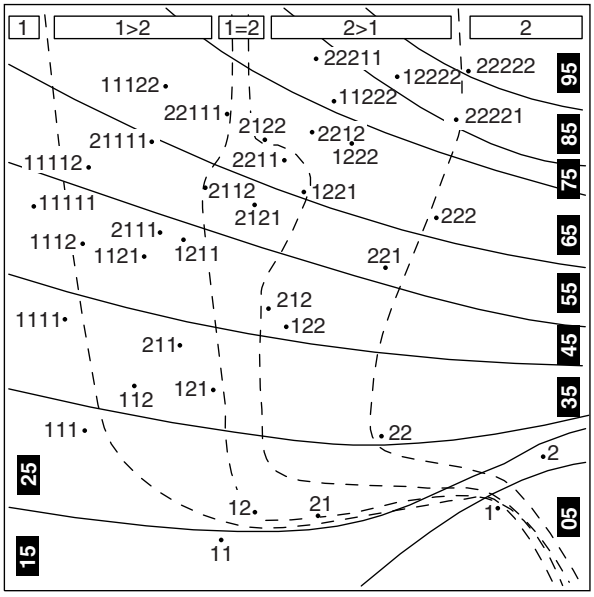
be explained by individually stretching or compressing a common MDS space along a fixed set of dimensions. That is,

$$d_{ij}^{(k)}(\mathbf{X}) = \sqrt{\sum_{a=1}^{m} w_a^{(k)} (x_{ia} - x_{ja})^2}, w_a^{(k)} \geq 0, \qquad (3)$$

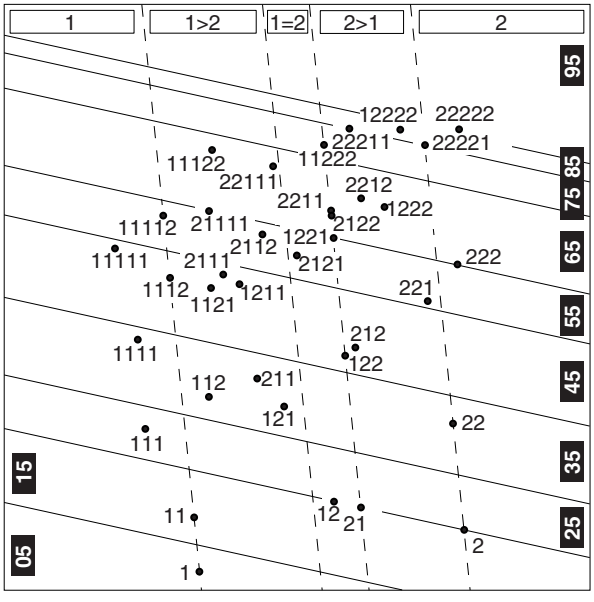
where $k = 1, \ldots, N$. The weight $w_a^{(k)}$ is interpreted as the salience of dimension a for individual k. Carroll and Wish (1974) used INDSCAL on the overall similarity ratings of different individuals for a set of countries, similar to the data discussed above. What they find is that one group of persons ("doves") pays much attention to economic development, while the other group ("falcons") emphasizes almost only political alignment of the countries with the West. Note, though, that these interpretations depend on the norming of $\mathbf{X}$. A more transparent way to analyze such data is to scale each individual's data matrix by itself, and then proceed by Procrustean fittings of the various solutions to each other, followed by finding optimal dimensions for an INDSCAL-type weighting model (Lingoes and Borg 1978).

A second popular variety of MDS is *Unfolding*. The prototypical data for this model are preference ratings of a set of persons for a set of objects. These data are mapped into distances between person-points and object-points in a "joint" space. The person-points are interpreted as "ideal" points that express the persons' points of maximal preference in the object space.

MDS solutions can be interpreted in different ways. The most popular approach is interpreting dimensions, but this is just a special case of interpreting regions. Regions are partitions of the MDS space which sort its points into subgroups that are equivalent in terms of substance. A systematic method for that purpose is *facet theory* (Borg and Shye 1995), an approach that offers methods to cross-classify the objects into substantively meaningful cells of a Cartesian product. The facets used for these classifications induce, one by one, partitions into the MDS space if they are empirically valid. The facets themselves are often based on theoretical considerations, but they can also be attributes that the objects possess by construction. Figure 3 shows an example. Here, (symmetrized) confusion probabilities of 36 Morse signals are represented as distances of a 2-dimensional MDS configuration. The space is partitioned by dashed lines into five regions that contain signals with only short beeps (coded as 1's); signals with more short than long (coded as 2's) beeps; etc. The solid lines cut the space into ten regions that each contain signals with equal duration (0.15 seconds to 0.95 seconds).



Multidimensional Scaling. Fig. 3 Exploratory MDS for confusion probabilities of 36 Morse signals



Multidimensional Scaling. Fig. 4 Confirmatory MDS for the Morse signals, enforcing linearized regions

The solution in Fig. 3 is found by *exploratory* ordinal MDS. There also exist various methods for *confirmatory* MDS that impose additional external constraints onto the MDS model. Figure 4 shows an example of an ordinal MDS with the additional constraint $\mathbf{X}$=$\mathbf{YC}$,

where **Y** is a 36 × 2 matrix of composition and duration codes, respectively, assigned to the 36 Morse signals; **C** is an unknown matrix of weights that re-scales **Y**'s columns monotonically. The confirmatory MDS procedure optimally represents the proximities in the sense of ordinal MDS while satisfying **X**=**YC**. The resulting configuration linearizes the regions of the MDS configuration which makes the solution easier to interpret. Provided its Stress is still acceptable, this is the preferred MDS representation, because it reflects a clear law of formation that is more likely to be replicable than an ad-hoc system of regions. Many alternative side constraints are conceivable. For example, an obvious modification is to require that **C** is diagonal. This enforces an orthogonal lattice of partitioning lines onto the solution in Fig. 4.

Many computer programs exist for doing MDS (for an overview, see Borg and Groenen (2005)). All large statistics packages offer MDS modules. One of the most flexible programs is Proxscal, one of the two MDS modules in Spss. The Spss package also offers Prefscal, a powerful program for unfolding. For R, De Leeuw and Mair (2009) have written a comprehensive MDS program called Smacof which can be freely downloaded from http://CRAN.R-project.org.

About the Author

Dr Ingwer Borg is Professor of Applied Psychological Methods at the University of Giessen (Giessen, Germany), and Scientific Director of the Department of Survey Design & Methodology at GESIS (Mannheim, Germany). He is Past President of the Facet Theory Association and of the International Society for the Study of Work and Organizational Values. He has published some 170 papers and 17 books, including *Modern Multidimensional Scaling* (with Patrick Groenen, Springer, 2005).

Cross References

►Data Analysis
►Distance Measures
►Multidimensional Scaling: An Introduction
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Sensometrics

References and Further Reading

Borg I, Groenen PJF (2005) Modern multidimensional scaling, 2nd edn. Springer, New York

Borg I, Shye S (1995) Facet theory: form and content. Sage, Newbury Park

Carroll JD, Chang JJ (1970) Analysis of individual differences in multidimensional scaling via an *N*-way generalization of 'Eckart-Young' decomposition. Psychometrika 35:283–320

Carroll JD, Wish M (1974) Multidimensional perceptual models and measurement methods. In: Carterette EC, Friedman MP (eds) Handbook of perception. Academic, New York, pp 391–447

Kruskal JB, Wish M (1978) Multidimensional scaling. Sage, Beverly Hills

Lingoes JC, Borg I (1978) A direct approach to individual differences scaling using increasingly complex transformations. Psychometrika, 43:491–519

Wish M (1971) Individual differences in perceptions and preferences among nations. In: King CW, Tigert D (eds) Attitude research reaches new heights. American Marketing Association, Chicago

Multidimensional Scaling: An Introduction

Nataša Kurnoga Živadinović
Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

►Multidimensional scaling (MDS), also called perceptual mapping, is based on the comparison of objects (persons, products, companies, services, ideas, etc.). The purpose of MDS is to identify the relationships between objects and to represent them in geometrical form. MDS is a set of procedures that allows the researcher to map distances between objects in a multidimensional space into a lower-dimensional space in order to show how the objects are related.

MDS was introduced by Torgerson (1952). It has its origins in psychology where it was used to understand respondents' opinions on similarities or dissimilarities between objects. MDS is also used in marketing, management, finance, sociology, information science, political science, physics, biology, ecology, etc. For example, it can be used to understand the perceptions of respondents, to identify unrecognized dimensions, for segmentation analysis, to position different brands, to position companies, and so on (for descriptions of various examples, see Borg and Groenen 2005 and Hair et al. 2010).

MDS starts from the proximities between the objects that express the similarity between them. There are different types of MDS: metric MDS (the similarities data are quantitative; input and output matrices are metric) and nonmetric MDS (the similarities data are qualitative; input matrix is nonmetric).

The steps involved in conducting MDS consist of problem formulation, selection of MDS procedure, determination of the number of dimensions, interpretation, and

validation. Problem formulation includes several tasks. First, the objectives of MDS should be identified. The nature of the variables to be included in MDS should be specified. Also, an appropriate number of variables should be chosen as the number of variables influences the resulting solution. The selection of MDS procedure depends on the nature of the input data (metric or nonmetric). Nonmetric MDS procedures assume that the input data is ordinal, but the resulting output is metric. Metric MDS procedures assume that both input and output data are metric. MDS procedures estimate the relative position of each object in a multidimensional space. The researcher must decide on a number of dimensions. The objective is to achieve an MDS solution that best fits the data in the smallest number of dimensions. Though the fit improves as the number of dimensions increases, the interpretation becomes more complicated. The interpretation of the dimensions and the configuration require subjective judgment, including some elements of judgment on the part of both the researcher and the respondent. The objectives of MDS are not achieved if an appropriate interpretation is lacking. Ultimately, the researcher must consider the quality of the MDS solution. (For detailed descriptions of MDS steps, see Cox and Cox 2001, Hair et al. 2010, and Kruskal and Wish 1978.)

To apply MDS, the distances between objects must first be calculated. The Euclidean distance is the most commonly used distance measure. The distance between objects A and B is given by $d_{AB} = \sqrt{\sum_{i=1}^{v}(x_{Ai} - x_{Bi})^2}$. MDS begins with a matrix $(n \times n)$ consisting of the distances between objects. From the calculated distances, a graph showing the relationship among objects is constructed.

The graphical representation used in MDS is a perceptual map, also called a spatial map. It represents the respondent's perceptions of objectives and shows the relative positioning of all analyzed objects. Let us suppose that there are five objects, A, B, C, D, and E. If objects A and B are judged by the respondents as most similar in comparison to all other pairs of objects (AC, AD, AE, BC, BD, etc.), the MDS procedures will position the objects A and B so that their distance is smaller than the distance of any other two objects. A perceptual map is constructed in two or more dimensions. In a two-dimensional map, objects are represented by points on a plane. In the case of a higher number of dimensions, graphical representation becomes more complicated.

MDS can be conducted at the individual or group level. At the individual level, perceptual maps should be constructed on a respondent-by-respondent base. At the group level, the average judgment of all respondents within a group should be established and the perceptual maps of one or more groups constructed.

Statistical packages such as statistical analysis system (SAS), statistical package for the social sciences (SPSS), Stata, and STATISTICA are suitable for MDS.

Methods closely related to MDS are factor analysis (see ►Factor Analysis and Latent Variable Modelling), ►correspondence analysis, and cluster analysis (see Borg and Groenen 2005, Hair et al. 2010; see also the entry ►Cluster Analysis: An Introduction).

Cross References

►Data Analysis
►Distance Measures
►Multidimensional Scaling
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis

References and Further Reading

Borg I, Groenen PJF (2005) Modern multidimensional scaling: theory and applications. Springer Series in Statistics. 2nd edn. Springer, New York

Cox TF, Cox AA (2001) Multidimensional scaling, 2nd edn. Chapman and Hall/CRC, Boca Raton

Hair JF, Black WC, Babin BJ, Anderson RE (2010) Multivariate data analysis: a global perspective, 7th edn. Pearson Education, Upper Saddle River

Kruskal JB, Wish M (1978) Multidimensional scaling. SAGE University Paper Series: Quantitative Applications in the Social Sciences. SAGE, Newbury Park

Torgerson WS (1952) Multidimensional scaling: I. Theory and method. Psychometrika, 17(4):401–419

Multilevel Analysis

Tom A. B. Snijders
Professor of Statistics
University of Oxford, Oxford, UK
Professor of Methodology and Statistics, Faculty of Behavioral and Social Sciences
University of Groningen, Groningen, Netherlands

Multilevel Analysis, Hierarchical Linear Models

The term "Multilevel Analysis" is mostly used interchangeably with "Hierarchical Linear Modeling," although strictly speaking these terms are distinct. Multilevel Analysis may be understood to refer broadly to the methodology of

research questions and data structures that involve more than one type of unit. This originated in studies involving several levels of aggregation, such as individuals and counties, or pupils, classrooms, and schools. Starting with Robinson's (1950) discussion of the *ecological fallacy*, where associations between variables at one level of aggregation are mistakenly regarded as evidence for associations at a different aggregation level (see Alker 1969, for an extensive review), this led to interest in how to analyze data including several aggregation levels. This situation arises as a matter of course in educational research, and studies of the contributions made by different sources of variation such as students, teachers, classroom composition, school organization, etc., were seminal in the development of statistical methodology in the 1980s (see the review in Chap. 1 of de Leeuw and Meijer 2008). The basic idea is that studying the simultaneous effects of variables at the levels of students, teachers, classrooms, etc., on student achievement requires the use of regression-type models that comprise error terms for each of those levels separately; this is similar to mixed effects models studied in the traditional linear models literature such as Scheffé (1959).

The prototypical statistical model that expresses this is the *Hierarchical Linear Model*, which is a mixed effects regression model for nested designs. In the two-level situation – applicable, e.g., to a study of students in classrooms – it can be expressed as follows. The more detailed level (students) is called the lower level, or level 1; the grouping level (classrooms) is called the higher level, or level 2. Highlighting the distinction with regular regression models, the terminology speaks of *units* rather than cases, and there are specific types of unit at each level. In our example, the level-1 units, students, are denoted by i and the level-2 units, classrooms, by j. Level-1 units are nested in level-2 units (each student is a member of exactly one classroom) and the data structure is allowed to be unbalanced, such that j runs from 1 to N while i runs, for a given j, from 1 to n_j. The basic two-level hierarchical linear model can be expressed as

$$Y_{ij} = \beta_0 + \sum_{h=1}^{r} \beta_h x_{hij} + U_{0j} + \sum_{h=1}^{p} U_{hj} z_{hij} + R_{ij}; \qquad (1a)$$

or, more succinctly, as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\ \mathbf{U} + \mathbf{R}. \qquad (1b)$$

Here Y_{ij} is the dependent variable, defined for level-1 unit i within level-2 unit j; the variables x_{hij} and z_{hij} are the explanatory variables. Variables R_{ij} are residual terms, or error terms, at level 1, while U_{hj} for $h = 0, \ldots, p$ are residual terms, or error terms, at level 2. In the case $p = 0$ this is called a *random intercept model*, for $p \geq 1$ it is called a *random slope model*. The usual assumption is that all R_{ij} and all vectors $U_j = (U_{0j}, \ldots, U_{pj})$ are independent, R_{ij} having a normal $\mathcal{N}(0, \sigma^2)$ and U_j having a multivariate normal $\mathcal{N}_{p+1}(\mathbf{0}, \mathbf{T})$ distribution. Parameters β_h are regression coefficients (fixed effects), while the U_{hj} are random effects. The presence of both of these makes (1) into a mixed linear model. In most practical cases, the variables with random effects are a subset of the variables with fixed effects ($x_{hij} = z_{hij}$ for $h \leq p$; $p \leq r$), but this is not necessary.

More Than Two Levels

This model can be extended to a three- or more-level model for data with three or more nested levels by including random effects at each of these levels. For example, for a three level structure where level-3 units are denoted by $k = 1, \ldots, M$, level-2 units by $j = 1, \ldots, N_k$, and level-1 units by $i = 1, \ldots, n_{ij}$, the model is

$$Y_{ijk} = \beta_0 + \sum_{h=1}^{r} \beta_h x_{hijk} + U_{0jk} + \sum_{h=1}^{p} U_{hjk} z_{hijk} + V_{0k} + \sum_{h=1}^{q} V_{hk} w_{hijk} + R_{ijk}, \qquad (2)$$

where the U_{hjk} are the random effects at level 2, while the V_{hk} are the random effects at level 3. An example is research into outcome variables Y_{ijk} of students (i) nested in classrooms (j) nested in schools (k), and the presence of error terms at all three levels provides a basis for testing effects of pupil variables, classroom or teacher variables, as well as school variables.

The development both of inferential methods and of applications was oriented first to this type of nested models, but much interest now is given also to the more general case where the restriction of nested random effects is dropped. In this sense, multilevel analysis refers to methodology of research questions and data structures that involve several sources of variation – each type of units then refers to a specific source of variation, with or without nesting. In social science applications this can be fruitfully applied to research questions in which different types of *actor* and *context* are involved; e.g., patients, doctors, hospitals, and insurance companies in health-related research; or students, teachers, schools, and neighborhoods in educational research. The word "level" then is used for such a type of units. Given the use of random effects, the most natural applications are those where each "level" is associated with some population of units.

Longitudinal Studies

A special area of application of multilevel models is longitudinal studies, in which the lowest level corresponds to repeated observations of the level-two units. Often the level-two units are individuals, but these may also be organizations, countries, etc. This application of mixed effects models was pioneered by Laird and Ware (1982). An important advantage of the hierarchical linear model over other statistical models for longitudinal data is the possibility to obtain parameter estimates and tests also under highly unbalanced situations, where the number of observations per individual, and the time points where they are measured, are different between individuals. Another advantage is the possibility of seamless integration with nesting if individuals within higher-level units.

Model Specification

The usual considerations for model specification in linear models apply here, too, but additional considerations arise from the presence in the model of the random effects and the data structure being nested or having multiple types of unit in some other way. An important practical issue is to avoid the ecological fallacy mentioned above; i.e., to attribute fixed effects to the correct level. In the original paper by Robinson (1950), one of the examples was about the correlation between literacy and ethnic background as measured in the USA in the 1930s, computed as a correlation at the individual level, or at the level of averages for large geographical regions. The correlation was .203 between individuals, and .946 between regions, illustrating how widely different correlations at different levels of aggregation may be.

Consider a two-level model (1) where variable X_1 with values x_{1ij} is defined as a level-1 variable – literacy in Robinson's example. For "level-2 units" we also use the term "groups." To avoid the ecological fallacy, one will have to include a relevant level-2 variable that reflects the composition of the level-2 units with respect to variable X_1. The mostly used composition variable is the group mean of X_1,

$$\bar{x}_{1.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{1ij}.$$

The usual procedure then is to include x_{1ij} as well as $\bar{x}_{1.j}$ among the explanatory variables with fixed effects. This allows separate estimation of the within-group regression (the coefficient of x_{1ij}) and the between-group regression (the sum of the coefficients of x_{1ij} and $\bar{x}_{1.j}$).

In some cases, notably in many economic studies (see Greene 2003), researchers are interested especially in the within-group regression coefficients, and wish to control for the possibility of unmeasured heterogeneity between the groups. If there is no interest in the between-group regression coefficients one may use a model with fixed effects for all the groups: in the simplest case this is

$$Y_{ij} = \beta_0 + \sum_{h=1}^{r} \beta_h x_{hij} + \gamma_j + R_{ij}. \qquad (3)$$

The parameters γ_j (which here have to be restricted, e.g., to have a mean 0 in order to achieve identifiability) then represent all differences between the level-two units, as far as these differences apply as a constant additive term to all level-1 units within the group. For example in the case of longitudinal studies where level-2 units are individuals and a linear model is used, this will represent all time-constant differences between individuals. Note that (3) is a linear model with only one error term.

Model (1) implies the distribution

$$\mathbf{y} \sim \mathcal{N}_p \left(\mathbf{X}\beta, \mathbf{Z}\,\mathbf{T}\,\mathbf{Z}' + \sigma^2 I \right).$$

Generalizations are possible where the level-1 residual terms R_{ij} are not i.i.d.; they can be heteroscedastic, have time-series dependence, etc. The specification of the variables Z having random effects is crucial to obtain a well-fitting model. See Chap. 9 of Snijders and Bosker (1999), Chap. 9 of Raudenbush and Bryk (2002), and Chap. 3 of de Leeuw and Meijer (2008).

Inference

A major reason for the take-off of multilevel analysis in the 1980s was the development of algorithms for maximum likelihood estimation for unbalanced nested designs. The EM algorithm (Dempster et al. 1981), Iteratively Reweighted Least Squares (Goldstein 1986), and Fisher Scoring (Longford 1987) were applied to obtain ML estimates for hierarchical linear models. The MCMC implementation of Bayesian procedures has proved very useful for a large variety of more complex multilevel models, both for non-nested random effects and for generalized linear mixed models; see Browne and Draper (2000) and Chap. 2 of de Leeuw and Meijer (2008).

Hypothesis tests for the fixed coefficients β_h can be carried out by Wald or Likelihood Ratio tests in the usual way. For testing parameters of the random effects, some care must be taken because the estimates of the random effect variances τ^2_{hh} (the diagonal elements of $\mathbf{T}$) are not approximately normally distributed if $\tau^2_{hh} = 0$. Tests for these parameters can be based on estimated fixed effects, using least squares estimates for U_{hj} in a specification where these are treated as fixed effects (Bryk and Raudenbush 2002, Chap. 3); based on appropriate distributions of the log likelihood ratio; or obtained as score tests (Berkhof and Snijders 2001).

About the Author

Professor Snijders is Elected Member of the European Academy of Sociology (2006) and Elected Correspondent of the Royal Netherlands Academy of Arts and Sciences (2007). He was awarded the Order of Knight of the Netherlands Lion (2008). Professor Snijders was Chairman of the Department of Statistics, Measurement Theory, and Information Technology, of the University of Groningen (1997–2000). He has supervised 52 Ph.D. students. He has been associate editor of various journals, and Editor of *Statistica Neerlandica* (1986–1990). Currently he is co-editor of *Social Networks*, Associate editor of *Annals of Applied Statistics*, and Associate editor of *Journal of Social Structure*. Professor Snijders has (co-)authored about 100 refereed papers and several books, including *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. (with Bosker, R.J., London etc.: Sage Publications, 1999). In 2005, he was awarded an honorary doctorate in the Social Sciences from the University of Stockholm.

Cross References

►Bayesian Statistics
►Cross Classified and Multiple Membership Multilevel Models
►Mixed Membership Models
►Moderating and Mediating Variables in Psychological Research
►Nonlinear Mixed Effects Models
►Research Designs
►Statistical Analysis of Longitudinal and Correlated Data
►Statistical Inference in Ecology

References and Further Reading

To explore current research activities and to obtain information training materials etc., visit the website www.cmm.bristol.ac.uk. There is also an on-line discussion group at www.jiscmail.ac.uk/lists/multilevel.html.

There is a variety of textbooks, such as Goldstein (2003), Longford (1993), Raudenbush and Bryk (2003), and Snijders and Bosker (1999). A wealth of material is contained in de Leeuw and Meijer (2008).

Alker HR (1969) A typology of ecological fallacies. In: Dogan M, Rokkan S (eds) Quantitative ecological analysis in the social sciences. MIT Press, Cambridge, pp 69–86

Berkhof J, Snijders TAB (2001) Variance component testing in multilevel models. J Educ Behav Stat 26:133–152

Browne WJ, Draper D (2000) Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. Computational Stat 15:391–420

de Leeuw J, Meijer E (2008) Handbook of multilevel analysis. Springer, New York

Dempster AP, Rubin DB, Tsutakawa RK (1981) Estimation in covariance components models. J Am Stat Assoc 76:341–353

Goldstein H (1986) Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 73:43–56

Goldstein H (2003) Multilevel statistical models, 3rd edn. Edward Arnold, London

Greene W (2003) Econometric analysis, 5th edn. Prentice Hall, Upper Saddle River

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38:963–974

Longford NT (1987) A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika 74:812–827

Longford NT (1993) Random coefficient models. Oxford University Press, New York

Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods, 2nd edn. Sage, Thousand Oaks

Robinson WS (1950) Ecological correlations and the behavior of individuals. Am Sociol Rev 15:351–357

Scheffé H (1959) The analysis of variance. Wiley, New York

Snijders TAB, Bosker RJ (1999) Multilevel analysis: an introduction to basic and advanced multilevel modeling. Sage, London

Multinomial Distribution

George A. F. Seber
Emeritus Professor of Statistics
Auckland University, Auckland, New Zealand

The Multinomial distribution arises as a model for the following experimental situation. An experiment or "trial" is carried out and the outcome occurs in one of k mutually exclusive categories with probabilities p_i, $i = 1, 2, \ldots, k$. For example, a person may be selected at random from a population of size N and their ABO blood phenotype recorded as A, B, AB, or O ($k = 4$). If the trial is repeated n times such that the trials are mutually independent, and if x_i is the frequency of occurrence in the ith category, then the joint probability function of the x_i is

$$P_1(x_1, x_1, \ldots, x_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

where $\sum_{i=1}^{k} x_i = n$ and $\sum_{1=1}^{k} p_i = 1$. This would be the correct probability function for the genetics example if further people were chosen with replacement. In practice, sampling is without replacement and the correct distribution is the multivariate hypergeometric, a difficult distribution to deal with. Fortunately, all is not lost, as when the sampling fraction $f = n/N$ is small enough (say less than 0.1 or preferably less than 0.05), the Multinomial distribution

is a good approximation and is used extensively in genetics (e.g., Greenwood and Seber 1992). We note that when $k = 2$ we have the ▶Binomial distribution. Also the terms of P_1 can be obtained by expanding $(p_1 + p_2 + \cdots + p_k)^n$.

Various properties of the Multinomial distribution can be derived using extensive algebra. However, they are more readily obtained by noting that any subset of a multinomial distribution is also Multinomial. We simply group the categories relating to the remaining variables into a single category. For example x_i will have a Binomial distribution as there are just two categories, the ith and the rest combined. Hence the mean and variance of x_i are

$$\mathrm{E}(x_i) = np_i \text{ and } \mathrm{var}(x_i) = np_iq_i,$$

where $q_i = 1 - p_i$. Also, if we combine the ith and jth category and then combine the rest into single category, we see that $x_i + x_j$ is Binomial with probability parameter $p_i + p_j$ and variance $n(p_i + p_j)(1 - p_i - p_j)$. Hence the covariance of x_i and x_j is

$$\mathrm{cov}(x_i, x_j) = \frac{1}{2}[\mathrm{var}(x_i + x_j) - \mathrm{var}(x_i) - \mathrm{var}(x_j)] = -np_ip_j.$$

Another useful result that arises in comparing proportions p_i and p_j in a ▶questionnaire is

$$\begin{aligned}\mathrm{var}(x_i - x_j) &= \mathrm{var}(x_i) + \mathrm{var}(x_j) - 2\mathrm{cov}(x_i, x_j)\\ &= n[p_i + p_j - (p_i - p_j)^2]. \end{aligned} \quad (1)$$

It should be noted that the Multinomial distribution given above is a "singular" distribution as the random variables satisfy the linear constraint $\sum_{i=1}^{k} x_i = n$, which leads to a singular variance-covariance matrix. We can instead use the "nonsingular" version

$$P_2(x_1, x_2, \ldots, x_{k-1}) = \frac{n!}{x_1!x_2!\cdots\left(n - \sum_{i=1}^{k-1} x_i\right)!} \times p_1^{x_1}p_2^{x_2}\cdots p_k^{n-\sum_{i=1}^{k-1} x_i}.$$

We note that the joint ▶moment generating function of $\boldsymbol{x}$ is

$$M(\boldsymbol{t}) = (p_1e^{t_1} + p_2e^{t_2} + \cdots + p_{k-1}e^{t_{k-1}} + p_k)^n,$$

which can also be used to derive the above properties of the Multinomial distribution as well as the ▶asymptotic normality properties described next.

Let $\hat{p}_i = x_i/n$ be the usual estimate of p_i. Given the vectors $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{k-1})'$ and $\boldsymbol{p} = (p_1, p_2, \ldots, p_{k-1})'$, then the mean of $\hat{\boldsymbol{p}}$ is $\boldsymbol{p}$ and its variance-covariance matrix is $n^{-1}\boldsymbol{V}$, where $\boldsymbol{V} = (\mathrm{diag}\,\boldsymbol{p} - \boldsymbol{p}\boldsymbol{p}')$ and $\mathrm{diag}\,\boldsymbol{p}$ is a diagonal matrix with diagonal elements $p_1, p_2, \ldots p_{k-1}$. In the same way that a Binomial random variable is asymptotically normal for large n, $\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p})$ is asymptotically multivariate Normal with mean vector 0 and variance-covariance matrix $\boldsymbol{V}$. If $\boldsymbol{V}^{-1}$ is the inverse of $\boldsymbol{V}$, then $\boldsymbol{V}^{-1} = n^{-1}\left((\mathrm{diag}\,\boldsymbol{p})^{-1} + p_k^{-1}1_{k-1}'1_{k-1}\right)$, where 1_{k-1} is a column $k-1$ ones (cf. Seber, 2008, 15.7). From the properties of the multivariate Normal distribution (cf. Seber 2008, 20.25),

$$n(\hat{\boldsymbol{p}} - \boldsymbol{p})'\boldsymbol{V}^{-1}(\hat{\boldsymbol{p}} - \boldsymbol{p}) = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i} \quad (2)$$

will be asymptotically distributed as the ▶Chi-square distribution with $k-1$ degrees of freedom. If we use the singular version and include x_k to expand $\boldsymbol{V}$ to $\boldsymbol{V}_k$, we can obtain the result more quickly using a generalized inverse (cf. Seber, 2008, 20.29b using $\boldsymbol{A} = \boldsymbol{V}_k^{-} = (\mathrm{diag}\,(\boldsymbol{p}', p_k)')^{-1}$). This link with the Chi-square distribution forms the basis of a number of tests involving the Multinomial distribution mentioned below.

We see that $P_1(\cdot)$ above can be regarded conceptually as a nonsingular distribution for the x_i $(i = 1, 2, \ldots, k)$ with probabilities π_i, but conditional on $\sum_{i=1}^{k} x_i = n$ with $p_i = \pi_i / \sum_{i=1}^{k} \pi_i$. It therefore follows that the joint distribution of any subset of multinomial variables conditional on their sum is also multinomial. For example, the distribution of x_1 and x_2 given $x_1 + x_2 = n$ is Binomial with probability parameter $p_1/(p_1 + p_2)$. We get a similar result in ecology where we have a population of plants divided up into k areas with x_i in the ith area being distributed as the Poisson distribution with mean μ_i. If the x_i are mutually independent, then the joint distribution of the x_i conditional on the sum $\sum_{i=1}^{k} x_i$ is Multinomial with probabilities $p_i = \mu_i / \sum_{j=1}^{k} \mu_j$.

The last topic I want to consider briefly is inference for the multinomial distribution. Estimating p_i by $\hat{p}_i = x_i/n$, using the normal approximation, and applying (1), we can obtain a confidence interval for any particular p_i or any particular difference $p_i - p_j$. Simultaneous confidence interval procedures are also available for all the p_i or all differences using the Bonferroni method. We can also test $\boldsymbol{p} = \boldsymbol{p}_0$ using (2).

A common problem is testing the hypothesis H_0 : $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta})$, where $\boldsymbol{p}$ is a known function of some unknown t-dimensional parameter $\boldsymbol{\theta}$ (e.g., the genetics example above). This can be done using a derivation like the one that led to (2) above, giving the so-called "goodness of fit" statistic, but with $\boldsymbol{p}$ replaced by $\boldsymbol{p}(\hat{\boldsymbol{\theta}})$. Here $\hat{\boldsymbol{\theta}}$, the maximum likelihood estimate of $\boldsymbol{\theta}$, is asymptotically Normal so that $\boldsymbol{p}(\hat{\boldsymbol{\theta}})$ is also asymptotically Normal. Under H_0, it can be shown that the test statistic is approximately Chi-square with degrees of freedom now $k - 1 - t$.

One application of the above is to the theory of contingency tables. We have an $r \times c$ table of observations x_{ij}

$(i = 1, 2, \ldots, r; j = 1, 2, \ldots, c)$ and p_{ij} is the probability of falling in the (i, j)th category. Treating the whole array as a single Multinomial distribution, one hypothesis of interest is $H_0 : p_{ij} = \alpha_i \beta_j$, where $\sum_{i=1}^{r} \alpha_i = 1$ and $\sum_{j=1}^{c} \beta_j = 1$. In this hypothesis of row and column independence, we have $\boldsymbol{\theta}' = (\alpha_1, \ldots \alpha_{r-1}, \beta_1, \ldots, \beta_{c-1})$ with maximum likelihood estimates $\hat{\alpha}_i = R_i/n$ and $\hat{\beta}_j = c_j/n$, where r_i is the ith row sum of the table and c_j the jth column sum. The statistic for the test of independence is therefore

$$\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_i - r_i c_j/n)^2}{r_i c_j/n}, \tag{3}$$

which, under H_0, is approximately Chi-square with $rc - 1 - (r-1) - (c-1) = (r-1)(c-1)$ degrees of freedom. If the rows of the $r \times c$ table now represents r independent Multinomial distributions with $\sum_{j=1}^{c} p_{ij} = 1$ for $i = 1, 2, \ldots, r$, then the hypothesis that the distributions are identical is $H_0 : p_{ij} = \gamma_j$ for $i = 1, 2, \ldots, r$, where $\sum_{j=1}^{c} \gamma_j = 1$. Pooling the common distributions, the maximum likelihood estimate of γ_j is $\hat{\gamma}_j = c_j/n$ so that the term $np_{ij}(\hat{\theta})$ becomes $r_i \hat{\gamma}_j$ and the test statistic for testing homogeneity turns out to be the same as (3) with the same degrees of freedom.

The above chi-squared tests are not particularly powerful and need to be backed up with various confidence interval procedures. Other asymptotically equivalent tests are the likelihood ratio test and the so-called "score" (Lagrange multiplier) test. Log linear models can also be used. For further properties of the Multinomial distribution see Johnson et al. (1997, Chap. 35) and asymptotic background theory for the chi-squared tests is given by Bishop et al. (1975, Chap. 14). More recent developments are given by Agresti (2002).

About the Author

For biography *see* the entry ►Adaptive Sampling.

Cross References

►Binomial Distribution
►Categorical Data Analysis
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Divisible Statistics
►Entropy and Cross Entropy as Diversity and Distance Measures
►Geometric and Negative Binomial Distributions
►Multivariate Statistical Distributions
►Statistical Distributions: An Overview

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, New York
Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge
Greenwood SR, Seber GAF (1992) Estimating blood phenotypes probabilities and their products. Biometrics 48:143–154
Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Wiley, New York
Seber GAF (2008) A matrix handbook for statisticians. Wiley, New York

Multi-Party Inference and Uncongeniality

Xiao-Li Meng
Professor, Chair
Harvard University, Cambridge, MA, USA

► *"Life is more complicated when you have three uncongenial models involved."*

The Multi-Party Inference Reality

Much of the statistical inference literature uses the familiar framework of "God's model versus my model." That is, an unknown model, "God's model," generates our data, and our job is to infer this model or at least some of its characteristics (e.g., moments, distributional shape) or implications (e.g., prediction). We first postulate one or several models, and then use an array of estimation, testing, selection, and refinement methods to settle on a model that we judge to be acceptable – according to some sensible criterion, hopefully pre-determined – for the inference goals at hand, even though we almost never can be sure that our chosen model resembles God's model in critical ways. Indeed, philosophically even the existence of God's model is not a universally accepted concept, just as theologically the existence of God is not an unchallenged notion.

Whether one does or does not adopt the notion of God's model, it is repeatedly emphasized in the literature that to select a reasonable model, an iterative process is necessary and hence multiple models are typically considered (e.g., see Box and Tiao 1973, Chap. 1; Gelman and Meng 1996). By *multiple models* we mean multiple sets of mathematically quantifiable assumptions (hence, not necessarily parametric models), which are compatible within each set but not across different sets. Indeed, if they are not incompatible across different sets then one is simply postulating a larger model; see McCullagh (2002). In this

sense we automatically take a "monotheistic" point of view that there is only one God's model; we assume God's model contains no self-contradiction (or at least none detectable by a human modeler). However, we do not go so far as to suggest that the modeler can always embed everything into one model, e.g., as in Bayesian model averaging, because contrasting models sometimes is as useful as, if not more so than, combining models.

Whereas many models may be entertained, the commonly accepted paradigm involves only two parties: the (hypothetical) God, and "me" – the modeler. Unfortunately, reality is far more complicated. To explain the complication, we must distinguish the *modeler's data* from *God's data*. The modeler's data are the data available to the modeler, whereas God's data are the realizations from God's model that the modeler's data were collected to *approximate*. Whereas any attempt to mathematically define such concepts is doomed to fail, it is useful to distinguish the two forms of data because the *approximation* process introduces an additional inference party (or parties).

For example, in the physical sciences, the modeler's data typically are results of a series of pre-processing steps to deal with limitations or irregularities in recording God's data (e.g., discarding "outliers" (see ►Outliers); recalibration to account for instrument drift), and typically the modeler at best only has partial information about this process. For the social and behavioral sciences, some variables are not even what we normally think they are, such as responses to a questionnaire survey. Rather, they are so-called "constructed variables," typically from a deterministic algorithm converting a set of answers to an index that indicates, say, whether a subject is considered to suffer major depression. The algorithm is often a black box, and in some cases it is pitch black because the modeler is not even informed of what variables were used as inputs to produce the output. In the context of public-use data files, virtually all data sets contain imputations of some sort (see ►Imputation) because of non-responses or other forms of missing data (e.g., missingness by design such as with matrix sampling), which means someone has "fixed the holes" in the data before they reach the modeler.

In all these examples, the key issue is not that there is data pre-processing step per se, but rather that during the journey from God's data to modeler's data, a set of assumptions has been introduced. There is no such thing as "assumption-free" pre-processing; any attempt to make the data "better" or "more usable" implies that a judgment has been made. Under the God-vs.-me paradigm, this intermediate "data cleaning" process has to be considered either as part of God's model, or of the modeler's model, or of both by somehow separating aspects of the process (e.g., one could argue that a refused answer to an opinion question is an opinion itself, whereas a refusal to an income question is a non-response). Regardless of how we conceptualize, we find ourselves in an extremely muddy – if not hopeless – situation. For example, if aspects of this intermediate process are considered to be part of God's model, then the modeler's inference is not just about God's model but also about someone else's assumptions about it. If we relegate the pre-processing to the modeler's model, then the modeler will need good information on the process. Whereas there has been an increasing emphasis on understanding the entire mechanism that leads to the modeler's data, the reality is that for the vast majority of real-life data sets, especially large-scale ones, it is simply impossible to trace back how the data were collected or pre-processed. Indeed, many such processes are nowhere documented, and some are even protected by confidentiality constraints (e.g., confidential information may be used for imputation by a governmental agency).

This intermediate "data cleaning" process motivates the *multi-party inference* paradigm. The term is self-explanatory: we acknowledge that there is more than one party involved in reaching the final inference. The key distinction between the multi-party paradigm and the God-vs.-me paradigm is not that the former involves more sets of assumptions, i.e., models – indeed under the latter we still almost always (should) consider multiple models. Rather, in the multi-party paradigm, we explicitly acknowledge the *sequential nature* of the parties' involvement, highlighted by how the intermediate party's assumptions impact the final inference, because typically they are necessarily incompatible with the modeler's assumptions, due both to the parties' having access to different amounts of information and to their having different objectives.

This situation is most vividly demonstrated by multiple imputation inference (Rubin 1987), where the intermediate party is the imputer. (There is often more than one intermediate party even in the imputation context, but the case of a single imputer suffices to reveal major issues.) In such a setting, the concept of *congeniality* (Meng 1994) is critical. In a nutshell, congeniality means that the imputation model and the analysis model are compatible for the purposes of predicting the missing data. In real life, this typically is not the case, even if the imputer and analyst are the same entity, because of the different aims of imputation (where one wants to use as many variables as possible even if causal directions are incorrectly specified) and of analysis (where one may be only interested in a subset of variables with specified causal directions). The next section demonstrates the importance

of recognizing *uncongeniality*, which directly affects the validity of the final inferences. The concept of uncongeniality was originally defined and has thus far been investigated in the context of multiple imputation inference, the most well-studied case of multi-party inference. However, its general implication is broad: to reach valid inference when more than one party is involved, we must consider the incompatibility/uncongeniality among their assumptions/models, even if each party has made assumptions that are consistent with God's model and has carried out its task in the best possible way given the information available at the time.

Uncongeniality in Multiple Imputation Inference

A common method for dealing with non-response in surveys and incomplete data in general is imputation (Little and Rubin 2002). Briefly, imputation is a prediction of the missing data from a posited (not necessarily parametric) model $p_I(Y_{mis}|Y_{obs})$, where Y_{mis} denotes the missing data and Y_{obs} the observed data. The trouble with single imputation, however sophisticated, is that the resulting data set cannot be analyzed in the same way as would an authentic complete data set, without sacrificing the validity of the inference. Multiple imputation (MI; Rubin 1987) attempts to circumvent this problem by providing multiple predictions from $p_I(Y_{mis}|Y_{obs})$, thereby permitting, via genuine replications, a direct assessment of uncertainties due to imputation.

Specifically, in the MI framework, we draw independently m times from $p_I(Y_{mis}|Y_{obs})$, resulting in m completed-data sets: $Y_{com}^{(\ell)} = \left\{Y_{obs}, Y_{mis}^{(\ell)}\right\}, \quad \ell = 1, \ldots, m.$ Suppose our complete-data analysis can be summarized by a point estimator $\hat{\theta}(Y_{com})$ and an associated variance estimator $U(Y_{com})$, where Y_{com} denotes $\{Y_{mis}, Y_{obs}\}$. The MI inference procedure consists of the following steps:

Step 1: Perform m complete-data analyses as if each $Y_{com}^{(\ell)}$ were real data:

$$\hat{\theta}_\ell \equiv \hat{\theta}\left(Y_{com}^{(\ell)}\right), \text{ and } U_\ell \equiv U\left(Y_{com}^{(\ell)}\right), \quad \ell = 1, \ldots, m.$$

Step 2: Use Rubin's Combining Rules:

$$\bar{\theta}_m = \frac{1}{m}\sum_{\ell=1}^{m} \hat{\theta}_\ell, \text{ and } T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m,$$

where

$$\bar{U}_m = \frac{1}{m}\sum_{\ell=1}^{m} U_\ell \text{ and } B_m = \frac{1}{m-1}\sum_{\ell=1}^{m} (\hat{\theta}_\ell - \bar{\theta}_m)(\hat{\theta}_\ell - \bar{\theta}_m)^\top$$

are respectively the within-imputation variance and the between-imputation variance, to reach the MI inference $\{\bar{\theta}_m, T_m\}$, with T_m the variance estimator of $\bar{\theta}_m$.

The justification of Rubin's combining rules is most straightforward under strict congeniality, which means that both the analyst and the imputer use (effectively) Bayesian models, and their Bayesian models are compatible. That is, we assume:

(I) The complete-data analysis procedure can be embedded into a Bayesian model, with

$$\hat{\theta}(Y_{com}) = E_A(\theta|Y_{com}) \text{ and } U(Y_{com}) = \mathrm{V}_A(\theta|Y_{com}),$$

where the subscript A indexes expectation with respect to the embedded analysis model;

(II) The imputer's model and the (embedded) analysis model are the same for the purposes of predicting missing data:

$$P_I(Y_{mis}|Y_{obs}) = P_A(Y_{mis}|Y_{obs}), \quad \textit{for all } Y_{mis} \textit{ (but the given } Y_{obs}).$$

Then for $\bar{\theta}_m$ as $m \to \infty$, we have

$$\begin{aligned}\bar{\theta}_\infty &= \mathrm{E}_I\left[\hat{\theta}(Y_{com})\middle|Y_{obs}\right] \\ &< \text{by (I)} > = \mathrm{E}_I\left[\mathrm{E}_A(\theta|Y_{com})\middle|Y_{obs}\right] \\ &< \text{by (II)} > = \mathrm{E}_A\left[\mathrm{E}_A(\theta|Y_{com})\middle|Y_{obs}\right] = \mathrm{E}_A(\theta|Y_{obs}).\end{aligned}$$

That is, the MI estimator $\bar{\theta}_m$ simply is a consistent (Monte Carlo) estimator of the posterior mean under the analyst's model based on the observed data Y_{obs}. The critical role of (II) is also vivid in establishing the validity of $T_m = \bar{U}_m + (1 + m^{-1})B_m$ as $m \to \infty$:

$$\begin{aligned}\bar{U}_\infty + B_\infty &= E_I[U(Y_{com})|Y_{obs}] + \mathrm{V}_I[\hat{\theta}(Y_{com})|Y_{obs}] \\ &< by\ (I) > = E_I[\mathrm{V}_A(\theta|Y_{com})|Y_{obs}] \\ &\quad + \mathrm{V}_I[E_A(\theta|Y_{com})|Y_{obs}] \\ &< by\ (II) > = E_A[\mathrm{V}_A(\theta|Y_{com})|Y_{obs}] \\ &\quad + \mathrm{V}_A[E_A(\theta|Y_{com})|Y_{obs}] = \mathrm{V}_A(\theta|Y_{obs}).\end{aligned}$$

Therefore, as $m \to \infty$, $\{\bar{\theta}_m, T_m\}$ reproduces the posterior mean and posterior variance under the analyst's model given Y_{obs}, because $\bar{\theta}_\infty = E_A(\theta|Y_{obs})$ and $T_\infty = \mathrm{V}_A(\theta|Y_{obs})$.

When congeniality fails, either because the analyst's procedure does not correspond to any Bayesian model or because the corresponding Bayesian model is incompatible with the imputer's model, the MI variance estimator T_m can overestimate or underestimate the variance of $\bar{\theta}_m$ even as $m \to \infty$. However, depending on the relationships

among God's model, the analyst's model and the imputer's model, we may still reach valid inference under uncongeniality. For example, under the assumption that the analyst's complete-data procedure is self-efficient (Meng 1994), if God's model is nested in the analyst's model, which in turn is nested in the imputer's model, then the MI confidence interval based on $\{\bar{\theta}_\infty, T_\infty\}$ is valid (asymptotically with respect to the size of the observed data). However, the MI estimator $\bar{\theta}_\infty$ may not be as efficient as the analyst's estimator (e.g., MLE) directly based on the observed data, because the additional assumptions built into the analysis model are not used by the imputer. But this comparison is immaterial when the analyst is unable to analyze the observed data directly, and therefore multiple imputation inference is needed (see ►Multiple Imputation).

However, the situation becomes more complicated if we assume God's model is nested in the imputer's model, which in turn is nested in the analyst's model. In such cases, it is possible to identify situations where the multiple imputation interval estimator is conservative in its own right, yet it is narrower than analyst's interval estimator (with the correct nominal coverage) directly based on the observed data (Xie and Meng 2010). This seemingly paradoxical phenomenon is due to the fact the imputer has introduced "secret" model assumptions into the MI inference, making it more efficient than the analyst's inference directly based on the observed data, which does not benefit from the imputer's assumptions. At the same time, since the analyst's complete-data procedure $\{\hat{\theta}(Y_{com}), U(Y_{com})\}$ is determined irrespective of the imputer's model, the imputer's secret assumption introduces uncongenality, which leads to the conservativeness of the MI interval. However, this is not to suggest that MI tends to be conservative, but rather to demonstrate the impact of imputation models on the MI inference and hence to provide practical guidelines on how to regulate the imputation models.

Even more complicated are situations where the analyst's and imputer's models do not nest, or where at least one of them does not contain God's model as a sub-model. Consequences of such are virtually undetermined at the present time, but one thing is clear. These complications remind us the importance of recognizing the multi-party inference paradigm, because the God-vs.-me paradigm sweeps all of them under the rug, or more precisely buries our heads in the sand, leaving our posteriors exposed without proper coverage.

Acknowledgment

The author thanks NSF for partial support, and Joseph Blitzstein, Yves Chretien and Xianchao Xie for very helpful comments and proofreading.

About the Author

Dr Xiao-Li Meng started his outstanding career in 1982 as Instructor of Mathematics in China Textile University and 22 years later has become Professor and Chair of Statistics at one of the most prestigious universities in the world, Harvard University (2004–Present), USA. In July 2007 he was appointed as Whipple V.N. Jones Professor of Statistics at his department. In 2001 he was awarded for "the outstanding statistician under the age of forty" by the Committee of Presidents of Statistical Societies. In 2002 he was ranked (by Science Watch) among the world top 25 most cited mathematicians for articles published and cited during 1991–2000. Professor Meng was Editor of *Bayesian Analysis* (2003–2005), and Co-Chair Editor, *Statistica Sinica* (2005–2008). He was an Associate editor for following journals: *Bernoulli* (2004–2005), *Biometrika* (2002–2005), The *Annals of Statistics* (1997–2003), *Journal of the American Statistical Association* (1996–2002) and *Statistica Sinica* (1992–1997). Currently, he is Editor of *Statistics Series, IMS Monograph and Textbook Series*. He is an Elected Fellow of the Institute of Mathematical Statistics (1997) and American Statistical Association (2004). Professor Meng is a recipient of the University of Chicago Faculty Award for Excellence in Graduate Teaching (1997–1998). He has published over 100 papers in leading statistical journals, and is widely known for his contributions in statistical analysis with missing data, Bayesian modeling, statistical computation, in particular Markov chain Monte Carlo and EM-type algorithms. (written by ML)

Cross References

►Data Analysis
►Data Privacy and Confidentiality
►Data Quality (Poor Quality Data: The Fly in the Data Analytics Ointment)
►Imputation
►Model Selection
►Multiple Imputation
►Nonresponse in Surveys

References and Further Reading

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Wiley, New York
Gelman AE, Meng X-L (1996) Model checking and model improvement. In: Gilks W, Richardson S, Spiegelhalter D (eds) Practical Markov chain Monte Carlo, Chapman & Hall, London, pp 189–201
Little R, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
McCullagh P (2002) What is a statistical model? (with discussion). Ann Stat 30:1225–1310

Meng X-L (1994) Multiple-imputation inference with uncongenial sources of input (with discussion). Stat Sci 9: 538–573

Rubin DB (1987) Multiple imputaiton for nonresponse in surveys. Wiley, New York

Xie X, Meng X-L (2010) Multi-party inferences: what happens when there are three uncongenial models involved? Techincal Report, Department of Statistics, Harvard University

Multiple Comparison

Toshihiko Morikawa[1], Takeharu Yamanaka[2]
[1]Former Professor
Kurume University, Kurume, Japan
[2]Section Head
National Kyushu Cancer Center, Fukuoka, Japan

Multiplicity Issues

Statistical evidence is obtained by rejecting the null hypothesis at a "small" prespecified significance level α, say 0.05 or 0.01, which is an acceptable level of probability of the type I error (the error of rejecting the "true" null hypothesis). If we have a family of multiple hypotheses in a confirmatory experiment and test them simultaneously at each level α, the *overall* or *familywise type I error rate* (*FWER*), i.e., the probability of rejecting at least one "true" null hypothesis in the family, may inflate and exceed α, even if there exist no treatment differences. We call such inflation of the *FWER* a *multiplicity issue*.

Usually there may be some correlation structure between test statistics, and the inflation of the *FWER* might not be so remarkable. However, if we have multiple hypotheses to be tested for confirmatory purpose, we should adjust for multiplicity so as to control the *FWER* within α. This is called *multiplicity adjustment*. Testing procedures for multiplicity adjustment are called *multiple comparison procedures* (*MCPs*) or more generally *multiple testing procedures* (*MTPs*).

Multiplicity issues may arise in (1) multiple treatments (multiple comparisons), (2) multiple response variables (multiple endpoints), (3) multiple time points (longitudinal analysis), (4) multiple subgroups (subgroup analysis), and (5) multiple looks (interim analysis with group sequential methods or adaptive designs).

Hereafter we mainly concentrate on the multiple treatment comparisons, i.e., multiple comparisons in a traditional sense.

Multiple Comparisons

In a two group comparison of treatments A and B on their response means μ_A and μ_B, we have just one null hypothesis $H_0 : \mu_A = \mu_B$ to be tested and there is no need to adjust for multiplicity. However, when we compare three treatment groups, e.g., there are three treatments A, B and C, we may typically want to compare their means pairwisely, i.e., μ_A vs μ_B, μ_A vs μ_C and μ_B vs μ_C. Then there are three test hypotheses to be adjusted for multiplicity; namely, we need multiple comparison procedures.

All Pairwise Comparisons

The method to exactly control the *FWER* by adjusting the critical value in the above "all" pairwise comparisons is called *Tukey's method* (or *Tukey's multiple comparison test*). The method was developed for equal sample sizes, but even if the sample sizes are different between groups, the same critical value could be used conservatively, and such a method is known as the *Tukey-Kramer method*. The nonparametric version of Tukey's method is called the *Steel-Dwass test*.

Comparisons with a Control

The above three treatment example may have a structure that A and B are two (high and low) doses of a drug and C is a placebo (zero-dose). Then main interest in a formal analysis may be focused on the comparisons between each active dose and the placebo, i.e., μ_A vs μ_C and μ_B vs μ_C. This type of multiple comparison on treatment means can be performed by *Dunnett's method* (or *Dunnett's multiple comparison test*), and the common reference C is called a *control* or *control group*. The nonparametric version of Dunnett's method is called *Steel's test*.

If we assume the monotonicity of response means, such as $\mu_A \geq \mu_B \geq \mu_C$ or $\mu_A \leq \mu_B \leq \mu_C$, then in the comparison with a control, we can apply the *Williams test*, which is more powerful than Dunnett's test when the monotone dose-response relationship holds. The nonparametric version of the Williams test is known as the *Shirley-Williams test*.

Any Contrast Comparisons

More generally in a $k(\geq 3)$ treatment comparison, various hypotheses on any contrasts, such as, $\sum_{i=1}^{k} c_i \mu_i = 0$ where $\sum_{i=1}^{k} c_i = 0$, can be tested using *Scheffe's method* to control the *FWER*. For all pairwise comparisons or comparisons with a control, Scheffe's method is not recommended because it is "too" conservative in such cases. A nonparametric version of the Scheffe type multiple comparison method can be easily constructed.

Fixed Number of Comparisons

When the number of comparisons is fixed, the *Bonferroni method* (or *Dunn's method*) is simpler and easier to apply. The method only adjusts the significance level to α/m for each single test, where m is the number of interested comparisons. It is known that the method controls the *FWER* because the well-known *Bonferroni inequality*, $Pr\left(\bigcup_{i=1}^{m} E_i\right) \leq \sum_{i=1}^{m} Pr(E_i)$ holds, where E_i is an event to reject hypothesis H_i. In the above three treatment example, the Bonferroni method could be applied with $m = 3$ for Tukey-type, and with $m = 2$ for Dunnett-type multiple comparisons, although it might be rather conservative.

Stepwise Procedures

All the methods described above (except the Williams test) are called "*simultaneous tests*" or "*single step tests*", because none of tests considered are affected by the results of others, and statistical testing for each hypothesis can be done simultaneously or in a single step manner. They control the *FWER* and can be used to easily construct the corresponding simultaneous confidence intervals, but there is some tradeoff in that they have a low statistical power in compensation for controlling the *FWER*.

Recently, more powerful test procedures than single step or simultaneous test procedures have been developed and become popular. Most of them are based on the *closed testing procedure* (*CTP*) proposed by Marcus, Peritz and Gabriel (1976) and they have a stepwise property in their nature. *CTPs* give a very general scheme of stepwise *MCPs* (or *MTPs*).

Closed Testing Procedures (CTPs)

Suppose that we have a family of m null hypotheses $F = \{H_1, H_2, \ldots, H_m\}$ to be tested and let $N = \{1, 2, \ldots, m\}$ be an *index set* that indicates the set of hypotheses considered. Then there are $2^m - 1$ possible intersections of null hypotheses H_i. We denote a set or family of such *intersection hypotheses* by $G = \{H_I = \bigcap_{i \in I} H_i : I \subseteq N,\ I \neq \emptyset\}$, where $\emptyset$ is an empty set and each intersection hypothesis H_I means that all hypotheses $H_i, i \in I$ hold simultaneously and thus H_I represents one possibility of the "true" null hypothesis. Because we do not know which H_I is true, a given *MCP* (or *MTP*) should control the *FWER* under any H_I. This is called a *strong control of the FWER*. If we control the *FWER* only under the *complete* or *global null hypothesis*, $H_N = \bigcap_{i \in N} H_i$, it is called a *weak control of the FWER*.

CTPs are testing procedures in which each *elementary hypothesis* $H_i, i = 1, \ldots, m$, is rejected only if all the intersection hypotheses including H_i, i.e., all $H_I = \bigcap_{j \in I} H_j, i \in I$, are rejected by the *size* α *test*. It is easily shown that any *CTP* controls the *FWER* in a strong sense. The procedure is equivalent to a test that starts with the test of *complete null hypothesis* H_N at level α and then proceeds in a stepwise manner that any *intersection hypothesis* H_I, $I \subset N$, is tested at level α only if all the intersection hypotheses $H_J = \bigcap_{i \in J} H_i$ which *imply* H_I, i.e., $J \supset I$, are rejected.

Some well known stepwise methods for the Tukey type multiple comparisons, e.g., *Fisher's protected LSD* (least significant difference) *test*, the *Newman-Keuls test*, and *Duncan's multiple range test*, control the *FWER* only in a weak sense, and should not be used. Instead, we can use the *Tukey-Welsch method* and *Peritz's method*. Also the *step-down Dunnett method* can be applied for the Dunnett type comparisons. They are *CTPs* and control the *FWER* in a strong sense. Note that the Williams test is also a *CTP*.

Modified Bonferroni Procedures (MBPs)

Modified Bonferroni procedures (*MBPs*) are extensions of the classical Bonferroni procedure, which use the Bonferroni's or similar criterion to test the intersection hypotheses H_I in *CTPs*. They use only *individual p-values* for multiplicity adjustment and are easy to apply. *Holm*, *Hochberg*, *Hommel* and *Rom procedures* are some of typical *MBPs*.

Gatekeeping Procedures (GKPs)

Most recently the new methods called the *gatekeeping procedures* (*GKPs*) have been rapidly developed. *GKPs* utilize the order and logical relationship between hypotheses or families of hypotheses and construct a *MTP* satisfying these relationships. They are usually based on *CTPs* and control the *FWER* in a strong sense. They include *serial GKP*, *parallel GKP*, *tree GKP*, and *truncated GKP*, etc. *GKPs* are especially useful for multiple endpoints and various combination structures of multiple comparisons, multiple endpoints and other multiplicities.

About the Authors

Dr. Toshihiko Morikawa is former professor of Kurume University, Japan. He is well-known as an author of the paper on a combined test of non-inferiority and superiority (Morikawa and Yoshida, *J. Biopharm. Statist.* 5, 297–306, 1995). He contributed to ICH as an expert working group (EWG) member of ICH E10 guideline. He is an elected member of ISI.

Dr. Takeharu Yamanaka is Chief Researcher in the Cancer Biostatistics Laboratory, National Kyushu Cancer Center, Japan. He has worked primarily on the design and analysis of clinical trials in areas including cancer. He has also served on the Data Safety Monitoring Boards for several international multi-center clinical trials.

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►False Discovery Rate
►Multiple Comparisons Testing from a Bayesian Perspective
►Simes' Test in Multiple Testing

References and Further Reading

Dmitrienko A et al (2005) Analysis of clinical Trials Using SAS: A Practical Guide. SAS Press, Cary, NC
Dmitrienko A et al (2010) Multiple Testing Problems in Pharmaceutical Statistics Chapman & Hall/CRC, Boca Raton, FL
Hochberg Y, Tamhane AC (1987) Multiple Comparison Procedures John Wiley and Sons, New York
Hsu JC (1996) Multiple comparisons: Theory and Methods. Chapman & Hall, London
Miller RG (1981) Simultaneous Statistical Inference, 2nd edn. Springer-Verlag, New York
Morikawa T, Terao A, Iwasaki M (1996) Power evaluation of various modified Bonferroni procedures by a Monte Carlo study. J Biopharm Stat 6:343–359

Multiple Comparisons Testing from a Bayesian Perspective

Andrew A. Neath[1], Joseph E. Cavanaugh[2]
[1]Professor
Southern Illinois University Edwardsville, Edwardsville, IL, USA
[2]Professor
The University of Iowa, Iowa City, IA, USA

A General Multiple Comparisons Problem

In this note, we examine a general multiple comparisons testing problem from a Bayesian viewpoint. Suppose we observe independent random samples from I normally distributed populations with equal variances. The goal of our problem is to determine which pairs of groups have equal means.

Write

$$\{X_{ij}\} \,|\, \{\mu_i\}, \sigma^2 \sim \text{ indep } N\left(\mu_i, \sigma^2\right). \tag{1}$$

We are interested in testing $H^{(a,b)}: \mu_a = \mu_b$ for each (a, b); a total of $I(I-1)/2$ distinct, but related hypotheses. A typical frequentist test is based on the decision rule of accept $H^{(a,b)}$ when

$$\left|\overline{X}_b - \overline{X}_a\right| \le Q_{a,b}. \tag{2}$$

The overall error rate is the probability of falsely rejecting any of the true hypotheses in the set $\left\{H^{(a,b)}\right\}$. The determination of $Q_{a,b}$ in (2) depends on how the overall error rate is to be controlled. A classical book featuring this multiple comparisons problem in detail is Scheffé (1959). For an applied review, see, for example, Kutner et al. (2004) or Montgomery (2008). A modern theoretical treatment is offered by Christensen (2002).

An overview to multiple comparisons under the Bayesian framework is given by Berry and Hochberg (1999). Westfall et al. (1997) consider the preceding problem of controlling the overall error rate from a Bayesian perspective. Here, our main focus is to show how a Bayesian approach can offer a logically pleasing interpretation of multiple comparisons testing.

A major point of difficulty to multiple comparisons procedures based on an accept / reject $H^{(a,b)}$ philosophy is illustrated by a case where one decides to accept $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$, but reject $\mu_1 = \mu_3$. Such an outcome is possible under decision rule (2), but an interpretation is difficult to provide since the overall decision is not logically consistent. Employing a Bayesian philosophy, we may restate the goal of the problem as quantifying the evidence from the data in favor of each hypothesis $H^{(a,b)}$.

To implement this philosophy, we will require a measure of prior/posterior belief in $H^{(a,b)}$, represented by point mass probabilities. The construction of prior probabilities over the set of hypotheses $\left\{H^{(a,b)}\right\}$ must account for the fact that the collection does not consist of mutually exclusive events. For example, $H^{(1,2)}$ true $(\mu_1 = \mu_2)$ may occur with $H^{(2,3)}$ true $(\mu_2 = \mu_3)$ or with $H^{(2,3)}$ false $(\mu_2 \neq \mu_3)$. One cannot develop a prior by comparing relative beliefs in each of the pairwise hypotheses. Furthermore, certain combinations of hypotheses in the set $\left\{H^{(a,b)}\right\}$ represent impossibilities. For example, the event with $H^{(1,2)}$ true $(\mu_1 = \mu_2)$, $H^{(2,3)}$ true $(\mu_2 = \mu_3)$, $H^{(1,3)}$ false $(\mu_1 \neq \mu_3)$ should be assigned zero probability.

Allowable decisions can be reached through the formation of equal mean clusters among the I populations. For example, the clustering $\mu_1 = \mu_2$, $\mu_3 = \mu_4$ implies $H^{(1,2)}$ true, $H^{(3,4)}$ true, and all others false. Designating a clustering of equal means will define a model nested within (1). When two or more means are taken as equal, we merely combine all relevant samples into one. The smaller model is of the same form as (1), only for $I' < I$. The problem can now be stated in terms of Bayesian ►model selection, where each allowable combination of hypotheses will correspond to a candidate model.

We provide a short review of Bayesian model selection in the general setting using the notation of Neath

and Cavanaugh (1997). Let Y_n denote the observed data. Assume that Y_n is to be described using a model M_k selected from a set of candidate models $\{M_1, \ldots, M_L\}$. Assume that each M_k is uniquely parameterized by θ_k, an element of the parameter space $\Theta(k)$. In the multiple comparisons problem, the class of candidate models consists of all possible mean clusterings. Each candidate model is parameterized by the mean vector $\mu = (\mu_1, \ldots, \mu_I)$ and the common variance σ^2, with the individual means restricted by the model-defined clustering of equalities. That is, each model determines a corresponding parameter space where particular means are taken as equal.

Let $L(\theta_k|Y_n)$ denote the likelihood for Y_n based on M_k. Let $\pi(k)$, $k = 1, \ldots, L$, denote a discrete prior over the models $M_1, \ldots, M_L$. Let $g(\theta_k|k)$ denote a prior on θ_k given the model M_k. Applying Bayes' Theorem, the joint posterior of M_k and θ_k can be written as

$$f(k, \theta_k|Y_n) = \frac{\pi(k)g(\theta_k|k)L(\theta_k|Y_n)}{h(Y_n)},$$

where $h(Y_n)$ denotes the marginal distribution of Y_n.

The posterior probability on M_k is given by

$$\pi(k|Y_n) = h(Y_n)^{-1}\pi(k)\int_{\Theta(k)} g(\theta_k|k)L(\theta_k|Y_n)\,d\theta_k. \quad (3)$$

The integral in (3) requires numerical methods or approximation techniques for its computation. Kass and Raftery (1995) provide a discussion of the various alternatives. An attractive option is one based upon the popular Bayesian information criterion (Schwarz 1978). Define

$$B_k = -2\ln L(\hat{\theta}_k|Y_n) + dim(\theta_k)\ln(n),$$

where $\hat{\theta}_k$ denotes the maximum likelihood estimate obtained by maximizing $L(\theta_k|Y_n)$ over $\Theta(k)$. It can be shown under certain nonrestrictive regularity conditions (Cavanaugh and Neath 1999) that

$$\pi(k|Y_n) \approx \frac{\exp(-B_k/2)}{\sum_{l=1}^{L}\exp(-B_l/2)}. \quad (4)$$

The advantages to computing the posterior model probabilities as (4) include computational simplicity and a direct connection with a popular and well-studied criterion for Bayesian model selection. The justification of approximation (4) is asymptotic for the general case of prior $g(\theta_k|k)$, but Kass and Wasserman (1995) argue how the approximation holds under a noninformative prior on θ_k even for moderate and small sample sizes.

Regardless of which technique is used for computing $\pi(k|Y_n)$, we compute the probability on hypothesis $H^{(a,b)}$ by summing over the probabilities on those models for which $\mu_a = \mu_b$. This gives a nice approach to determining the evidence in favor of each of the pairwise equalities. The probability approach to presenting results for multiple comparisons testing provides more information than merely an accept / reject decision and is free of the potential contradictions alluded to earlier.

Example

We illustrate the Bayesian approach to multiple comparisons testing using data from Montgomery (2008). The $I = 5$ groups correspond to different cotton blends. Five fabric specimens are tested for each blend. The response measurements reflect tensile strength (in pounds per square inch). See Table 1 for the data and summary statistics. For ease of notation, treatments are identified in ascending order of the observed sample means.

A glance at the data suggests a potentially strong clustering of μ_1, μ_2 and a clustering to a lesser degree among μ_3, μ_4, μ_5. We shall see how these notions can be quantified by computing Bayesian posterior probabilities on the pairwise equalities. The top five most likely pairwise equalities are displayed in Table 2.

The hypothesis $\mu_1 = \mu_2$ is well-supported by the data ($P[H^{(1,2)}] \approx .8$), as was suspected. There is also some evidence in favor of $\mu_3 = \mu_4$ ($P[H^{(3,4)}] \approx .6$) and a non-negligible probability of $\mu_4 = \mu_5$ ($P[H^{(4,5)}] > .1$). Yet, there is good evidence against $\mu_3 = \mu_5$ ($P[H^{(3,5)}] < .02$).

Consider the clustering among μ_3, μ_4, μ_5. Tukey's multiple comparison procedure gives a critical range of $Q = 5.37$. A pair of means is deemed equal only if the corresponding sample difference is less than Q in magnitude. One reaches the decision of accept $\mu_3 = \mu_4$, accept $\mu_4 = \mu_5$, but reject $\mu_3 = \mu_5$. This decision is not logically consistent and is lacking any probabilistic detail. The proposed Bayesian approach bridges this probabilistic gap

Multiple Comparisons Testing from a Bayesian Perspective.

Table 1 Data for example

Group (cotton blend)	Response (tensile strength in lb/in^2)	Sample mean	Sample s.d.
1	7,7,9,11,15	9.8	3.35
2	7,10,11,11,15	10.8	2.86
3	12,12,17,18,18	15.4	3.13
4	14,18,18,19,19	17.6	2.07
5	19,19,22,23,25	21.6	2.61

Multiple Comparisons Testing from a Bayesian Perspective.

Table 2 Probabilities of pairwise equalities

Hypothesis	Posterior
$\mu_1 = \mu_2$	.7976
$\mu_3 = \mu_4$	.6015
$\mu_4 = \mu_5$	.1200
$\mu_2 = \mu_3$	.0242
$\mu_3 = \mu_5$	.0191

and provides a nice presentation for multiple comparisons. Bayesian inference has an advantage over traditional frequentist approaches to multiple comparisons in that degree of belief is quantified. One can avoid illogical conclusions which arise from an accept/reject decision process.

For computing details and continued analysis on this example, see Neath and Cavanaugh (2006).

About the Author

For the biographies *see* the entry ►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements.

Cross References

►Bayesian Statistics
►False Discovery Rate
►Multiple Comparison
►Simes' Test in Multiple Testing

References and Further Reading

Berry D, Hochberg Y (1999) Bayesian perspectives on multiple comparisons. J Stat Plan Infer 82:215–227
Cavanaugh J, Neath A (1999) Generalizing the derivation of the Schwarz information criterion. Commun Stat 28:49–66
Christensen R (2002) Plane answers to complex questions, 3rd edn. Springer, New York
Kass R, Raftery A (1995) Bayes factors. J Am Stat Assoc 90:773–795
Kass R, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J Am Stat Assoc 90:928–934
Kutner M, Nachtsheim C, Neter J, Li W (2004) Applied linear statistical models, 5th edn. McGraw-Hill/Irwin, New York
Montgomery D (2008) Design and analysis of experiments, 7th edn. Wiley, New York
Neath A, Cavanaugh J (1997) Regression and time series model selection using variants of the Schwarz information criterion. Commun Stat 26:559–580
Neath A, Cavanaugh J (2006) A Bayesian approach to the multiple comparisons problem. J Data Sci 4:131–146
Scheffé H (1959) The analysis of variance. Wiley, New York
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Westfall P, Johnson W, Utts J (1997) A Bayesian perspective on the Bonferroni adjustment. Biometrika 84:419–427

Multiple Imputation

Christian Heumann
Ludwig-Maximilian University, Munich, Germany

Multiple Imputation and Combining Estimates

Missing data substantially complicates the statistical analysis of data. A common approach to circumvent the problem of analyzing a data set with missing data is to replace/impute the missing values by some estimates or auxiliary values. Subsequently, the data are then analyzed as if they would have been complete. While it is often straightforward to get a point estimate $\hat{\theta}$ for a quantity or parameter of interest, θ, an estimate for the variance of $\hat{\theta}$ is typically difficult to obtain, since the uncertainty due to the imputed values is not reflected correctly. This is exactly where multiple imputation (Rubin 1978, 1996) steps in: by creating several datasets by imputing several values for each missing position in the dataset, multiple imputation tries to reflect the uncertainty due to the imputed values. Note, that this uncertainty is additional to the usual uncertainty arising from the sampling process. Finally, the estimate $\hat{\theta}$ is computed for each of the completed datasets and these estimates are then combined into a single estimate for θ. In the following we give the algorithmic scheme for computing the combined point estimate and an estimated covariance matrix of it, that is, we directly address the case of a vector valued parameter θ. Strategies on how proper imputations can be created are discussed in the next paragraph.

Algorithm for inference under multiple imputation

1. Create m imputed datasets.
2. For each imputed dataset, $j = 1, \ldots, m$, compute the point estimate $Q^{(j)} = \hat{\theta}^{(j)}$ and its corresponding estimated (probably asymptotic) covariance matrix $U^{(j)} = \widehat{\mathrm{Cov}}(\hat{\theta}^{(j)})$. Usually, the "MI"-paradigm (Schafer 1999) assumes that $Q^{(j)}$ is asymptotically normal.
3. The multiple-imputation point estimate for θ is then

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m} Q^{(j)} = \frac{1}{m}\sum_{j=1}^{m} \hat{\theta}^{(j)}. \qquad (1)$$

4. The estimated covariance matrix of $\overline{Q}$ consists of two components, the within-imputation covariance and the between-imputation covariance. The within-imputation covariance $\overline{U}$ is given by

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m} U^{(j)} = \frac{1}{m}\sum_{j=1}^{m} \widehat{\mathrm{Cov}}(\hat{\theta}^{(j)}). \qquad (2)$$

The between-imputation covariance B is given by

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(Q^{(j)} - \overline{Q}\right)\left(Q^{(j)} - \overline{Q}\right)^{T}, \qquad (3)$$

where T means the transposed vector, i.e. B is a quadratic matrix where the dimensions are equal to the length of the vector θ. Now we can combine the two estimates to the total variance T which is our estimated covariance matrix of $\overline{Q}$:

$$T = \widehat{\mathrm{Cov}}(\overline{Q}) = \overline{U} + (1 + m^{-1})B. \qquad (4)$$

5. A problem is that while the distribution of $T^{-\frac{1}{2}}(\theta - \overline{Q})$ can be approximated by a t-distribution with ν degrees of freedom,

$$\nu = (m-1)\left[1 + \frac{\overline{U}}{1 + m^{-1}B}\right]^2, \qquad (5)$$

in the *scalar* case, the same is not trivial for the vector valued case, see Schafer (1997).

Approaches to Create Multiple Imputations

So far we have discussed how MI works in principal and how the estimates for the completed datasets can be combined. Now we address how the imputations can be generated. We assume a missing data process that is ignorable. This relates essentially to a missing at random mechanism (MAR) plus the assumption that the parameters of the data model and the parameters of the missing data process are distinct (in likelihood inference this means that the combined parameter space is the product of the two parameter spaces, in a Bayesian analysis this means roughly that the prior distributions are independent). We note, that extensions to the case of nonignorable data situations are possible (although in general this is not easy), especially if one uses a Bayesian approach. The following subsections cannot reflect the whole research which has been done in the past. They only represent a small number of methods selected by the authors.

MI from Parametric Bayesian Models

Let D^{obs} be the observed data and D^{mis} the missing part of a dataset D, with $D = (D^{\mathrm{obs}}, D^{\mathrm{mis}})$. Then, m proper multiple imputations can be obtained via the predictive posteriori distribution of the missing data given the observed data

$$p(D^{\mathrm{mis}}|D^{\mathrm{obs}}) = \int p(D^{\mathrm{mis}}|D^{\mathrm{obs}};\theta)\, p(\theta|D^{\mathrm{obs}})\, d\theta \qquad (6)$$

or an approximation thereof. Note, that $p(\theta|D^{\mathrm{obs}})$ denotes the posteriori distribution of θ. Typically, two distinct approaches are considered to generate multiple imputations from (6): joint modeling and fully conditional modeling. The first approach assumes that the data follow a specific multivariate distribution, e.g. $D \sim N(\mu, \Sigma)$. Under a Bayesian framework draws from $p(D^{\mathrm{mis}}|D^{\mathrm{obs}})$ can be either generated directly (in some trivial cases) or simulated via suitable algorithms (in most cases) such as the IP-algorithm (see, e.g., Schafer [1997]). The second approach specifies an individual conditional distribution $p(D_j|D_{-j}, \theta_j)$ for each variable $D_j \in D$ and creates imputations as draws from these univariate distributions. It can be shown that the process of iteratively drawing and updating the imputed values from the conditional distributions can be viewed as a Gibbs sampler, that converges to draws from the (theoretical) joint distribution (if it exists). Further discussions and details on these issues can be found, e.g., in Drechsler and Rässler (2008) and the references therein.

An additional important remark refers to the fact that the imputations are called improper if we only draw imputations from

$$p(D^{\mathrm{mis}}|D^{\mathrm{obs}}, \tilde{\theta}),$$

where $\tilde{\theta}$ is a reasonable point estimate of θ (such as maximum likelihood, posterior mode or posterior mean), see also section "Other Pragmatic Approaches". That is why the above mentioned IP algorithm always includes the P-Step which samples also a new value of θ from $p(\theta|D^{\mathrm{obs}})$ before using this value to create a new imputed data set.

Nonparametric Methods

Another method to create proper multiple imputations is the so-called ABB (Approximate Bayesian Bootstrap). We refer the reader to Litte and Rubin (2002, Chap. 5.4).

Bootstrap EM

If the EM (Expectation-Maximization) algorithm is applied to an incomplete dataset, then a common problem is that only a point estimate (maximum likelihood estimate) is generated, but not an estimated (co-)variance matrix of this estimate. A typical approach to handle that issue corresponds to the use of the bootstrap (see ▶Bootstrap Methods) to create multiple imputations which then can be used to calculate such an estimate as shown in section "Multiple

Imputation and Combining Estimates". The following steps are repeated for $j = 1, \ldots, m$:

1 Draw a bootstrap sample $D^{(j)}$ from the data with replacement (including all data, complete and incomplete) with the same sample size as the original data. Obtain the maximum likelihood estimate $\hat{\theta}^{(j)}$ from the EM algorithm applied to $D^{(j)}$.
2 Use $\hat{\theta}^{(j)}$ to create an imputed dataset j from $p(D^{\text{mis}}|D^{\text{obs}}; \hat{\theta}^{(j)})$.

Other Pragmatic Approaches

Since Rubin introduced the MI paradigm in the late 1970s, there have been proposed several more or less ad-hoc methods to create multiple imputations that do not rely directly on random draws of the predictive posteriori distribution (6). A common approach refers to types of regression imputation (see, e.g., Little and Rubin [2002]), whereby missing values are replaced by predicted values from a regression of the missing item on the items observed based upon the subsample of the complete cases. This may be interpreted as an approximation to $p(D^{\text{mis}}|D^{\text{obs}}; \theta)$ from (6) with the simple constraint, that the uncertainty due to estimation of θ is not sufficiently reflected and hence $p(\theta|D^{\text{obs}})$ is apparently neglected. As an approach to consider this source of uncertainty anyhow and generate pragmatic multiple imputations (PMI), one might add an stochastic error to the imputation value and/or draw a random value from the conditional estimated distribution resulting from the prediction of the regression. Further extensions on regression imputation, e.g. the use of flexible nonparametric models and a recursive algorithm (GAMRI, Generalized Additive Model based Recursive Imputation), are discussed in Schomaker et al. (2010). Of course, the combination of values form different single imputation procedures might be seen as another type of PMI as well. Various strategies, such as nearest neighbor imputation (Chen and Shao 2000), Hot Deck imputations (Little and Rubin 2002) and others can be used for that approach.

Proper Versus Pragmatic Multiple Imputation

We recommend to create proper multiple imputations based on the predictive posteriori distribution of the missing data given the observed data. As mentioned in section "Software", a variety of statistical software packages nowadays provide fast and reliable tools to create proper multiple imputations even for users with less statistical expertise in missing-data-procedures. In situations where numerical algorithms fail to do so (sparse data, small datasets) pragmatic multiple imputations can be seen as a first approach to model imputation uncertainty.

Problems and Extensions

A number of problems arise along with multiple imputation procedures. Often they are not exclusively related to multiple imputation but to the general problem of misspecification in statistical models. If, e.g., the data model is misspecified because it assumes independent observations on the sampling units, but the observations are temporally or/and spatially correlated, also the results based on MI may become erroneous. An additional problem is ►model selection in general, especially if it is applied on high dimensional data. Also fully Bayesian inference, which often takes a lot of time for one specific model, is often too time consuming to be realistically applied to such problems. The same applies to model averaging (Frequentist or Bayesian) which may be thought of being an alternative to model selection.

Software

Recent years have seen the emergence on software that not only allows for valid inference with multiple imputation but also enables users with less statistical expertise to handle missing-data problems. We shortly introduce two packages that highlight the important progresses that lately have been made in easy-to-use Open-Source-Software. A broader description, discussion and comparison on MI-software can be found in Horton and Kleinman (2007).

- `Amelia II` (Honaker et al. 2008) is a package strongly related to the statistical Software *R* (R Development Core Team 2009) and performs proper multiple imputations by using an new, bootstrapping-based EM-algorithm that is both fast and reliable. All imputations are created via the `amelia()` function. For valid inference the quantities of the m imputed data sheets can be combined (i) in *R* using the `zelig()` command of `Zelig` (Imai et al. 2006), (ii) by hand using (1) and (4), respectively, or (iii) in separate software such as SAS, Stata etc. The Amelia II Software (named after the famous "missing" pilot Amelia Mary Earhart) is exceedingly attractive as it provides many useful options, such as the analysis of time-series data, the specification of priors on individual missing cell values, the handling of ordinal and nominal variables, the choice of suitable transformations and other useful tools. For further details see King et al. (2001) and Honaker and King (2010).

- `MICE` (Multiple Imputations by Chained Equations, van Buuren and Oudshoorn (2007)) is another package provided for *R* and *S-Plus*. It implements the chained equation approach proposed from van Buuren et al. (1999), where proper multiple imputations are generated via Fully Conditional Specification and Gibbs Sampling. The imputation step is carried out using the `mice()` function. As bugs of earlier versions seem to be removed, the MICE software can be attractive especially to the advanced user since he/she may specify his/her own imputation functions without much additional effort.

Cross References

►Imputation
►Incomplete Data in Clinical and Epidemiological Studies
►Multi-Party Inference and Uncongeniality
►Multivariate Statistical Distributions
►Nonresponse in Surveys
►Nonsampling Errors in Surveys
►Sampling From Finite Populations
►Statistical Software: An Overview

References and Further Reading

Chen JH, Shao J (2000) Nearest neighbor imputation for survey data. J Off Stat 16:113–131
R Development Core Team (2009) R: a language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org
Drechsler J, Rässler S (2008) Does convergence really matter? In: Shalabh, Heumann C (eds) Recent advances in linear models and related areas. Physica, pp 341–355
Honaker and King (2010) What to do about missing data in time series cross-secttion data. Am J Polit Sci 54(2):561–581
Honaker J, King G, Blackwell M (2008) Amelia II: a program for missing data. http://gking.harvard.edu/amelia
Horton NJ, Kleinman KP (2007) Much ado about nothing: a comparison of missing data methods and software to fit incomplete regression models. Am Stat 61:79–90
Imai K, King G, Lau O (2009) Zelig software website. http://gking.harvard.edu/zelig/
King G, Honaker J, Joseph A, Scheve K (2001) Analyzing incomplete political science data: an alternative algorithm for multiple imputation. Am Polit Sci Rev 95:49–69
Little R, Rubin D (2002) Statistical analysis with missing data. Wiley, New York
Rubin DB (1978) Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. In: American Statistical Association Proceedings of the Section on Survey Research Methods, pp 20–40
Rubin DB (1996) Multiple imputation after 18+ years. J Am Stat Assoc 91:473–489
Schafer J (1997) Analysis of incomplete multivariate data. Chapman & Hall, London
Schafer J (1999) Multiple imputation: a primer. Stat Meth Med Res 8:3–15
Schomaker M, Wan ATK, Heumann C (2010) Frequentist model averaging with missing observations. Comput Stat Data Anal, in press
Van Buuren S, Oudshoorn CGM (2007) MICE: multivariate imputation by chained equations. R package version 1.16. http://web.inter.nl.net/users/S.van.Buuren/mi/hmtl/mice.htm
van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of blood pressure covariates in survival analysis. Stat Med 18:681–694

Multiple Statistical Decision Theory

Deng-Yuan Huang
Professor
Fu Jen Catholic University, Taipei, Taiwan

In the theory and practice of statistical inference, multiple decision problems are encountered in many experimental situations. The classical methods for analyzing data customarily employ hypothesis testing in most situations. In such cases, when the hypothesis is rejected, one wants to know on which of a number of possible ways the actual situations fit our goal. If in the formulation of the problem, we consider only two decisions (reject or not reject the hypothesis), we will not only neglect to differentiate between certain alternative decisions but may also be using an inappropriate acceptance region for the hypothesis. Moreover, the traditional approach to hypothesis testing problems is not formulated in a way to answer the experimenter's question, namely, how to identify the hypothesis that satisfies the goal. Furthermore, when performing a test one may commit one of two errors: rejecting the hypothesis when it is true or accepting it when it is false. Unfortunately, when the number of observations is given, both probabilities cannot be controlled simultaneously by the classical approach (Lehmann 1959). Kiefer (1977) gave an example to show that for some sample values an appropriate test does not exhibit any detailed data-dependent measure of conclusiveness that conveys our strong feeling in favor of the alternative hypothesis. To enforce Kiefer's point, Schaafsma (1969) pointed out the Neyman–Pearson formulation is not always satisfactory and reasonable (Gupta and Huang 1981).

In the preceding paragraphs, we have discussed various difficulties associated with the hypothesis testing formulation. Thus, there arises the need for a modification of this theory and for alternative ways to attack such problems.

The approach in terms of Wald's decision theory (1950) provides an effective tool to overcome the above-mentioned difficulties in some reasonable ways. Actually, the problems of hypothesis testing can be formulated as general multiple decision problems. To this end, we first define that the space A of actions of the statistician consists of a finite number ($k \geq 2$) of elements, $A = \langle a_1, a_2, \ldots, a_k \rangle$. In practice, there are two distinct types of multiple decision problems. In one the parameter space Θ is partitioned into k subsets $\Theta_1, \Theta_2, \ldots, \Theta_k$, according to the increasing value of a real-valued function $r(\underline{\theta})$, $\underline{\theta} \in \Theta$. The action a_i is preferred if $\underline{\theta} \in \Theta_i$. This type of multiple decision problem is called monotone. This approach has been studied by Karlin and Rubin (1956) and Brown et al. (1976). For example, in comparing two treatments with means θ_1 and θ_2, an experimenter may have only a finite number of actions available, among these the experimenter might have preference based on the magnitudes of the differences of the means $\theta_2 - \theta_1$: A particular case occurs when one may choose from the three alternatives:

1. Prefer treatment 1 over treatment 2
2. Prefer treatment 2 over treatment 1
3. No preference (Ferguson 1967)

Another important class of multiple decision problems arises – selection problems where the treatments are classified into a superior category (the selected items) and an inferior one. In general, selection problems have been treated under several different formulations (Gupta and Panchapakesan 1979).

Recently, the modification of the classical hypothesis testing is considered the null hypothesis and several alternative hypotheses. Some multiple decision procedures are proposed to test the hypotheses. Under controlling the type I error, the type II error is the probability of incorrect decision. The type I and type II errors are given, the sample size can be determined. In general, one's interest is not just testing H_0 against the global alternative. Formulating the problem as one of choosing a subset of a set of alternatives has been studied (Lin and Huang 2007).

About the Author

Dr. Deng-Yuan Huang is Professor and Director, Institute of Applied Statistics, and Dean of the College of Management at Fu-Jen Catholic University in Taipei, Taiwan. He received his Ph.D. degree in Statistics from Purdue University in 1974. He is a renowned scholar in multiple decision theory, and has published numerous books and journal articles. Professor Huang has held positions of great honor in the research community of his country. He has also served as a member of the Committee on Statistics and the Committee on the Census of the Directorate General of Budget Accounting and Statistics of Taiwan. Before beginning his doctoral studies under Professor Shanti Gupta, he received the B.S. in mathematics from National Taiwan Normal University and the M.S. in Mathematics from National Taiwan University. Professor Huang is a member of the Institute of Mathematical Statistics, the Chinese Mathematical Association, and the Chinese Statistical Association. In 2002, he received the Distinguished Alumnus Award from Purdue University. In his honor, the International Conference on Multiple Decision Theory was held in Taiwan in 2007.

Cross References

►Decision Theory: An Introduction
►Decision Theory: An Overview

References and Further Reading

Brown LD, Cohen A, Strawderman WE (1976) A complete class theorem for strict monotone likelihood ratio with applications. Ann Stat 4:712–722
Ferguson TS (1967) Mathematical statistics: a decision theoretic approach. Academic, New York
Gupta SS, Huang DY (1981) Multiple decision theory: recent developments. Lecture notes in statistics, vol 6. Springer, New York
Gupta SS, Panchapakesan S (1979) Multiple decision procedures: theory and methodology of selecting and ranking populations. Wiley, New York, Republished by SIAM, Philadelphia, 2002
Karlin S, Rubin H (1956) The theory of decision procedures for distribution rules. Ann Math Stat 27:272–299
Kiefer J (1977) Conditional confidence statements and confidence estimators. JASA 72:789–827 (with comments)
Lehmann L (1959) Testing statistical hypotheses. Wiley, New York
Lin CC, Huang DY (2007) On some multiple decision procedures for normal variances Communication in statistics. Simulat Comput 36:265–275
Schaafsma W (1969) Minimal risk and unbiaseness for multiple decision procedures of type I. Ann Math Stat 40:1684–1720
Wald A (1950) Statistical decision function. Wiley, New York

Multistage Sampling

David Steel
Professor, Director of Centre for Statistical and Survey Methodology
University of Wollongong, Wollongong, NSW, Australia

Probability and Single Stage Sampling

In probability sampling each unit in the finite population of interest has a known, non-zero, chance of selection, π_i. In

single stage sampling the units in the sample, *s*, are selected directly from the population and information is obtained from them. For example, the finite population of interest may consist of businesses and a sample of businesses is selected. In these cases the population units and sampling units are the same. To obtain a single stage sample a sampling frame consisting of a list of the population units and means of contacting them are usually required. Simple random sampling (SRS) can be used, in which each possible sample of a given size has the same chance of selection. SRS leads to each unit in the population having the same chance of selection and is an equal probability selection method (EPSEM). Other EPSEMs are available. A probability sampling method does not need to be an EPSEM. As long as the selection probabilities are known it is possible to produce an estimator that is design unbiased, that is unbiased over repeated sampling. For example the ►Horvitz-Thompson estimator of the population total can be used, $\hat{T}_y = \sum_{i \in s} \pi_i^{-1} y_i$.

Stratification is often used, in which the population is divided into strata according to the values of auxiliary variables known for all population units. An independent sample is then selected from each stratum. The selection probabilities may be the same in each stratum, but often they are varied to give higher sampling rates in strata that are more heterogeneous and/or cheaper to enumerate. Common stratification variables are geography, size and type, for example industry of a business.

Cluster and Multistage Sampling

Instead of selecting a sample of population units directly it may be more convenient to select sampling units which are groups that contain several population units. The sampling unit and the population unit differ. The groups are called Primary Sampling Units (PSUs). If we select all population units from each selected PSU we have ►cluster sampling. If we select a sample of the units in the selected PSUs we have multistage sampling. Each population unit must be uniquely associated with only one PSU through coverage rules. These methods are often used when there is some geographical aspect to the sample selection and there are significant travel costs involved in collecting data and/or when there is no suitable population list of the population units available. A common example of a PSU is a household, which contains one or more people (Clark and Steel 2002). Another common example is area sampling (see Kish 1963, Chap. 9).

In a multistage sample the sample is selected in stages, the sample units at each stage being sampled from the larger units chosen at the previous stage. At each successive stage smaller sampling units are defined within those selected at the previous stage and further selections are made within each of them. At each stage a list of units from which the selections are to be made is required only within units selected at the previous stage.

For example, suppose we wish to select a sample of visitors staying overnight in the city of Wollongong. No list of such people exists, but if we confine ourselves to people staying in hotels or motels then it would be possible to construct a list of such establishments. We could then select a sample of hotels and motels from this list and select all guests from the selected establishments, in which case we have a cluster sample. It would probably be better to select a sample from the guests in each selected establishment allowing selection of more establishments, in which case we have a multi-stage sampling scheme. The probability of a particular guest being selected in the sample is the product of the probability of the establishment being selected and the probability the guest is selected given the establishment is selected. Provided the selection of establishments and guests within selected establishments is done using probability sampling, the sampling method is a valid probability sample. It would also be worthwhile stratifying according to the size of the establishment and its type.

Cluster and multistage sampling are used because a suitable sampling frame of population units does not exist but a list of PSUs does, or because they are less costly than a single stage sample of the same size in terms of population units. In multistage sampling the probability a population unit is selected is the probability the PSU containing the unit is selected multiplied by the conditional probability that the unit is selected given that the PSU it is in is selected.

Cluster and multistage sampling are often cheaper and more convenient than other methods but there is usually an increase in standard errors for the same sample size in terms of number of finally selected population units. It is important that the estimation of sampling error reflects the sample design used (See Lohr 1999, Chap. 9).

In many situations, the problems of compiling lists of population units and travel between selected population units are present even within selected PSUs. Consideration is then given to selecting the sample of population units within a selected PSU by grouping the population units into second stage units, a sample of which is selected. The population units are then selected from selected second stage units. This is called three-stage sampling. This process can be continued to any number of stages. The set of all selected population units in a selected PSU is called an ultimate cluster.

Multistage sampling is very flexible since many aspects of the design have to be chosen including the number of

stages and, for each stage, the unit of selection, the method of selection and number of units selected. Stratification and ratio or other estimation techniques may be used. This flexibility means that there is large scope for meeting the demands of a particular survey in an efficient way.

For a multistage sample the sampling variance of an estimator of a mean or total has a component arising from each stage of selection. The contribution of a stage of selection is determined by the number of units selected at that stage and the variation between the units at that stage, within the units at the next highest level. The precise formula depends on the selection and estimation methods used (See Lohr 1999, Chaps. 5–6; Cochran 1977, Chaps. 9, 9A, 10–11; Kish 1963, Chaps. 5–7, 9–10).

If PSUs vary appreciably in size then it can be useful to control the impact of this variation using ratio estimation or Probability Proportional to Size (PPS) sampling using the number of units in the PSU. For two-stage sampling a common design involves PPS selection of PSUs and selection of an equal number of units in each selected PSU. This gives each population unit the same chance of selection, which is usually a sensible feature for a sample of people, and an equal workload within each selected PSU, which has operational benefits. The first stage component of variance is determined by the variation of the PSU means. To use PPS sampling we need to know the population size of each PSU in the population. For ratio estimation we only need to know the total population size.

Optimal Design in Multistage Sampling

One of the main problems in designing multistage samples is to determine what size sample within selected PSUs to take to optimally balance cost and sampling error. In a two stage sampling scheme in which m PSUs are to be selected and the average number of units selected in each PSU is $\overline{n}$ the sampling variance is minimized for fixed sample size when $\overline{n} = 1$, since then the sample includes the largest number of PSUs. However, costs will be minimized when as few PSUs as possible are selected. Costs and variances are pulling in opposite directions and we must try to optimally balance them. In a two-stage sample several types of costs can be distinguished: overhead costs, costs associated with the selection of PSUs and costs associated with the selection of 2nd stage units. This leads to specifying a cost function of the form

$$C_0 + C_1 m + C_2 m\overline{n}.$$

For some of the common two-stage sampling and estimation methods used in practice the variance of the estimator of total or mean can be written as

$$V_0^2 + \frac{V_1^2}{m} + \frac{V_2^2}{m\overline{n}}.$$

For fixed cost the variance is minimized by choosing

$$\overline{n} = \sqrt{\frac{C_1}{C_2}\frac{V_2^2}{V_1^2}}.$$

The optimum choice of $\overline{n}$ thus depends on the ratios of costs and variances. As the first stage costs increase relative to the second stage costs the optimum $\overline{n}$ increase, so we are led to a more clustered sample. As the second stage component of variance increases relative to the first stage we are also led to a more clustered design.

The optimum value of $\overline{n}$ can be expressed in terms of the measure of homogeneity $\delta = \frac{V_1^2}{V_1^2 + V_2^2}$, as $\overline{n} = \sqrt{\frac{C_1}{C_2}\frac{1-\delta}{\delta}}$. As δ increases the optimal choice of $\overline{n}$ decreases. For example if $C_1/C_2 = 10$ and $\delta = 0.05$ then the optimal $\overline{n} = 14$. To determine the optimal choice of $\overline{n}$ we only need to obtain an idea of the ratio of first stage to second stage cost coefficients and δ.

About the Author

Dr David Steel is a Professor in the School of Mathematics and Applied Statistics, University of Wollongong, Australia. He was the Head of the School of Mathematics and Applied Statistics (2000–2004) and Associate Dean (Research) for the Faculty of Informatics (2004–2006). He is foundation Director of the Center for Statistical and Survey Methodology (2007–). He has authored and co-authored more than 60 papers and books chapters. Professor Steel is currently an Associate Editor for the *Journal of the Royal Statistical Society (Series A)* and *Survey Methodology*. He is a foundation member of the Methodological Advisory Committee of the Australian Bureau of Statistics (1995–).

Cross References

►Cluster Sampling
►Sample Survey Methods
►Sampling From Finite Populations
►Stratified Sampling

References and Further Reading

Clark R, Steel DG (2002) The effect of using household as a sampling unit. Int Stat Rev 70:289–314
Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York
Lohr S (1999) Sampling: design and analysis. Duxbury, Pacific Grove
Kish L (1965) Survey sampling. Wiley, New York

Multivariable Fractional Polynomial Models

Willi Sauerbrei[1], Patrick Royston[2]
[1]Professor
University Medical Center Freiburg, Freiburg, Germany
[2]Professor
University College London, London, UK

Fractional Polynomial Models

Suppose that we have an outcome variable, a single continuous covariate X, and a suitable regression model relating them. Our starting point is the straight line model, $\beta_1 X$ (for simplicity, we suppress the constant term, β_0). Often a straight line is an adequate description of the relationship, but other models must be investigated for possible improvements in fit. A simple extension of the straight line is a power transformation model, $\beta_1 X^p$. The latter model has often been used by practitioners in an *ad hoc* way, utilising different choices of p. Royston and Altman (1994) formalize the model slightly by calling it a first-degree fractional polynomial or FP1 function. The power p is chosen from a pragmatically chosen restricted set $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where X^0 denotes $\log X$.

As with polynomial regression, extension from one-term FP1 functions to the more complex and flexible two-term FP2 functions follows immediately. Instead of $\beta_1 X^1 + \beta_2 X^2$, FP2 functions with powers (p_1, p_2) are defined as $\beta_1 X^{p_1} + \beta_2 X^{p_2}$ with p_1 and p_2 taken from S. If $p_1 = p_2$ Royston and Altman proposed $\beta_1 X^{p_1} + \beta_2 X^{p_1} \log X$, a so-called repeated-powers FP2 model.

For a more formal definition, we use the notation from Royston and Sauerbrei (2008). An FP1 function or model is defined as $\varphi_1(X, p) = \beta_0 + \beta_1 X^p$, the constant (β_0) being optional and context-specific. For example, β_0 is usually included in a normal-errors regression model but is always excluded from a Cox proportional-hazards model. An FP2 transformation of X with powers $\mathbf{p} = (p_1, p_2)$, or when $p_1 = p_2$ with repeated powers $\mathbf{p} = (p_1, p_1)$ is the vector $X^{\mathbf{p}}$ with

$$X^{\mathbf{p}} = X^{(p_1,p_2)} = \begin{cases} (X^{p_1}, X^{p_2}), & p_1 \neq p_2 \\ (X^{p_1}, X^{p_1} \log X), & p_1 = p_2 \end{cases}$$

An FP2 function (or model) with parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathrm{T}}$ and powers $\mathbf{p}$ is $\varphi_2(X, \mathbf{p}) = \beta_0 + X^{\mathbf{p}}\boldsymbol{\beta}$. With the set S of powers as just given, there are 8 FP1 transformations, 28 FP2 transformations with distinct powers $(p_1 \neq p_2)$ and 8 FP2 transformations with equal powers $(p_1 = p_2)$. The best fit among the combinations of powers from S is defined as that with the highest likelihood.

The general definition of an FPm function with powers $\mathbf{p} = (p_1 \leq \cdots \leq p_m)$ is conveniently written as a recurrence relation. Let $h_0(X) = 1$ and $p_0 = 0$. Then

$$\varphi_m(X, \mathbf{p}) = \beta_0 + X^{\mathbf{p}}\beta = \beta_0 + \sum_{j=1}^{m} \beta_j h_j(X)$$

where for $j = 1, \ldots, m$

$$h_j(X) = \begin{cases} X^{p_j}, & p_{j-1} \neq p_j \\ h_{j-1}(X) \log X, & p_{j-1} = p_j \end{cases}$$

For example, for $m = 2$ and $\mathbf{p} = (-1, 2)$ we have $h_1(X) = X^{-1}$, $h_2(X) = X^2$. For $\mathbf{p} = (2, 2)$ we have $h_1(X) = X^2$, $h_2(X) = X^2 \log X$.

Figure 1 shows some FP2 curves, chosen to indicate the flexibility available with a few pairs of powers (p_1, p_2). The ability to fit a variety of curve shapes, some of which have asymptotes or which have both a sharply rising or falling portion and a nearly flat portion, to real data is a particularly useful practical feature of FP2 functions.

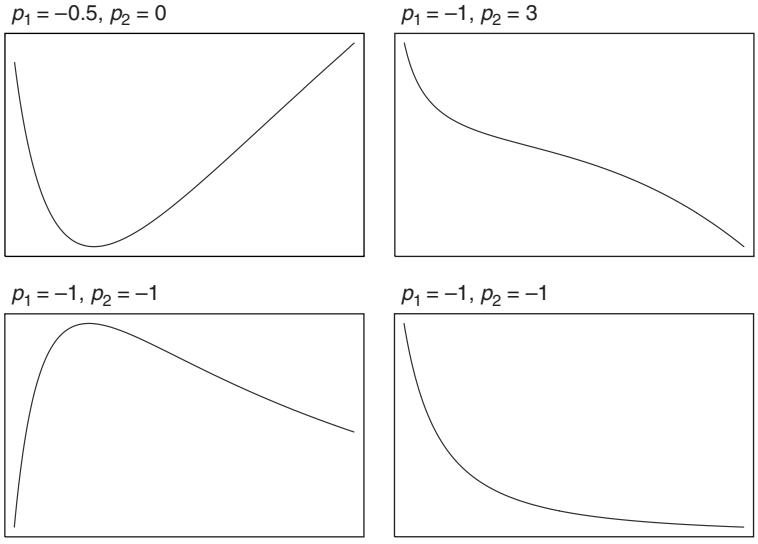
Function Selection Procedure (FSP)

Choosing the best FP1 or FP2 function by mininizing the deviance (minus twice the maximized log likelihood) is straightforward. However, having a sensible default function is important for increasing the parsimony, stability and general usefulness of selected functions. In most of the algorithms implementing FP modelling, the default function is linear – arguably, a natural choice. Therefore, unless the data support a more complex FP function, a straight line model is chosen. There are occasional exceptions; for example, in modelling time-varying regression coefficients in the Cox model, Sauerbrei et al. (2007a) chose a default time transformation of $\log t$ rather than t.

It is assumed in what follows that the null distribution of the difference in deviances between an FPm and an FP$(m-1)$ model is approximately central χ^2 on two degrees of freedom. Justification of this result is given in Sect. 4.9.1 of Royston and Sauerbrei (2008) and supported by simulation results (Ambler and Royston 2001).

For FP model selection, Royston and Sauerbrei (2008) proposed using the following closed test procedure (although other procedures are possible). It runs as follows:

1. Test the best FP2 model for X at the α significance level against the null model using four d.f. If the test is not significant, stop, concluding that the effect of X is "not significant" at the α level. Otherwise continue.



Multivariable Fractional Polynomial Models. Fig. 1 Examples of FP2 curves for different powers (p_1, p_2)

2. Test the best FP2 for X against a straight line at the α level using three d.f. If the test is not significant, stop, the final model being a straight line. Otherwise continue.
3. Test the best FP2 for X against the best FP1 at the α level using two d.f. If the test is not significant, the final model is FP1, otherwise the final model is FP2. End of procedure.

The test at step 1 is of overall association of the outcome with X. The test at step 2 examines the evidence for non-linearity. The test at step 3 chooses between a simpler or more complex non-linear model. Before applying the procedure, the analyst must decide on the nominal P-value (α) and on the degree (m) of the most complex FP model allowed. Typical choices are $\alpha = 0.05$ and FP2 ($m = 2$).

Multivariable Fractional Polynomial (MFP) Procedure

In many studies, a relatively large number of predictors is available and the aim is to derive an interpretable multivariable model which captures the important features of the data: the stronger predictors are included and plausible functional forms are found for continuous variables.

As a pragmatic strategy to building such models, a systematic search for possible non-linearity (provided by the FSP) is added to a backward elimination (BE) procedure. For arguments to combine FSP with BE, see Royston and Sauerbrei (2008). The extension is feasible with any type of regression model to which BE is applicable. Sauerbrei and Royston (1999) called it the multivariable fractional polynomial (MFP) procedure, or simply MFP. Using MFP successfully requires only general knowledge about building regression models.

The nominal significance level is the main tuning parameter required by MFP. Actually, two significance levels are needed: α_1 for selecting variables with BE, and α_2 for comparing the fit of functions within the FSP. Often, $\alpha_1 = \alpha_2$ is a good choice. A degree greater than 2 ($m > 2$) is rarely if ever needed in a multivariable context. Since the model is derived data-dependently, parameter estimates are likely to be somewhat biased.

As with any multivariable selection procedure checks of the underlying assumptions and of the influence of single observations are required and may result in model refinement. To improve robustness of FP models in the univariate and multivariable context Royston and Sauerbrei (2007) proposed a preliminary transformation of X. The transformation shifts the origin of X and smoothly pulls in extreme low and extreme high values towards the center of the distribution. The transformation is linear in the central bulk of the observations.

If available, subject-matter knowledge should replace data-dependent model choice. Only minor modifications are required to incorporate various types of subject-matter knowledge into MFP modelling. For the discussion of a detailed example, see Sauerbrei and Royston (1999).

For model-building by selection of variables and functional forms for continuous predictors, MFP has several advantages over spline-based models (the most important alternatives). For example, MFP models exhibit fewer

artefacts in fitted functions, and are more transportable, mathematically concise and generally more useful than spline models (Royston and Sauerbrei 2008; Sauerbrei et al. 2007b). Residual analysis with spline models may be used to check whether the globally defined functions derived by MFP analysis have missed any important local features in the functional form for a given continuous predictor (Binder and Sauerbrei 2010).

Recommendations for practitioners of MFP modelling are given in Royston and Sauerbrei (2008) and Sauerbrei et al. (2007b).

Extensions of MFP to Investigate for Interactions

MFP was developed to select main effects of predictors on the outcome. If a variable X_2 explains (at least partially) the relationship between a predictor X_1 and the outcome Y then confounding is present. Another important issue is interaction between two or more predictors in a multivariable model. An interaction between X_1 and X_2 is present if X_2 modifies the relationship between X_1 and the outcome. That means that the effect of X_1 is different in subgroups determined by X_2. Extensions of MFP have been proposed to handle two-way interactions involving at least one continuous covariate (Royston and Sauerbrei 2004). Higher order interactions, which typically play a role in factorial experiments, are a further extension, but not one that has yet been considered in the FP context.

To investigate for a possible interaction between a continuous predictor and two treatment arms in a randomized controlled trial, the multivariable fractional polynomial interaction (MFPI) procedure was introduced (Royston and Sauerbrei 2004). In a first step, the FP class is used to model the prognostic effect of the continuous variable separately in the two treatment arms, usually under some restrictions such as the same power terms in each arm. In a second step, a test for the equality of the prognostic functions is conducted. If significant, an interaction is present and the difference between two functions estimates the influence of the prognostic factor on the effect of treatment. The difference function is called a treatment effect function (and should be plotted). For interpretation, it is important to distinguish between the two cases of a predefined hypothesis and of searching for hypotheses (Royston and Sauerbrei 2004, 2008).

For more than two groups, extensions to investigate continuous by categorical interactions are immediate. Furthermore, MFPI allows investigation of treatment-covariate interactions in models with or without adjustment for other covariates. The adjustment for other covariates enables the use of the procedure in observational studies, where the multivariable context is more important than in an RCT.

Continuous-by-continuous interactions are important in observational studies. A popular approach is to assume linearity for both variables and test the multiplicative term for significance. However, the model may fit poorly if one or both of the main effects is non-linear. Royston and Sauerbrei (2008, Chap. 7) introduced an extension of MFPI, known as MFPIgen, in which products of selected main effect FP functions are considered as candidates for an interaction between a pair of continuous variables. Several continuous variables are usually available, and a test of interaction is conducted for each such pair. If more than one interaction is detected, interactions are added to the main-effects model in a step-up manner.

The MFPT(ime) algorithm (Sauerbrei et al. 2007a) combines selection of variables and of the functional form for continuous variables with determination of time-varying effects in a Cox proportional hazards model for ►survival data. A procedure analogous to the FSP was suggested for investigating whether the effect of a variable varies in time, i.e., whether a time-by-covariate interaction is present.

Further Contributions to Fractional Polynomial Modelling

Methods based on fractional polynomials have been reported recently, aiming to improve or extend the modelling of continuous covariates in various contexts. For example, Faes et al. (2007) applied model averaging to fractional polynomial functions to estimate a safe level of exposure; Lambert et al. (2005) considered time-dependent effects in regression models for relative survival; and Long and Ryoo (2010) used FPs to model non-linear trends in longitudinal data. For further topics and references, see Sect. 11.3 of Royston and Sauerbrei (2008).

About the Authors

Willi Sauerbrei, Ph.D., is a senior statistician and professor in medical biometry at the University Medical Center Freiburg. He has authored many research papers in biostatistics, and has published over 150 articles in leading statistical and clinical journals. He worked for more than 2 decades as an academic biostatistician and has extensive experience of cancer research. Together with Patrick Royston, he has written a book on modeling (*Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*, Wiley 2008).

Patrick Royston, D.Sc., is a senior statistician at the MRC Clinical Trials Unit, London, an honorary professor of statistics at University College London and a Fellow of the Royal Statistical Society. He has authored many research papers in biostatistics, including over 150 articles in leading statistical journals. He is co-author (with Willi Sauerbrei, see above) of a book on multivariable modeling. He is also an experienced statistical consultant, Stata programmer and software author.

Cross References

►Interaction
►Measurement Error Models
►Model Selection
►Nonparametric Regression Using Kernel and Spline Methods

References and Further Reading

Ambler G, Royston P (2001) Fractional polynomial model selection procedures: investigation of Type I error rate. J Stat Comput Simul 69:89–108

Binder H, Sauerbrei W (2010) Adding local components to global functions for continuous covariates in multivariable regression modeling. Stat Med 29:808–817

Faes C, Aerts M, Geys H, Molenberghs G (2007) Model averaging using fractional polynomials to estimate a safe level of exposure. Risk Anal 27:111–123

Lambert PC, Smith LK, Jones DR, Botha JL (2005) Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. Stat Med 24:3871–3885

Long J, Ryoo J (2010) Using fractional polynomials to model nonlinear trends in longitudinal data. Br J Math Stat Psychol 63: 177–203

Royston P, Altman DG (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Appl Stat 43(3):429–467

Royston P, Sauerbrei W (2004) A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med 23: 2509–2525

Royston P, Sauerbrei W (2007) Improving the robustness of fractional polynomial models by preliminary covariate transformation. Comput Stat Data Anal 51:4240–4253

Royston P, Sauerbrei W (2008) Multivariable model-building – a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Wiley, Chichester

Sauerbrei W, Royston P (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. J R Stat Soc A 162:71–94

Sauerbrei W, Royston P, Look M (2007a) A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. Biomet J 49:453–473

Sauerbrei W, Royston P, Binder H (2007b) Selection of important variables and determination of functional form for continuous predictors in multivariable model-building. Stat Med 26: 5512–5528

Multivariate Analysis of Variance (MANOVA)

Barbara G. Tabachnick, Linda S. Fidell
California State University, Northridge, CA, USA

ANOVA (►analysis of variance) tests whether mean differences among groups on a single DV (dependent variable) are likely to have occurred by chance. MANOVA (multivariate analysis of variance) tests whether mean differences among groups on a *combination* of DVs are likely to have occurred by chance. For example, suppose a researcher is interested in the effect of different types of treatment (the IV; say, desensitization, relaxation training, and a waiting-list control) on anxiety. In ANOVA, the researcher chooses one measure of anxiety from among many. With MANOVA, the researcher can assess several types of anxiety (say, test anxiety, anxiety in reaction to minor life stresses, and so-called free-floating anxiety). After random assignment of participants to one of the three treatments and a subsequent period of treatment, participants are measured for test anxiety, stress anxiety, and free-floating anxiety. Scores on all three measures for each participant serve as DVs. MANOVA is used to ask whether a combination of the three anxiety measures varies as a function of treatment. (MANOVA is statistically identical to discriminant analysis. The difference between the techniques is one of emphasis. MANOVA emphasizes the mean differences and statistical significance of differences among groups. Discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems) emphasizes prediction of group membership and the dimensions on which groups differ.)

MANOVA developed in the tradition of ANOVA. Traditionally, MANOVA is applied to experimental situations where all, or at least some, IVs are manipulated and participants are randomly assigned to groups, usually with equal cell sizes. The goal of research using MANOVA is to discover whether outcomes, as reflected by the DVs, are changed by manipulation (or other action) of the IVs.

In MANOVA, a new DV is created from the set of DVs that maximizes group differences. The new DV is a linear combination of measured DVs, combined so as to separate the groups as much as possible. ANOVA is then performed on the newly created DV. As in ANOVA, hypotheses about means are tested by comparing variances between means relative to variances in scores within groups-hence multivariate analysis of variance.

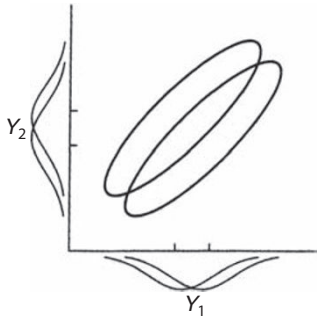
In factorial or more complicated MANOVA, a different linear combination of DVs is formed for each IV and

interaction. If gender of participant is added to type of treatment as a second IV, one combination of the three DVs maximizes the separation of the three treatment groups, a second combination maximizes separation of women and men, and a third combination maximizes separation of the six cells of the interaction. Further, if an IV has more than two levels, the DVs can be recombined in yet other ways to maximize the separation of groups formed by comparisons.

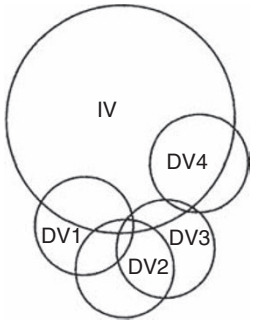
MANOVA has a number of advantages over ANOVA. First, by measuring several DVs instead of only one, the researcher improves the chance of discovering what it is that changes as a result of different IVs and their interactions. For instance, desensitization may have an advantage over relaxation training or waiting-list control, but only on test anxiety; the effect is missed in ANOVA if test anxiety is not chosen as the DV. A second advantage of MANOVA over a series of ANOVAs (one for each DV) is protection against inflated Type I error due to multiple tests of (likely) correlated DVs. (The linear combinations themselves are usually of interest in discriminant analysis, but not in MANOVA.)

Another advantage of MANOVA is that, under certain, probably rare conditions, it may reveal differences not shown in separate ANOVAs (Maxwell 2001). Such a situation is shown in Fig. 1 for a one-way design with two levels. In this figure, the axes represent frequency distributions for each of two DVs, Y_1 and Y_2. Notice that from the point of view of either axis, the distributions are sufficiently overlapping that a mean difference might not be found in ANOVA. The ellipses in the quadrant, however, represent the distributions of Y_1 and Y_2 for each group separately. When responses to two DVs are considered in combination, group differences become apparent. Thus, MANOVA, which considers DVs in combination, may occasionally be more powerful than separate ANOVAs.

The goal in MANOVA is to choose a small number of DVs where each DV is related to the IV, but the DVs are not related to each other. Good luck. In the usual situation there are correlations among the DVs, resulting in some ambiguity in interpretation of the effects of IVs on any single DV and loss of power relative to ANOVA. Figure 2 shows a set of hypothetical relationships between a single IV and four DVs. DV1is highly related to the IV and shares some variance with DV2 and DV3. DV2 is related to both DV1 and DV3 and shares very little unique variance with the IV. DV3 is somewhat related to the IV, but also to all of the other DVs. DV4 is highly related to the IV and shares only a little bit of variance with DV3. Thus, DV2 is completely redundant with the other DVs, and DV3 adds only a bit of unique variance to the set. (However, DV2 might be useful as a covariate if that use is conceptually viable because it reduces the total variances in DVs 1 and 3 that are not overlapping with the IV.)



Multivariate Analysis of Variance (MANOVA). Fig. 1 Advantage of MANOVA, which combines DVs, over ANOVA. Each axis represents a DV; frequency distributions projected to axes show considerable overlap, while ellipses, showing DVs in combination, do not



Multivariate Analysis of Variance (MANOVA). Fig. 2 Hypothetical relationships among a single IV and four DVs

Although computing procedures and programs for MANOVA and MANCOVA are not as well developed as for ANOVA and ANCOVA, there is in theory no limit to the generalization of the model. The usual questions regarding main effects of IVs, interactions among IVs, importance of DVs, parameter estimates (marginal and cell means), specific comparisons and trend analysis (for IVs with more than two levels), effect sizes of treatments, and effects of covariates, if any, are equally interesting with MANOVA as with ANOVA. There is no reason why all types of designs - one-way, factorial, repeated measures, nonorthogonal, and so on - cannot be extended to research with several DVs.

For example, multivariate analysis of covariance (MANCOVA) is the multivariate extension of ANCOVA. MANCOVA asks if there are statistically significant mean differences among groups after adjusting the newly created DV for differences on one or more covariates. To extend the example, suppose that before treatment participants are

pretested on test anxiety, minor stress anxiety, and free-floating anxiety; these pretest scores are used as covariates in the final analysis. MANCOVA asks if mean anxiety on the composite score differs in the three treatment groups, after adjusting for preexisting differences in the three types of anxieties.

MANOVA is also a legitimate alternative to repeated-measures ANOVA in which differences between pairs of responses to the levels of the within-subjects IV are simply viewed as separate DVs.

Univariate analyses are also useful following a MANOVA or MANCOVA. For example, if DVs can be prioritized, ANCOVA is used after MANOVA (or MANCOVA) in Roy-Bargmann stepdown analysis where the goal is to assess the contributions of the various DVs to a significant effect (Bock 1971; Bock and Haggard 1968). One asks whether, after adjusting for differences on higher-priority DVs serving as covariates, there is any significant mean difference among groups on a lower-priority DV. That is, does a lower-priority DV provide additional separation of groups beyond that of the DVs already used? In this sense, ANCOVA is used as a tool in interpreting MANOVA results. Results of stepdown analysis are reported in addition to individual ANOVAs.

However, MANOVA is a substantially more complicated analysis than ANOVA because there are several important issues to consider. MANOVA has all of the complications of ANOVA (e.g., homogeneity of variance; equality of sample sizes within groups; absence of ►outliers; power, cf. Woodward et al. 1990; normality of sampling distributions, independence of errors) and several more besides (homogeneity of variance-covariance matrices; multivariate normality, cf. Mardia 1971 and Seo et al. 1995; linearity, absence of ►multicollinearity and singularity; and choice among statistical criteria, cf. Olson 1979). These are not impossible to understand or test prior to analysis, but they are vital to an honest analysis.

Comprehensive statistical software packages typically include programs for MANOVA. The major SPSS module is GLM, however the older MANOVA module remains available through syntax and includes Roy-Bargmann stepdown analysis as an option. NCSS and SYSTAT have specific MANOVA modules, whereas SAS provides analysis of MANOVA through its GLM module. Analysis is also available through BMDP4V, STATA, and Statistica.

For more information about MANOVA, see Chaps. 7 and 8 of Tabachnick and Fidell (2007).

About the Authors

Dr Barbara Tabachnick is a Professor Emerita at California State University, Northridge. She has authored and co-authored more than 100 papers and chapters, as well as two books, including *Using Multivariate Statistics* (5th edition, Allyn & Bacon, 2007) and *Experimental Designs Using ANOVA* (Duxbury 2007), both with Dr. Linda Fidell. She continues to consult on research grants.

Dr. Linda Fidell is a Professor Emerita at California State University, Northridge. She has authored and co-authored more than 60 papers and chapters, as well as two books, including *Using Multivariate Statistics* (5th edition, Allyn & Bacon, 2007) and *Experimental Designs Using ANOVA* (Duxbury 2007), both with Dr. Barbara Tabachnick. She continues to consult on research grants.

Cross References

►Analysis of Variance
►Discriminant Analysis: An Overview
►Discriminant Analysis: Issues and Problems
►General Linear Models
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Nonparametric Models for ANOVA and ANCOVA Designs
►Statistical Fallacies: Misconceptions, and Myths

References and Further Reading

Bock RD, Haggard EA (1968) The use of multivariate analysis of variance in behavioral research. McGraw-Hill, New York

Mardia KV (1971) The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. Biometrika 58(1):105–121

Maxwell S (2001) When to use MANOVA and significant MANOVAs and insignificant ANOVAs or vice versa. J Consum Psychol 10(1–2):29–30

Olson CL (1976) On choosing a test statistic in multivariate analysis of variance. Psychol Bull 83(4):579–586

Seo T, Kanda T, Fujikoshi Y (1995) The effects of nonnormality on tests for dimensionality in canonical correlation and MANOVA models. J Multivariate Anal 52:325–337

Tabachnick BG, Fidell LS (2007) Using multivariate statistics. Allyn & Bacon, Boston

Woodward JA, Overall JE (1975) Multivariate analysis of variance by multiple regression methods. Psychol Bull 82(1):21–32

Multivariate Data Analysis: An Overview

Joseph F. Hair
Professor of Marketing
Kennesaw State University, Kennesaw, GA, USA

Most business problems involve many variables. Managers look at multiple performance measures and related metrics

when making decisions. Consumers evaluate many characteristics of products or services in deciding which to purchase. Multiple factors influence the stocks a broker recommends. Restaurant patrons consider many factors in deciding where to dine. As the world becomes more complex, more factors influence the decisions managers and customers make. Thus, increasingly business researchers, as well as managers and customers, must rely on more sophisticated approaches to analyzing and understanding data.

Analysis of data has previously involved mostly univariate and bivariate approaches. Univariate analysis involves statistically testing a single variable, while bivariate analysis involves two variables. When problems involve three or more variables they are inherently multidimensional and require the use of multivariate data analysis. For example, managers trying to better understand their employees might examine job satisfaction, job commitment, work type (part-time vs. full-time), shift worked (day or night), age and so on. Similarly, consumers comparing supermarkets might look at the freshness and variety of produce, store location, hours of operation, cleanliness, prices, courtesy and helpfulness of employees, and so forth. Managers and business researchers need multivariate statistical techniques to fully understand such complex problems.

Multivariate data analysis refers to all statistical methods that simultaneously analyze multiple measurements on each individual respondent or object under investigation. Thus, any simultaneous analysis of more than two variables can be considered multivariate analysis. Multivariate data analysis is therefore an extension of univariate (analysis of a single variable) and bivariate analysis (cross-classification, correlation, and simple regression used to examine two variables).

Figure 1 displays a useful classification of statistical techniques. Multivariate as well as univariate and bivariate techniques are included to help you better understand the similarities and differences. As you can see at the top, we divide the techniques into dependence and interdependence depending on the number of dependent variables. If there is one or more dependent variables a technique is referred to as a dependence method. That is, we have both dependent and independent variables in our analysis. In contrast, when we do not have a dependent variable we refer to the technique as an interdependence method. That is, all variables are analyzed together and our goal is to form groups or give meaning to a set of variables or respondents.

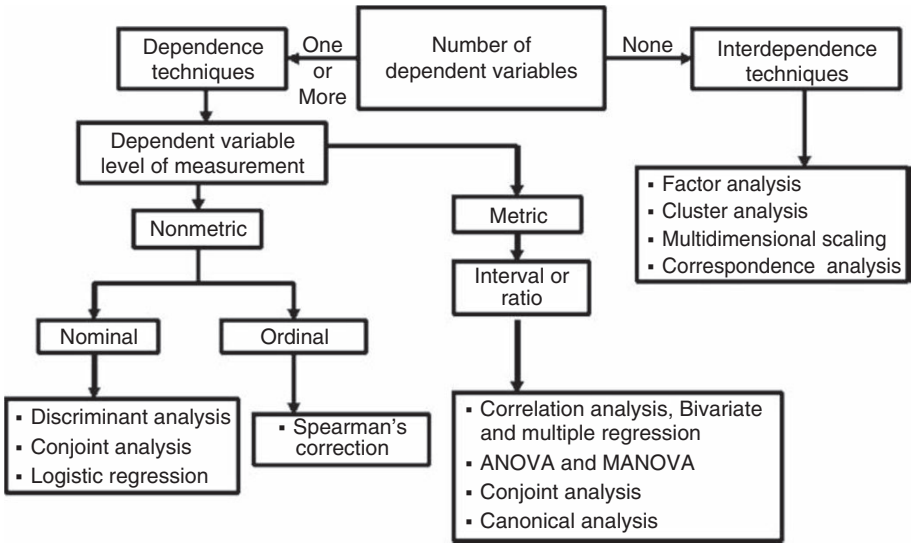
The classification can help us understand the differences in the various statistical techniques. If a research problem involves association or prediction using both dependent and independent variables, one of the dependence techniques on the left side of the diagram is appropriate. The choice of a particular statistical technique depends on whether the dependent variable is metric or nonmetric, and how many dependent variables are involved. With a nonmetric, ordinally measured dependent we would use the Spearman correlation. With a nonmetric, nominal dependent variable we use discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems), conjoint analysis or ►logistic regression. On the other hand, if our dependent variable is metric, we can use correlation, regression, ANOVA or MANOVA, canonical correlation, and conjoint analysis (the statistical technique of conjoint analysis can be formulated to handle both metric and nonmetric variables). The various statistical techniques are defined in Fig. 2. For more information on multivariate statistical techniques see Hair et al. (2010).

Concluding Observations

Today multivariate data analysis is being used by most medium and large sized businesses, and even some small businesses. Also, most business researchers rely on multivariate analysis to better understand their data. Thus, in today's business environment it's just as important to understand the relationship between variables, which often requires multivariate analysis, as it is to gather the information in the first place. The importance of multivariate statistical methods that help us to understand relationships has increased dramatically in recent years. What can we expect in the future as applications of multivariate data analysis expand: (1) data will continue to increase exponentially, (2) data quality will improve as will data cleaning techniques and data maintenance, (3) data analysis tools will be more powerful and easier to use, and (4) there will be many more career opportunities involving examining and interpreting data using multivariate data analysis.

About the Author

Professor Joe Hair is a member of the American Marketing Association, Academy of Marketing Science, and Society for Marketing Advances. He has authored 55 books, monographs, and cases, and over 80 articles in scholarly journals. He is a co-author (with William C. Black, Barry Babin and Rolph Anderson) of the well known applications-oriented introduction to multivariate analysis text *Multivariate Data Analysis* (Prentice Hall, 7th edition, 2010). He serves on the editorial review boards of several journals and was the 2009 Academy of Marketing Science/Harold Berkman Lifetime Service Award recipient, the KSU Coles College Foundation Distinguished Professor in 2009, the Marketing Management Association Innovative Marketer



Multivariate Data Analysis: An Overview. Fig. 1 Classification of statistical techniques

ANOVA – ANOVA stands for analysis of variance. It is used to examine statistical differences between the means of two or more groups. The dependent variable is metric and the independent variable(s) is nonmetric. One-way ANOVA has a single non-metric independent variable and two-way ANOVA can have two or more non-metric independent variables. ANOVA is bivariate while MANOVA is the multivariate extension of ANOVA.

Bivariate Regression – this is a type of regression that has a single metric dependent variable and a single metric independent variable.

Cluster Analysis – this type of analysis enables researchers to place objects (e.g., customers, brands, products) into groups so that objects within the groups are similar to each other. At the same time, objects in any particular group are different from objects in all other groups.

Correlation – correlation examines the association between two metric variables. The strength of the association is measured by the correlation coefficient. **Canonical correlation** analyzes the relationship between multiple dependent and multiple independent variables, most often using metric measured variables.

Conjoint Analysis – this technique enables researchers to determine the preferences individuals have for various products and services, and which product features are valued the most.

Discriminant Analysis – enables the researcher to predict group membership using two or more metric dependent variables. The group membership variable is a non-metric dependent variable.

Factor Analysis – this technique is used to summarize the information from a large number of variables into a much smaller number of variables or factors. This technique is used to combine variables whereas cluster analysis is used to identify groups with similar characteristics.

Logistic Regression – logistic regression is a special type of regression that involves a non-metric dependent variable and several metric independent variables.

Multiple Regression – this type of regression has a single metric dependent variable and several metric independent variables.

MANOVA – same technique as ANOVA but it can examine group differences across two or more metric dependent variables at the same time.

Perceptual Mapping – this approach uses information from other statistical techniques (e.g., multidimensional scaling) to map customer perceptions of products, brands, companies, and so forth.

Multivariate Data Analysis: An Overview. Fig. 2 Definitions of statistical techniques

of the Year in 2007, and the 2004 recipient of the Academy of Marketing Science Excellence in Teaching Award.

Cross References

►Canonical Correlation Analysis
►Cluster Analysis: An Introduction
►Correspondence Analysis
►Data Analysis
►Discriminant Analysis: An Overview
►Discriminant Analysis: Issues and Problems
►Factor Analysis and Latent Variable Modelling
►Linear Regression Models
►Logistic Regression
►Multidimensional Scaling
►Multidimensional Scaling: An Introduction
►Multivariate Analysis of Variance (MANOVA)
►Multivariate Rank Procedures: Perspectives and Prospectives
►Multivariate Reduced-Rank Regression
►Multivariate Statistical Analysis
►Multivariate Statistical Process Control
►Principal Component Analysis
►Scales of Measurement
►Scales of Measurement and Choice of Statistical Methods
►Structural Equation Models

References and Further Reading

Esbensen KH (2006) Multivariate data analysis. IM Publications, Chichester
Hair J et al (2010) Multivariate data analysis, 7th edn. Prentice-Hall
Ho R (2006) Handbook of univariate and multivariate data analysis and interpretation with SPSS. Chapman & Hall, CRC, Boca Raton
Manly B (2005) Multivariate statistical methods a primer. Chapman & Hall, CRC, Boca Raton
Spicer J (2005) Making sense of multivariate data analysis: an intuitive approach. Sage Publications, Thousand Oaks

Multivariate Normal Distributions

Damir Kalpić[1] , Nikica Hlupić[2]
[1]Professor and Head, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia
[2]Assistant Professor, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia

The multivariate normal distribution is a generalization of the familiar univariate normal or Gaussian distribution (Hogg et al. 2005; Miller and Miller 1999) to $p \geq 2$ dimensions. Just as with its univariate counterpart, the importance of the multivariate normal distribution emanates from a number of its useful properties, and especially from the fact that, according to the central limit theorem (Anderson 2003; Johnson and Wichern 2007) under certain regularity conditions, sum of random variables generated from various (likely unknown) distributions tends to behave as if its underlying distribution were multivariate normal.

The need for generalization to the multivariate distribution naturally arises if we simultaneously investigate more than one quantity of interest. In that case, single observation (result of an experiment) is not value of a single variable, but the set of p values of $p \geq 2$ random variables. Therefore, we deal with $p \times 1$ random vector $\mathbf{X}$ and each single observation becomes $p \times 1$ vector $\mathbf{x}$ of single realizations of p random variables under examination. All these variables have their particular expected values that jointly constitute $p \times 1$ mean vector $\boldsymbol{\mu}$, which is expected value of random vector $\mathbf{X}$. Since analysis of collective behaviour of several quantities must take into account their mutual *correlations,* in multivariate analysis we also define $p \times p$ *variance-covariance matrix*

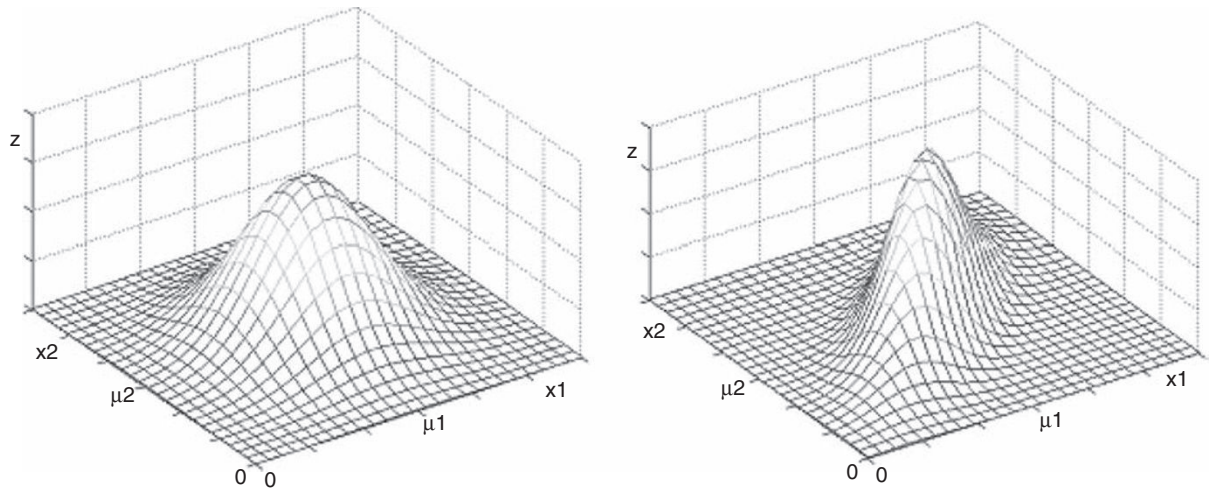
$$\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X}) = E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix}, \quad (1)$$

where σ_{ij} are covariances between ith and jth component of $\mathbf{X}$ and σ_{ii} are variances of ith variable (more commonly denoted σ_i^2). This matrix is symmetric because $\sigma_{ij} = \sigma_{ji}$ and it is assumed to be *positive definite.*

Conceptually, the development of multivariate normal distribution starts from the univariate *probability* density function of a normal random variable X with the mean μ and variance σ^2. Common notation is $X \sim N(\mu, \sigma^2)$ and *probability density function* (pdf) of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}z^2}; -\infty < x < +\infty. \quad (2)$$

Variable Z is so-called *standard normal variable* or *z-score* and it represents the square of the distance from a single observation (measurement) x to the population



Multivariate Normal Distributions. Fig. 1 Bivariate normal distribution with: *left* - $\sigma_1 = \sigma_2, \rho = 0$; *right* - $\sigma_1 = \sigma_2, \rho = 0,75$

mean μ, expressed in standard deviation units. It is this distance that directly generalizes to $p \geq 2$ dimensions, because in the univariate case we can write

$$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu), \tag{3}$$

and in the multivariate case, by analogy, we have the *Mahalanobis distance* (Johnson and Wichern 2007) expressed as

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}). \tag{4}$$

The multivariate normal probability density function is obtained (Anderson 2003; Hogg et al. 2005; Johnson and Wichern 2007) by replacing (3) by (4) in the density function (2) and substituting the normalizing constant by $(2\pi)^{-p/2}|\Sigma|^{-1/2}$, so that the p-dimensional normal probability density for the random vector $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]^T$ is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2} \tag{5}$$

where $x_i \in (-\infty, \infty)$ and $i = 1, 2, \ldots, p$. Again analogously to the univariate case, we write $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

As an example, consider bivariate $(p = 2)$ distribution in terms of the individual parameters $\mu_1, \mu_2, \sigma_1^2 = \sigma_{11}, \sigma_2^2 = \sigma_{22}$ and $\sigma_{12} = \sigma_{21}$. If we also introduce *correlation coefficient* $\rho = \rho_{12} = \text{corr}(X_1, X_2) = \sigma_{12}^2/(\sigma_1 \cdot \sigma_2)$, density (5) becomes

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2}\right]\right\}. \tag{6}$$

Formula (6) clearly indicates certain important general properties of multivariate normal distributions. First of all, if random variables X_1 and X_2 are uncorrelated, i.e., $\rho = 0$, it immediately follows that their joint density (6) can be factored as the product of two univariate normal densities of the form of (2). Since $f(x_1, x_2)$ factors as $f(x_1, x_2) = f(x_1) \cdot f(x_2)$, it follows that if X_1 and X_2 are uncorrelated, they are also *statistically independent*. This is a direct consequence of the general $(p \geq 2)$ multivariate normal property that uncorrelated variables are independent and have *marginal distributions* univariate normal. However, converse is not necessarily true for both of these statements and requires caution. Independent normal variables certainly are uncorrelated (this is true for any distribution anyway), but marginal distributions may be univariate normal without the joint distribution being multivariate normal. Similarly, marginally normal variables can be uncorrelated without being independent (Anderson 2003; Miller and Miller 1999).

Several other general properties of multivariate normal distribution are easier to conceive by studying the bivariate normal surface defined by (6) and illustrated in Fig. 1. Obviously, the bivariate (as well as multivariate) probability density function has a maximum at (μ_1, μ_2). Next, any intersection of this surface and a plane parallel to the z-axis has the shape of an univariate normal distribution, indicating that marginal distributions are univariate normal.

Finally, any intersection of this surface and a plane parallel to the x_1x_2 plane is an ellipse called *contour of constant probability density*. In the special case when variables are uncorrelated (independent) and $\sigma_1 = \sigma_2$ (Fig. 1 - left), contours of constant probability density are circles

and it is customary to refer to the corresponding joint density as a *circular normal density*. When variables are uncorrelated, but $\sigma_1 \neq \sigma_2$, contours are ellipses whose semi-axes are parallel to the x_1, x_2 axes of the coordinate system. In the presence of correlation, probability density concentrates along the line (Fig. 1 - right) determined by the coefficient of correlation and variances of variables, so the contours of constant probability density are ellipses rotated in a plane parallel to x_1x_2 plane (Anderson 2003; Miller and Miller 1999). All these properties are valid in p-dimensional spaces ($p > 2$) as well.

Here is the list of most important properties of the multivariate normal distribution (Anderson 2003; Johnson and Wichern 2007; Rao 2002).

1. Let $\mathbf{X}$ be a random vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{a}$ an arbitrary $p \times 1$ vector. Then the linear combination $\mathbf{a}^T\mathbf{X} = a_1X_1 + a_2X_2 + \ldots + a_pX_p$ is distributed as $N(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})$. In words, any linear combination of jointly normal random variables is normally distributed. Converse is also true: if $\mathbf{a}^T\mathbf{X}$ is $\sim N(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})$ for every $\mathbf{a}$, then $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
2. Generalization of property 1: Let $\mathbf{X}$ be a random vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let us form q linear combinations $\mathbf{AX}$, where $\mathbf{A}$ is an arbitrary $q \times p$ matrix. Then it is true that $\mathbf{AX} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$. Similarly, for any vector of constants $\mathbf{d}$ we have $\mathbf{X} + \mathbf{d} \sim N_p(\boldsymbol{\mu} + \mathbf{d}, \boldsymbol{\Sigma})$.
3. All subsets of variables constituting $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are (multivariate) normally distributed.
4. Multivariate normal $q_1 \times 1$ and $q_2 \times 1$ vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent if and only if they are uncorrelated, i.e., $\mathrm{cov}(\mathbf{X}_1, \mathbf{X}_2) = 0$ (a $q_1 \times q_2$ matrix of zeros).
5. If multivariate normal $q_1 \times 1$ and $q_2 \times 1$ vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent and distributed as $N_{q1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $N_{q2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, respectively, then $(q_1 + q_2) \times 1$ vector $[\mathbf{X}_1^T\ \mathbf{X}_2^T]^T$ has multivariate normal distribution

$$\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

6. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be mutually independent random vectors that are all multivariate normally distributed, each having its particular mean, but all having the same covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{X}_j \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$. Linear combination of these vectors $\mathbf{V}_1 = c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + \ldots + c_n\mathbf{X}_n$ is distributed as $N_p\left(\sum_{j=1}^{n} c_j\boldsymbol{\mu}_j, \left(\sum_{j=1}^{n} c_j^2\right)\boldsymbol{\Sigma}\right)$. Moreover, similarly to property 5, $\mathbf{V}_1$ and some other linear combination $\mathbf{V}_2 = b_1\mathbf{X}_1 + b_2\mathbf{X}_2 + \ldots + b_n\mathbf{X}_n$ are jointly multivariate normally distributed with covariance matrix

$$\begin{bmatrix} \left(\sum_{j=1}^{n} c_j^2\right)\boldsymbol{\Sigma} & (\mathbf{b}^T\mathbf{c})\boldsymbol{\Sigma} \\ (\mathbf{b}^T\mathbf{c})\boldsymbol{\Sigma} & \left(\sum_{j=1}^{n} b_j^2\right)\boldsymbol{\Sigma} \end{bmatrix}.$$

Thus, if $\mathbf{b}^T\mathbf{c} = 0$, i.e., vectors $\mathbf{b}$ and $\mathbf{c}$ are orthogonal, it follows that $\mathbf{V}_1$ and $\mathbf{V}_2$ are independent and vice versa.
7. All conditional distributions are multivariate normal. Formally, let $\mathbf{X}_1$ and $\mathbf{X}_2$ be any two subsets of a multivariate normal vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, and $|\boldsymbol{\Sigma}_{22}| > 0$. The conditional distribution of $\mathbf{X}_1$, given a fixed $\mathbf{X}_2 = \mathbf{x}_2$, is multivariate normal with

$$\mathrm{mean}(\mathbf{X}_1|\mathbf{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \text{ and } \mathrm{cov}(\mathbf{X}_1|\mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

8. Generalized distance $(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ of observations $\mathbf{x}$ of a vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from the mean $\boldsymbol{\mu}$ has a chi squared distribution with p *degrees of freedom* denoted χ_p^2.
9. With $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ as a set of n observations from a (multivariate) normal population with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, we have the following results:
   (a) $\overline{\mathbf{X}}$ is distributed as $N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$
   (b) $(n-1)\mathbf{S}$ has a *Wishart distribution;* with $n-1$ degrees of freedom
   (c) $\overline{\mathbf{X}}$ and $\mathbf{S}$ are independent.

Cross References

►Bivariate Distributions
►Central Limit Theorems
►Hotelling's T^2 Statistic
►Multivariate Rank Procedures: Perspectives and Prospectives
►Multivariate Statistical Analysis
►Multivariate Statistical Distributions
►Multivariate Statistical Simulation
►Normal Distribution, Univariate
►Statistical Distributions: An Overview
►Statistical Quality Control: Recent Advances

References and Further Reading

Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley, Hoboken

Ghurye SG, Olkin I (1962) A characterization of the multivariate normal distribution. Ann Math Stat 33:533–541

Green PE (1978) Analyzing multivariate data. Dryden Press, London
Hogg RV, McKean JW, Craig AT (2005) Introduction to mathematical statistics, 6th edn. Pearson Prentice Hall, Upper Saddle River
Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Pearson Prentice Hall, New York
Kagan A, Linnik YV, Rao CR (1972) Characterization problems of mathematical statistics. Wiley, New York
Miller I, Miller M (1999) John E. Freund's mathematical statistics, 6th edn. Pearson Prentice Hall, Upper Saddle River
Rao CR (2002) Linear statistical inference and its applications, 2nd edn. Wiley, New York
Seal HL (1967) Studies in the history of probability and statistics. XV The historical development of the Gauss linear model, Biometrika, 54:1–24

Multivariate Outliers

Isabel M. Rodrigues[1], Graciela Boente[2]
[1]Assistant Professor
Technical University of Lisbon (TULisbon), Lisboa, Portugal
[2]Professor, Facultad de Ciencias Exactas and Naturales Universidad de Buenos Aires and CONICET, Buenos Aires, Argentina

In the statistical analysis of data one is often confronted with observations that "appear to be inconsistent with the remainder of that set of data" (Barnett and Lewis 1994). Although such observations (the ▶outliers) have been the subject of numerous investigations, there is no general accepted formal definition of outlyingness. Nevertheless, the outliers describe abnormal data behavior, i.e., data that are deviating from the natural data variability (see, e.g., Peña and Prieto 2001, Filzmoser 2004, and Filzmoser et al. 2008 for a discussion).

Sometimes outliers can grossly distort the statistical analysis, while at other times their influence may not be as noticeable. Statisticians have accordingly developed numerous algorithms for the detection and treatment of outliers, but most of these methods were developed for univariate data sets. They are based on the estimation of location and scale, or on quantiles of the data. Since in a univariate sample outliers may be identified as an exceptionally large or small value, a simple plot of the data, such as scatterplot, stem-and-leaf plot, and QQ-plot can often reveal which points are outliers.

In contrast, for multivariate data sets the problem of outliers identification gives challenges that do not occur with univariate data since there is no simple concept of ordering the data. Furthermore, the multivariate case introduces a different kind of outlier, a point that is not extreme component wise but departs from the prevailing pattern of correlation structure. This departs causes that the observations appear as univariate outliers in some direction not easily identifiable. In this context, to detect an observation as possible outlier not only the distance from the centroid of the data is important but also the data shape. Also, as Gnanadesikan and Kettenring (1972) pointed out the visual detection of multivariate outliers is virtually impossible because the outliers do not "stick out on the end."

Since most standard multivariate analysis techniques rely on the assumption of normality, in 1963, Wilks proposed identifying sets of outliers of size j from $\{1, 2, \ldots, n\}$, in normal multivariate data, by checking the minimum values of the ratios $|A_{(I)}|/|A|$, where $|A_{(I)}|$ is the internal scatter of a modified sample in which the set of observations I of size j has been deleted and $|A|$ is the internal scatter of the complete sample. For $j = 1$ this method is equivalent to the classical way to declare a multivariate observation as a possible outlier by using the squared Mahalanobis' distance defined as

$$MD_i^2(\mathbf{x}_i, \mathbf{t}, \mathbf{V}) = ((\mathbf{x}_i - \mathbf{t})^T \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{t}))^{1/2}$$

where $\mathbf{t}$ is the estimated multivariate location and $\mathbf{V}$ the estimated scatter matrix. Usually $\mathbf{t}$ is the multivariate arithmetic mean, the centroid, and $\mathbf{V}$ the sample covariance matrix. Mahalanobis' distance identifies observations that lie far away from the center of the data cloud, giving less weight to variables with large variances or to groups of highly correlated variables. For a p-multivariate normally distributed data $MD_i^2(\mathbf{x}_i, \mathbf{t}, \mathbf{V})$ converge to χ_p^2, a chi-square distribution with p degree of freedom. Points with large $MD_i^2 \equiv MD_i^2(\mathbf{x}_i, \mathbf{t}, \mathbf{V})$, compared with some χ_p^2 quantile, are then considered outliers. Hence, to evaluate multivariate normality one may plot the ordered $MD_{(i)}^2$ against the expected order statistics of the ▶chi-square distribution with sample quantiles $\chi^2_{p[(i-1/2)/2]} = q_i$ where q_i $(i = 1, \ldots, n)$ is the $100(i - 1/2)/n$ sample quantile of χ_p^2. The plotted points $(MD_{(i)}, q_i)$ should be close to a line, so the points far from the line are potential outliers. Formal tests for multivariate outliers are considered by Barnett and Lewis (1994).

Clearly, the Mahalanobis distance relies on classical location and scatter estimators. The presence of outliers may distort arbitrarily the values of these estimators and render meaningless the results. This is particularly acute when there are several outliers forming a cluster, because

they will move the arithmetic mean toward them and inflate the classical tolerance ellipsoid in their direction. So this approach suffers from the *masking* and *swamping* effects by which multiple outliers do not have a large MD_i^2. A solution to this problem is well known in ►robust statistics: **t** and **V** have to be estimated in a robust manner, where the expression "robust"means resistance against the influence of outlying observations. Thus, the "robustified" ordered Mahalanobis distances, $RMD_{(i)}^2$ may be plotted to locate extreme outliers. This is the approach considered by Becker and Gather (2001), Filzmoser (2004), and Hardin and Rocke (2005) who studied outlier identification rules adapted to the sample size using different location and scatter robust estimators.

For a review on some of the robust estimators for location and scatter introduced in the literature see Maronna et al. (2006). The minimum covariance determinant (MCD) estimator – the procedure is due to Rousseeuw (1984) – is probably most frequently used in practice, partly because a computationally fast algorithm has been developped (Rousseeuw and Van Driessen 1999). The MCD estimator also benefits from the availability of software implementation in different languages, including R, S-Plus, Fortran, Matlab, and SAS. For these reasons the MCD estimator had gained much popularity, not only for outliers identification but also as an ingredient of many robust multivariate techniques.

Other currently popular multivariate outlier detection methods fall under projection pursuit techniques, originally proposed by Kruskal (1969). Projection pursuit searches for "interesting" linear projections of multivariate data sets, where a projection is deemed interesting if it minimizes or maximizes a projection index (typically a scale estimator). Therefore, the goal of projection pursuit methods is to find suitable projections of the data in which the outliers are readily apparent and may thus be downweighted to yield a estimator, which in turn can be used to identify the outliers. Since they do not assume the data to originate from a particular distribution but only search for useful projections, projection pursuit procedures are not affected by non-normality and can be widely applied in diverse data situations. The penalty for such freedom comes in the form of increased computational burden, since it is not clear which projections should be examined. An exact method would require to test over all possible directions.

The most well-known outlier identification method based upon the projection pursuit concept is the Stahel–Donoho (Stahel 1981; Donoho 1982) estimator. This was the first introduced high-breakdown and affine equivariant estimator of multivariate location and scatter that became better known after Maronna and Yohai (1995) published an analysis of it. It is based on a measure of the outlyingness of data points, which is obtained by projecting the observation on univariate directions. The Stahel–Donoho estimator then computes a weighted mean and covariance matrix, with weights inverse proportional to the outlyingness. This outlyingness measure is based upon the projection pursuit idea that if a point is a multivariate outlier, there must be some one-dimensional projection of the data in which this point is a univariate outlier. Using a particular observation as a reference point, the Stahel–Donoho algorithm determines which directions have optimal values for a pair of robust univariate location/scale estimators and then uses these estimators to assign weights to the other points. One way of reducing the computational cost of the Stahel–Donoho estimator is to reduce the number of projections that need to be examined.

In this direction, Peña and Prieto (2001) proposed a method, the Kurtosis1, which involves projecting the data onto a set of $2p$ directions. These directions are chosen to maximize and minimize the kurtosis coefficient of the data along them. A small number of outliers would cause heavy tails and lead to a larger kurtosis coefficient, while a larger number of outliers would start introducing bimodality and decrease the kurtosis coefficient. Viewing the data along projections that have maximum and minimum kurtosis values would therefore seem to display the outliers in a more recognizable representation.

For a much more detailed overview about outliers see Barnett and Lewis (1994) and also Rousseeuw et al. (2006) for a review on robust statistical methods and outlier detection.

Cross References

►Chi-Square Distribution
►Distance Measures
►Multivariate Normal Distributions
►Multivariate Technique: Robustness
►Outliers
►Robust Statistical Methods

References and Further Reading

Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, Chichester

Becker C, Gather U (2001) The largest nonidentifiable outlier: a comparison of multivariate simultaneous outlier identification rules. Comput Stat Data Anal 36:119–127

Donoho D (1982) Breakdown properties of multivariate location estimators. Ph.D. thesis, Harvard University

Filzmoser P (2004) A multivariate outlier detection method. In: Aivazian S, Filzmoser P, Kharin Yu (eds) Proceedings of the seventh international conference on computer data analysis and modeling, vol 1. Belarusian State University, Minsk, pp 18–22

Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Comput Stat Data Anal 52:1694–1711
Gnanadesikan R, Kettenring JR (1972) Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28:81–124
Hardin J, Rocke D (2005) The distribution of robust distances. J Comput Graph Stat 14:928–946
Kruskal JB (1969) Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation". In: Milton RC, Nelder JA (eds) Statistical computation. Academic, New York, pp 427–440
Maronna RA, Yohai VJ (1995) The behavior of the Stahel-Donoho robust multivariate estimator. J Am Stat Assoc 90:330–341
Maronna RA, Martin RD, Yohai V (2006) Robust statistics: theory and methods. Wiley, New York
Peña D, Prieto FJ (2001) Multivariate outlier detection and robust covariance matrix estimation. Technometrics 43:286–310
Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880
Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223
Rousseeuw PJ, Debruyne M, Engelen S, Hubert M (2006) Robustness and outlier detection in chemometrics. Cr Rev Anal Chem 36:221–242
Stahel WA (1981) Robust estimation: infinitesimal optimality and covariance matrix estimators. Ph.D. thesis in German, Swiss Federal Institute of Technology, Zurich, Switzerland
Wilks SS (1963) Multivariate statistical outliers. Sankhya 25:407–426

Multivariate Rank Procedures : Perspectives and Prospectives

Pranab K. Sen
Cary C. Boshamer Professor of Biostatistics and Professor of Statistics and Operations Research
University of North Carolina, Chapel Hill, NC, USA

Developments ►in multivariate statistical analysis have genesis in the parametrics surrounding the *multivariate normal* distribution (see ►Multivariate Normal Distributions) in the continuous case while the *product multinomial law* dominates in discrete multivariate analysis. Characterizations of multi-normal distributions have provided a wealth of rigid mathematical tools leading to a very systematic evolution of mathematical theory laying down the foundation of multivariate statistical methods. *Internal multivariate* analyses comprising of *principal component models, canonical correlation* and *factor analysis* are all based on appropriate *invariance structures* that exploit the underlying linearity of the interrelation of different characteristics, without depending much on underlying normality, and these tools are very useful in many areas of applied research, such as sociology, psychology, economics, and agricultural sciences. In the recent past, there has been a phenomenal growth of multivariate analysis in medical studies, clinical trials and ►bioinformatics, among others. The role of multinormality is being scrutinized increasingly in these contexts.

External multivariate analyses pertaining to ►*multivariate analysis of variance* (MANOVA) and covariance (MANOCOVA), *classification and discrimination,* among others, have their roots in the basic assumption of multinormal distribution, providing some optimal, or at least desirable, properties of statistical inference procedures. Such optimal statistical procedures generally exist only when the multinormality assumption holds. Yet, in real life applications, the postulation of multinormality may not be tenable in a majority of cases. Whereas in the univariate case, there are some other distributions, some belonging to the so-called *exponential family of densities* and some not, for which exact statistical inference can be drawn, often being confined to suitable subclass of statistical procedures. In the multivariate case, alternatives to multinormal distributions are relatively few and lack generality. As such, almost five decades ago, it was strongly felt that statistical procedures should be developed to bypass the stringent assumption of multinormality; this is the genesis of *multivariate nonparametrics*.

Whereas the classical normal theory likelihood based multivariate analysis exploited *affine invariance*, leading to some optimality properties, it has some shortcomings too. Affine invariance makes sense only when the different characteristics or variates are linearly combinable in a meaningful way. Further, such parametric procedures are quite vulnerable to even small departures from the assumed multinormality. Thus, they are generally *nonrobust* even in a local sense. Moreover, in many applications, different characteristics are recorded on different units and often on a relative scale (viz., ranking of n individuals on some multivariate traits) where linearly combinability may not be compatible. Rather, it is more important to have coordinatewise invariance under arbitrary strictly monotone transformations – a feature that favors ranks over actual measurements. Multivariate rank procedures have this basic advantage of invariance under coordinatewise arbitrary strictly monotone transformations, not necessarily linear. Of course, this way the emphasis on affine invariance is sacrificed, albeit, there are affine-invariant rank procedures too (see Oja 2010).

The basic difference between univariate and multivariate rank procedures is that for suitable *hypothesis of invariance*, in the univariate case, such procedures are genuinely distribution-free, whereas in the multivariate case,

even the hypothesis of invariance holds, these tests are usually *conditionally distribution-free*. This feature, known as the *rank-permutation principle*, was initially developed by Chatterjee and Sen (1964) and in a more general framework, compiled and reported in Puri and Sen (1971), the first text in multivariate nonparametrics. During the past four decades, a phenomenal growth of research literature in multivariate nonparametrics has taken place; specific entries in the *Encyclopedia of Statistical Science* and *Encyclopedia of Biostatistics* (both published from Wiley-Interscience, New York) provide detailed accounts of these developments.

In the recent past, *high-dimensional low sample size* (HDLSS) problems have cropped up in diverse fields of application. In this setup, the dimension is generally far larger than the number of sample observations, and hence, standard parametric procedures are untenable; nonparametrics fare much better. This is a new frontier of multivariate nonparametrics and there is a tremendous scope of prospective research with deep impact on fruitful applications. ►Data mining (or knowledge discovery and data mining) and statistical learning algorithms also rest on multivariate nonparametrics to a greater extent. Bioinformatics and environmetrics problems also involve such nonstandard multivariate nonparametric procedures. In a micro-array data model, an application of multivariate rank methods has been thoroughly explored in Sen (2008).

About the Author

Dr. Pranab Kumar Sen is a Cary C. Boshamer Professor of Biostatistics, University of North Carolina (1982–) and a lifelong Adjunct Professor, Indian Statistical Institute, Calcutta (1993–). He was born on November 7, 1937 in Calcutta, India. He had his school and college education (B.Sc. (1955), M.Sc. (1957) and Ph.D. (1962), all in Statistics) from Calcutta University. Professor Sen is Fellow of the Institute of Mathematical Statistics (1968), Fellow of the American Statistical Association (1969), and Elected Member of the International Statistical Institute (1973). Professor Sen has (co-)authored over 615 publications in Statistics, Probability Theory, Stochastic Processes, and Biostatistics in leading journals in these areas, and (co-)authored or (co-) edited 23 books and monographs in Statistics, Probability Theory and Biostatistics. He has (co-)supervised the Doctoral Dissertation of 82 students from University of North Carolina (1969–2009), many of whom have achieved distinction both nationally and internationally. In 1988 he was awarded the Boltzman Award in Mathematical Sciences from Charles University, Prague, and in 1998, the Commemoration Medal by the Czech Union of Mathematicians and Physicists, Prague. In 2002, he was awarded the Senior Noether Award from the American Statistical Association for his significant contributions to Nonparametrics, teaching as well as research. In 2010, Professor Sen has received the Wilks Medal, American Statistical Association. He was the Founding (joint) Editor of two international journals: *Sequential Analysis* (1982) and *Statistics and Decisions* (1983). Currently, he is the Chief Editor of *Sankhya* (Series A and B).

"Professor Sen's pioneering contributions have touched nearly every area of statistics. He is the first person who, in joint collaboration with Professor S. K. Chatterjee, developed multivariate rank tests as well as time-sequential nonparametric methods. He is also the first person who carried out in-depth research in sequential nonparametrics culminating in his now famous Wiley book *Sequential Nonparametrics: Invariance Principles and Statistical Inference* and SIAM monograph." (Malay Ghosh and Michael J. Schell, A Conversation with Pranab Kumar Sen, *Statistical Science*, Volume 23, Number 4 (2008), 548–564.

Cross References

►Data Mining
►Multivariate Data Analysis: An Overview
►Multivariate Normal Distributions
►Multivariate Reduced-Rank Regression
►Multivariate Statistical Analysis
►Nonparametric Statistical Inference

References and Further Reading

Chatterjee SK, Sen PK (1964) Nonparametric testing for the bivariate two-sample location problem. Calcutta Stat Assoc Bull 13:18–58
Oja H (2010) Springer book on multivariate rank procedure, August 2010
Puri ML, Sen PK (1971) Nonparametric methods in multivariate analysis. Wiley, New York
Sen PK (2008) Kendall's tau in high dimensional genomics parsimony. Institute of Mathematical Statistics, Collection Ser. 3 pp 251–266

Multivariate Reduced-Rank Regression

Alan J. Izenman
Senior Research Professor of Statistics, Director of the Center for Statistical and Information Science
Temple University, Philadelphia, PA, USA

Multivariate reduced-rank regression is a way of constraining the multivariate linear regression model so that the rank of the regression coefficient matrix has less than full

rank. Without the constraint, multivariate linear regression has no true multivariate content.

To see this, suppose we have a random r-vector $\mathbf{X} = (X_1, \ldots, X_r)^\tau$ of predictor (or input) variables with mean vector $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_{XX}$, and a random s-vector $\mathbf{Y} = (Y_1, \ldots, Y_s)^\tau$ of response (or output) variables with mean vector $\boldsymbol{\mu}_Y$ and covariance matrix $\boldsymbol{\Sigma}_{YY}$. Suppose that the $(r+s)$-vector $\mathbf{Z} = (\mathbf{X}^\tau, \mathbf{Y}^\tau)^\tau$ has a joint distribution with mean vector and covariance matrix,

$$\boldsymbol{\mu}_Z = \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \qquad \boldsymbol{\Sigma}_{ZZ} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix}, \tag{1}$$

respectively, where we assume that $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{YY}$ are both nonsingular. Now, consider the classical multivariate linear regression model,

$$\underset{s\times 1}{\mathbf{Y}} = \underset{s\times 1}{\boldsymbol{\mu}} + \underset{s\times r}{\boldsymbol{\Theta}}\ \underset{r\times 1}{\mathbf{X}} + \underset{s\times 1}{\mathcal{E}}, \tag{2}$$

where $\mathbf{Y}$ depends linearly on $\mathbf{X}$, $\boldsymbol{\mu}$ is the overall mean vector, $\boldsymbol{\Theta}$ is the multivariate regression coefficient matrix, and $\mathcal{E}$ is the error term. In this model, $\boldsymbol{\mu}$ and $\boldsymbol{\Theta}$ are unknown and are to be estimated. The least-squares estimator of $(\boldsymbol{\mu}, \boldsymbol{\Theta})$ is given by

$$(\boldsymbol{\mu}^*, \boldsymbol{\Theta}^*) = arg \min_{\boldsymbol{\mu},\boldsymbol{\Theta}} \mathrm{E}\{(\mathbf{Y} - \boldsymbol{\mu} - \boldsymbol{\Theta}\mathbf{X})(\mathbf{Y} - \boldsymbol{\mu} - \boldsymbol{\Theta}\mathbf{X})^\tau\}, \tag{3}$$

where

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_Y - \boldsymbol{\Theta}^*\boldsymbol{\mu}_X, \quad \boldsymbol{\Theta}^* = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}. \tag{4}$$

In (3), the expectation is taken over the joint distribution of $(X^\tau, Y^\tau)^\tau$. The minimum achieved is $\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$. The $(s \times r)$-matrix $\boldsymbol{\Theta}^*$ is called the (full-rank) regression coefficient matrix. This solution is identical to that obtained by performing a sequence of s ordinary least-squares multiple regressions. For the jth such multiple regression, Y_j is regressed on the r-vector $\mathbf{X}$, where $j = 1, 2, \ldots, s$. Suppose the minimizing regression coefficient vectors are the r-vectors $\boldsymbol{\beta}_j^*$, $j = 1, 2, \ldots, s$. Arranging the coefficient vectors as the columns, $(\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_r^*)$, of an $(r \times s)$-matrix, and then transposing the result, it follows from (4) that

$$\boldsymbol{\Theta}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_r^*)^\tau. \tag{5}$$

Thus, multivariate linear regression is equivalent to just carrying out a sequence of multiple regressions. This is why multivariate regression is often confused with multiple regression.

Now, rewrite the multivariate linear model as

$$\underset{s\times 1}{\mathbf{Y}} = \underset{s\times 1}{\boldsymbol{\mu}} + \underset{s\times r}{\mathbf{C}}\ \underset{r\times 1}{\mathbf{X}} + \underset{s\times 1}{\mathcal{E}}, \tag{6}$$

where the rank constraint is

$$\mathrm{rank}(\mathbf{C}) = t \le \min(r, s). \tag{7}$$

Equations (6) and (7) form the multivariate reduced-rank regression model. When the rank condition (7) holds, there exist two (nonunique) full-rank matrices $\mathbf{A}$ and $\mathbf{B}$, where $\mathbf{A}$ is an $(s \times t)$-matrix and $\mathbf{B}$ is a $(t \times r)$-matrix, such that

$$\underset{s\times r}{\mathbf{C}} = \underset{s\times t}{\mathbf{A}}\ \underset{t\times r}{\mathbf{B}}. \tag{8}$$

The multivariate reduced-rank regression model can now be written as

$$\underset{s\times 1}{\mathbf{Y}} = \underset{s\times 1}{\boldsymbol{\mu}} + \underset{s\times t}{\mathbf{A}}\ \underset{t\times r}{\mathbf{B}}\ \underset{r\times 1}{\mathbf{X}} + \underset{s\times 1}{\mathcal{E}}. \tag{9}$$

The rank condition has been embedded into the regression model. The goal is to estimate $\boldsymbol{\mu}$, $\mathbf{A}$, and $\mathbf{B}$ (and, hence, $\mathbf{C}$).

Let $\boldsymbol{\Gamma}$ be a positive-definite symmetric $(s \times s)$-matrix of weights. The weighted least-squares estimates of $(\boldsymbol{\mu}, A, B)$ are

$$(\boldsymbol{\mu}^*, \mathbf{A}^*, \mathbf{B}^*) = arg \min_{\boldsymbol{\mu},\mathbf{A},\mathbf{B}} \mathrm{E}\{(\mathbf{Y} - \boldsymbol{\mu} - \mathbf{AB}\mathbf{X})^\tau \boldsymbol{\Gamma} (\mathbf{Y} - \boldsymbol{\mu} - \mathbf{AB}\mathbf{X})\} \tag{10}$$

where

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}_Y - \mathbf{AB}\boldsymbol{\mu}_X \tag{11}$$

$$\mathbf{A}^* = \boldsymbol{\Gamma}^{-1/2}\mathbf{V} \tag{12}$$

$$\mathbf{B}^* = \mathbf{V}^\tau \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1}, \tag{13}$$

and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_t)$ is an $(s \times t)$-matrix, where the jth column, $\mathbf{v}_j$, is the eigenvector corresponding to the jth largest eigenvalue, λ_j, of the $(s \times s)$ symmetric matrix,

$$\boldsymbol{\Gamma}^{1/2} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Gamma}^{1/2}. \tag{14}$$

The multivariate reduced-rank regression coefficient matrix $\mathbf{C}$ with rank t is, therefore, given by

$$\mathbf{C}^* = \boldsymbol{\Gamma}^{-1/2} \left(\sum_{j=1}^{t} \mathbf{v}_j \mathbf{v}_j^\tau \right) \boldsymbol{\Gamma}^{1/2} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1}. \tag{15}$$

The minimum achieved is $tr\{\boldsymbol{\Sigma}_{YY}\boldsymbol{\Gamma}\} - \sum_{j=1}^{t} \lambda_j$.

The main reason that multivariate reduced-rank regression is so important is that it contains as special cases the classical statistical techniques of ▶principal component analysis, canonical variate and correlation analysis (see ▶Discriminant Analysis: An Overview, and ▶Discriminant Analysis: Issues and Problems), linear discriminant analysis, exploratory factor analysis, multiple correspondence analysis, and other linear methods of analyzing multivariate data. It is also closely related to artificial neural network models and to cointegration in the econometric literature.

For example, the special cases of principal component analysis, canonical variate and correlation analysis, and linear discriminant analysis are given by the following choices: For *principal component analysis*, set $\mathbf{X} \equiv \mathbf{Y}$

and $\mathbf{\Gamma} = \mathbf{I}_s$; for *canonical variate and correlation analysis*, set $\mathbf{\Gamma} = \mathbf{\Sigma}_{YY}^{-1}$; for *linear discriminant analysis*, use the canonical-variate analysis choice of $\mathbf{\Gamma}$ and set $\mathbf{Y}$ to be a vector of binary variables whose component values (0 or 1) indicate the group or class to which an observation belongs. Details of these and other special cases can be found in Izenman (2008). If the elements of $\mathbf{\Sigma}_{ZZ}$ in (1) are unknown, as will happen in most practical problems, they have to be estimated using sample data on $\mathbf{Z}$.

The relationships between multivariate reduced-rank regression and the classical linear dimensionality reduction techniques become more interesting when the meta-parameter t is unknown and has to be estimated. The value of t is called the *effective dimensionality* of the multivariate regression (Izenman 1980). Estimating t is equivalent to the classical problems of determining the number of principal components to retain, the number of canonical variate to retain, or the number of linear discriminant functions necessary for classification purposes. Graphical methods for estimating t include the scree plot, the rank trace plot, and heatmap plots. Formal hypothesis tests have also been developed for estimating t.

When the number of variables is greater than the number of observations, some adjustments to the results have to be made to ensure that $\mathbf{\Sigma}_{XX}$ and $\mathbf{\Sigma}_{YY}$ can be inverted. One simple way of doing this is to replace $\mathbf{\Sigma}_{XX}$ by $\mathbf{\Sigma}_{XX} + \delta\mathbf{I}_r$ and to replace $\mathbf{\Sigma}_{YY}$ by $\mathbf{\Sigma}_{YY} + \kappa\mathbf{I}_s$ as appropriate, where $\delta > 0$ and $\kappa > 0$. Other methods, including regularization, banding, tapering, and thresholding, have been studied for estimating large covariance matrices and can be used here as appropriate.

The multivariate reduced-rank regression model can also be developed for the case of nonstochastic (or fixed) predictor variables.

The multivariate reduced-rank regression model has its origins in Anderson (1951), Rao (1964, 1965), and Brillinger (1969), and its name was coined by Izenman (1972, 1975). For the asymptotic distribution of the estimated reduced-rank regression coefficient matrix, see Anderson (1999), who gives results for both the random-$\mathbf{X}$ and fixed-$\mathbf{X}$ cases. Additional references are the monographs by van der Leeden (1990) and Reinsel and Velu (1998).

About the Author

Professor Izenman was Director of the Statistics and Probability Program at the National Science Foundation (1992–1994). He has been an Associate Editor of the *Journal of the American Statistical Association*. He is Associate Editor of the journals *Law, Probability, and Risk* and *Statistical Analysis and Data Mining*. He is a Fellow of the American Statistical Association. He was Vice-President, ASA Philadelphia Chapter (1987–1988).

Cross References

►Canonical Correlation Analysis
►Discriminant Analysis: An Overview
►Multivariate Rank Procedures: Perspectives and Prospectives
►Multivariate Statistical Analysis
►Principal Component Analysis

References and Further Reading

Anderson TW (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. Ann Math Stat 22:327–351

Anderson TW (1999) Asymptotic distribution of the reduced-rank regression estimator under general conditions. Ann Stat 27:1141–1154

Brillinger DR (1969) The canonical analysis of stationary time series. In: Multivariate analysis II, Krishaiah PR (ed) Academic, New York, pp 331–350

Izenman AJ (1972) Reduced-rank regression for the multivariate linear model, its relationship to certain multivariate techniques, and its application to the analysis of multivariate data, Ph.D. dissertation, University of California, Berkeley

Izenman AJ (1975) Reduced-rank regression for the multivariate linear model. J Multivariate Anal 5:248–264

Izenman AJ (1980) Assessing dimensionality in multivariate regression. In: Handbook of statistics I, Krishnaiah PR (ed) North-Holland, Amsterdam, pp 571–591

Izenman AJ (2008) Modern multivariate statistical techniques: regression, classification, and manifold learning. Springer, New York

Rao CR (1964) The use and interpretation of principal components in applied research. Sankhya A 26:329–358

Rao CR (1965) Linear statistical inference and its applications. Wiley, New York

Reinsel GC, Velu RP (1998) Multivariate reduced-rank regression, Lecture notes in statistics, vol 136, Springer, New York

Van der Leeden R (1990) Reduced-rank regression with structured residuals. DSWO, Leiden

Multivariate Statistical Analysis

Nanny Wermuth
Professor of Statistics
Chalmers Technical University/University of Gothenburg, Gothenburg, Sweden

Classical multivariate statistical methods concern models, distributions and inference based on the Gaussian distribution. These are the topics in the first textbook for mathematical statisticians by T. W. Anderson that was published in 1958 and that appeared as a slightly expanded 3rd edition in 2003. Matrix theory and notation is used

there extensively to efficiently derive properties of the multivariate Gaussian or the Wishart distribution, of principal components, of canonical correlation and discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems) and of the general multivariate linear model in which a Gaussian response vector variable Y_a has linear least-squares regression on all components of an explanatory vector variable Y_b.

In contrast, many methods for analyzing sets of observed variables have been developed first within special substantive fields and some or all of the models in a given class were justified in terms of probabilistic and statistical theory much later. Among them are factor analysis (see ►Factor Analysis and Latent Variable Modelling), path analysis, ►structural equation models, and models for which partial-least squares estimation have been proposed. Other multivariate techniques such as cluster analysis (see ►Cluster Analysis: An Introduction) and ►multidimensional scaling have been often used, but the result of such an analysis cannot be formulated as a hypothesis to be tested in a new study and satisfactory theoretical justifications are still lacking.

Factor analysis was proposed by psychologist C. Spearman (1904), (1926) and, at the time, thought of as a tool for measuring human intelligence. Such a model has one or several latent variables. These are hidden or unobserved and are to explain the observed correlations among a set of observed variables, called items in that context. The difficult task is to decide how many and which of a possibly large set of items to include into a model. But, given a set of latent variables, a classical factor analysis model specifies for a joint Gaussian distribution mutual independence of the observed variables given the latent variables. This can be recognized to be one special type of a graphical Markov model; see Cox and Wermuth (1996), Edwards (2000), Lauritzen (1996), Whittaker (1990).

Path analysis was developed by geneticist S. Wright (1923), (1934) for systems of linear dependence of variables with zero mean and unit variance. He used what we now call directed acyclic graphs to represent hypotheses of how the variables he was studying could have been generated. He compared correlations implied for missing edges in the graph with corresponding observed correlations to test the goodness of fit of such a hypothesis.

By now it is known, under which condition for these models in standardized Gaussian variables, maximum-likelihood estimates of correlations coincide with Wright's estimates via path coefficients. The condition on the graph is simple: there should be no three-node-two-edge subgraph of the following kind $\circ\!\!\longrightarrow\!\circ\longleftarrow\!\!\circ$. Then, the directed acyclic graph is said to be decomposable and captures the same independences as the concentration graph obtained by replacing each arrow by an undirected edge. In such Gaussian concentration graph models, estimated variances are matched to the observed variances so that estimation of correlations and variances is equivalent to estimation of covariances and variances.

Wright's method of computing implied path coefficients by "tracing paths" has been generalized via a so-called separation criterion. This criterion, given by Geiger, Verma and Pearl (1990), permits to read off a directed acyclic graph all independence statements that are implied by the graph. The criterion takes into account that not only ignoring (marginalizing over) variables might destroy an independence, but also conditioning on common responses may render two formerly independent variables to be dependent. In addition, the separation criterion holds for any distribution generated over the graph.

The separation criterion for directed acyclic graphs has been translated into conditions for the presence of edge-inducing paths in the graph; see Marchetti and Wermuth (2009). Such an edge-inducing path is also association-inducing in the corresponding model, given some mild conditions on the graph and on the distributions generated over it; see Wermuth (2010). In the special case of only marginalizing over linearly related variables, these induced dependences coincide with the path-tracing results given by Wright provided the directed acyclic graph model is decomposable and the variables are standardized to have zero means and unit variances. This applies not only to Gaussian distributions but also to special distributions of symmetric binary variables; see Wermuth et al. (2009).

Typically however, directed acyclic graph models are defined for unstandardized random variables of any type. Then, most dependences are no longer appropriately represented by linear regression coefficients or correlations, but maximum-likelihood estimates of all measures of dependence can still be obtained by separately maximizing each univariate conditional distribution, provided only that its parameters are variation-independent from parameters of distributions in the past.

Structural equation models, developed in econometrics, can be viewed as another extension of Wright's path analyses. The result obtained by T. Haavelmo (1943) gave an important impetus. For his insight that separate linear least-squares estimation may be inappropriate for equations having strongly correlated residuals, Haavelmo received a Nobel prize in 1989. It led to a class of models defined by linear equations with correlated residuals and to responses called endogenous. Other variables conditioned on and considered to be predetermined were named

exogenous. Vigorous discussions of estimation methods for structural equations occurred during the first few Berkeley symposia on mathematical statistics and probability from 1945 to 1965.

Path analysis and structural equation models were introduced to sociological research via the work by O.D. Duncan (1966, 1975). Applications of structural equation models in psychological and psychometric research resulted from cooperations between A. Goldberger and K. Jöreskog; see Goldberger (1971, 1972) and Jöreskog (1973, 1981). The methods became widely used once a corresponding computer program for estimation and tests was made available; see also Kline (2010).

In 1962, A. Zellner published his results on seemingly unrelated regressions. He points out that two simple regression equations are not separate if the two responses are correlated and that two dependent endogenous variables need to be considered jointly and require simultaneous estimation methods. These models are now recognized as special cases of both linear structural equations and of multivariate regression chains, a subclass of graphical Markov models; see Cox and Wermuth (1993), Drton (2009), Marchetti and Lupparelli (2010).

But it was not until 40 years later, that a maximum-likelihood solution for the Gaussian distribution in four variables, split into a response vector Y_a and vector variable Y_b, was given and an example of a poorly fitting data set with very few observations for which the likelihood equations have two real roots; see Drton and Richardson (2004). For well-fitting data and reasonably large sample sizes, this is unlikely to happen; see Sundberg (2010). For such situations, a close approximation to the maximum-likelihood estimate has been given in closed form for the seemingly unrelated regression model, exploiting that it is a reduced model to the covering model that has closed-form maximum-likelihood estimates, the general linear model of Y_a given Y_b; see Wermuth et al. (2006), Cox and Wermuth (1990).

For several discrete random variables of equal standing, i.e., without splits into response and explanatory variables, maximum-likelihood estimation was developed under different conditional independence constraints in a path-breaking paper by M. Birch (1963). This led to the formulation of general log-linear models, which were studied intensively among others by Haberman (1974), Bishop et al. (1975), Sundberg (1975) and by L. Goodman, as summarized in a book of his main papers on this topic, published in 1978. His work was motivated mainly by research questions from the social and medical sciences.

For several Gaussian variables of equal standing, two different approaches to reducing the number of parameters in a model, were proposed at about the same time. T. W. Anderson put structure on the covariances, the moment parameters of a joint Gaussian distribution and called the resulting models, hypotheses linear in covariances; see Anderson (1973), while A. P. Dempster put structure on the canonical parameters with zero constraints on concentrations, the off-diagonal elements of the inverse of a covariance matrix, and called the resulting models covariance selection models; see Dempster (1972).

Nowadays, log-linear models and covariance selection models are viewed as special cases of concentration graph models and zero constraints on the covariance matrix of a Gaussian distribution as special cases of covariance graph models. Covariance and concentration graph models are graphical Markov models with undirected graphs capturing independences. A missing edge means marginal independence in the former and conditional independence given all remaining variables in the latter; see also Wermuth and Lauritzen (1990), Wermuth and Cox (1998), (2004), Wermuth (2010).

The largest known class of Gaussian models that is in common to structural equation models and to graphical Markov models are the recursive linear equations with correlated residuals. These include linear summary graph models of Wermuth (2010), linear maximal ancestral graph of Richardson and Spirtes (2002), linear multivariate regression chains, and linear directed acyclic graph models. Deficiencies of some formulations start to be discovered by using algebraic methods. Identification is still an issue to be considered for recursive linear equations with correlated residuals, since so far only necessary or sufficient conditions are known but not both. Similarly, maximum-likelihood estimation still needs further exploration; see Drton et al. (2009).

For several economic time series, it became possible to judge whether such fluctuating series develop nevertheless in parallel, that is whether they represent cointegrating variables because they have a common stochastic trend. Maximum-likelihood analysis for cointegrating variables, formulated by Johansen (1988, 2009), has led to many important applications and insights; see also Hendry and Nielsen (2007).

Algorithms and corresponding programs are essential for any widespread use of multivariate statistical methods and for successful analyses. In particular, iterative proportional fitting, formulated by Bishop (1964) for log-linear models, and studied further by Darroch and Ratcliff (1972), was adapted to concentration graph models for CG (conditional Gaussian)-distributions (Lauritzen and Wermuth 1989) of mixed discrete and continuous variables by Frydenberg and Edwards (1989).

The EM (expectation-maximization)-algorithm of Dempster et al. (1977) was adapted to Gaussian directed

acyclic graph models with latent variables by Kiiveri (1987) and to discrete concentration graph models with missing observation by Lauritzen (1995).

With the TM-algorithm of Edwards and Lauritzen (2001), studied further by Sundberg (2002), maximum-likelihood estimation became feasible for all chain graph models called blocked concentration chains in the case these are made up of CG (conditional Gaussian)-regressions (Lauritzen and Wermuth 1989).

For multivariate regression chains of discrete random variables, maximum-likelihood estimation has now been related to the multivariate logistic link function by Marchetti and Lupparelli (2010), where these link functions provide a common framework and corresponding algorithm for ▶generalized linear models, which include among others linear, logistic and probit regressions as special cases; see McCullagh and Nelder (1989), Glonek and McCullagh (1995).

Even in linear models, estimation may become difficult when some of the explanatory variables are almost linear functions of others, that is if there is a problem of ▶multicollinearity. This appears to be often the case in applications in chemistry and in the environmental sciences. Thus, in connection with consulting work for chemists, Hoerl and Kennard (1970) proposed the use of ridge-regression (see ▶Ridge and Surrogate Ridge Regressions) instead of linear least-squares regression. This means for regressions of vector variable Y on X, to add to X^TX some positive constant k along the diagonal before matrix inversion to give as estimator $\tilde{\beta} = (kI + X^TX)^{-1}X^TY$.

Both ridge-regression and partial-least-squares, (see ▶Partial Least Squares Regression Versus Other Methods) proposed as an estimation method in the presence of latent variables by Wold (1980), have been recognized by Björkström and Sundberg (1999) to be shrinkage estimators and as such special cases of Tykhonov (1963) regularization.

More recently, a number of methods have been suggested which combine adaptive skrinkage methods with variable selection. A unifying approach which includes the least-squares estimator, shrinkage estimators and various combinations of variable selection and shrinkage has recently been given via a least squares approximation by Wang and Leng (2007). Estimation results depend necessarily on the chosen formulations and the criteria for shrinking dependences and for selecting variables.

Many more specialized algorithms and programs have been made available within the open access programming environment R, also those aiming to analyze large numbers of variables for only few observed individuals. It remains to be seen, whether important scientific insights will be gained by their use.

About the Author

Dr Nanny Wermuth is Professor of Statistics, at the joint Department of Mathematical Sciences of Chalmers Technical University and the University of Gothenburg. She is a Past President, Institute of Mathematical Statistics (2008–2009) and Past President of the International Biometric Society (2000–2001). In 1992 she was awarded a Max Planck-Research Prize, jointly with Sir David Cox. She chaired the Life Science Committee of the International Statistical Institute (2001–2005) and was an Associate editor of the *Journal of Multivariate Analysis* (1998–2001) and *Bernoulli* (2007–2010). Professor Wermuth is an Elected member of the German Academy of Sciences and of the International Statistical Institute (1982), an elected Fellow of the American Statistical Association (1989), and of the Institute of Mathematical Statistics (2001). She is a co-author (with David R. Cox) of the text *Multivariate dependencies: models, analysis and interpretation* (Chapman and Hall, 1996).

Cross References

▶Canonical Correlation Analysis
▶Cluster Analysis: An Introduction
▶Correspondence Analysis
▶Discriminant Analysis: An Overview
▶Discriminant Analysis: Issues and Problems
▶Factor Analysis and Latent Variable Modelling
▶General Linear Models
▶Likelihood
▶Logistic Regression
▶Multidimensional Scaling
▶Multidimensional Scaling: An Introduction
▶Multivariate Analysis of Variance (MANOVA)
▶Multivariate Data Analysis: An Overview
▶Multivariate Normal Distributions
▶Multivariate Rank Procedures: Perspectives and Prospectives
▶Multivariate Reduced-Rank Regression
▶Multivariate Statistical Process Control
▶Multivariate Technique: Robustness
▶Partial Least Squares Regression Versus Other Methods
▶Principal Component Analysis
▶R Language
▶Ridge and Surrogate Ridge Regressions
▶Structural Equation Models

References and Further Reading

Anderson TW (1958) An introduction to multivariate statistical analysis. Wiley, New York; (2003) 3rd edn. Wiley, New York

Anderson TW (1973) Asymptotically efficient estimation of covariance matrices with linear structure. Ann Stat 1:135–141

Birch MW (1963) Maximum likelihood in three-way contingency tables. J Roy Stat Soc B 25:220–233

Bishop YMM (1967) Multidimensional contingency tables: cell estimates. Ph.D. dissertation, Department of Statistics, Harvard University

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis: theory and practice. MIT Press, Cambridge

Björkström A, Sundberg R (1999) A generalized view on continuum regression. Scand J Stat 26:17–30

Cox DR, Wermuth N (1990) An approximation to maximum-likelihood estimates in reduced models. Biometrika 77:747–761

Cox DR, Wermuth N (1993) Linear dependencies represented by chain graphs (with discussion). Stat Sci 8:204–218; 247–277

Cox DR, Wermuth N (1996) Multivariate dependencies: models, analysis, and interpretation. Chapman & Hall, London

Darroch JN, Ratcliff D (1972) Generalized iterative scaling for log-linear models. Ann Math Stat 43:1470–1480

Dempster AP (1972) Covariance selection Biometrics 28:157–175

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39:1–38

Drton M (2009) Discrete chain graph models. Bernoulli 15:736–753

Drton M, Richardson TS (2004) Multimodality of the likelihood in the bivariate seemingly unrelated regression model. Biometrika 91:383–392

Drton M, Eichler M, Richardson TS (2009) Computing maximum likelihood estimates in recursive linear models. J Mach Learn Res 10:2329–2348

Duncan OD (1966) Path analysis: sociological examples. Am J Sociol 72:1–12

Duncan OD (1975) Introduction to structural equation models. Academic, New York

Edwards D (2000) Introduction to graphical modelling, 2nd edn. Springer, New York

Edwards D, Lauritzen SL (2001) The TM algorithm for maximising a conditional likelihood function. Biometrika 88:961–972

Frydenberg M, Edwards D (1989) A modified iterative proportional scaling algorithm for estimation in regular exponential families. Comput Stat Data Anal 8:143–153

Frydenberg M, Lauritzen SL (1989) Decomposition of maximum likelihood in mixed interaction models. Biometrika 76:539–555

Geiger D, Verma TS, Pearl J (1990) Identifying independence in Bayesian networks. Networks 20:507–534

Glonek GFV, McCullagh P (1995) Multivariate logistic models. J Roy Stat Soc B 57:533–546

Goldberger AS (1971) Econometrics and psychometrics: a survey of communalities. Psychometrika 36:83–107

Goldberger AS (1972) Structural equation methods in the social sciences. Econometrica 40:979–1002

Goodman LA (1978) Analyzing qualitative/categorical data. Abt Books, Cambridge

Haberman SJ (1974) The analysis of frequency data. University of Chicago Press, Chicago

Haavelmo T (1943) The statistical implications of a system of simultaneous equations. Econometrica 11:1–12; Reprinted in: Hendry DF, Morgan MS (eds) (1995) The foundations of econometric analysis. Cambridge University Press, Cambridge

Hendry DF, Nielsen B (2007) Econometric modeling: a likelihood approach. Princeton University Press, Princeton

Hoerl AE, Kennard RN (1970) Ridge regression. Biased estimation for non-orthogonal problems. Technometrics 12:55–67

Johansen S (1988) Statistical analysis of cointegration vectors. J Econ Dyn Contr 12:231–254; Reprinted in: Engle RF, Granger CWJ (eds) (1991) Long-run economic relationships, readings in cointegration. Oxford University Press, Oxford, pp 131–152

Johansen S (2009) Cointegration: overview and development. In: Handbook of financial time series, Andersen TG, Davis R, Kreiss J-P, Mikosch T (eds), Springer, New York, pp 671–693

Jöreskog KG (1973) A general method for estimating a linear structural equation system. In: Structural equation models in the social sciences, Goldberger AS, Duncan OD (eds), Seminar, New York, pp 85–112

Jöreskog KG (1981) Analysis of covariance structures. Scan J Stat 8:65–92

Kiiveri HT (1987) An incomplete data approach to the analysis of covariance structures. Psychometrika 52:539–554

Kline RB (2010) Principles and practice of structural equation modeling, 3rd edn. Guilford, New York

Lauritzen SL (1995) The EM-algorithm for graphical association models with missing data. Comp Stat Data Anal 1:191–201

Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford

Lauritzen SL, Wermuth N (1989) Graphical models for association between variables, some of which are qualitative and some quantitative. Ann Stat 17:31–57

Marchetti GM, Lupparelli M (2010) Chain graph models of multivariate regression type for categorical data. Bernoulli, to appear and available on ArXiv, http://arxiv.org/abs/0906.2098v2

Marchetti GM, Wermuth N (2009) Matrix representations and independencies in directed acyclic graphs. Ann Stat 47:961–978

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall/CRC, Boca Raton

Richardson TS, Spirtes P (2002) Ancestral Markov graphical models. Ann Stat 30:962–1030

Spearman C (1904) General intelligence, objectively determined and measured. Am J Psych 15:201–293

Spearman C (1926) The abilities of man. Macmillan, New York

Sundberg R (1975) Some results about decomposable (or Markov-type) models for multidimensional contingency tables: distribution of marginals and partitioning of tests. Scand J Stat 2:71–79

Sundberg R (2002) The convergence rate of the TM algorithm of Edwards and Lauritzen. Biometrika 89:478–483

Sundberg R (2010) Flat and multimodal likelihoods and model lack of fit in curved exponential families. Scand J Stat, published online: 28 June 2010

Tikhonov AN (1963) Solution of ill-posed problems and the regularization method (Russian). Dokl Akad Nauk SSSR 153:49–52

Wang H, Leng C (2007) Unified lasso estimation via least square approximation. J Am Stat Assoc 102:1039–1048

Wermuth N (2010) Probability distributions with summary graph structure. Bernoulli, to appear and available on ArXiv, http://arxiv.org/abs/1003.3259

Wermuth N, Cox DR (1998) On association models defined over independence graphs. Bernoulli 4:477–495

Wermuth N, Cox DR (2004) Joint response graphs and separation induced by triangular systems. J Roy Stat Soc B 66:687–717
Wermuth N, Lauritzen SL (1990) On substantive research hypotheses, conditional independence graphs and graphical chain models (with discusssion). J Roy Stat Soc B 52:21–75
Wermuth N, Marchetti GM, Cox DR (2009) Triangular systems for symmetric binary variables. Electr J Stat 3:932–955
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester
Wold HOA (1954) Causality and econometrics. Econometrica 22:162–177
Wold HOA (1980) Model construction and evaluation when theoretical knowledge is scarce: theory and application of partial least squares. In: Evaluation of econometric models, Kmenta J, Ramsey J (eds), Academic, New York, pp 47–74
Wright S (1923) The theory of path coefficients: a reply to Niles' criticism. Genetics 8:239–255
Wright S (1934) The method of path coefficients. Ann Math Stat 5:161–215
Zellner A (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J Am Stat Assoc 57:348–368

Multivariate Statistical Distributions

Donald R. Jensen
Professor Emeritus
Virginia Polytechnic Institute and State University,
Blacksburg, VA, USA

Origins and Uses

Multivariate distributions (MDs) are defined on finite-dimensional spaces. Origins trace to early studies of ►*multivariate normal distributions* as models for dependent chance observations (Adrian 1808; Bravais 1846; Dickson 1886; Edgeworth 1892; Galton 1889; Gauss 1823; Helmert 1868; Laplace 1811; Pearson 1896; Plana 1813; Schols 1875; Spearman 1904; Student 1908); for two and three dimensions in Bravais (1846) and Schols (1875); and for finite dimensions in Edgeworth (1892) and Gauss (1823), advancing such now–familiar concepts as regression and partial correlation. Let $\boldsymbol{Y} = [Y_1, \ldots, Y_5]$ designate chance observations; in pharmacology as systolic (Y_1) and diastolic (Y_2) pressures, pulse rate (Y_3), and gross (Y_4) and fine (Y_5) motor skills. Strengths of materials may register moduli of elasticity (Y_1) and of rupture (Y_2), specific gravity (Y_3), coefficient of linear expansion (Y_4), and melting point (Y_5). A complete probabilistic description of each vector observation entails the joint distribution of $[Y_1, \ldots, Y_5]$. A sample of n such k–vectors, arranged as rows, yields a random matrix $\boldsymbol{Y} = [Y_{ij}]$ of order $(n \times k)$, its distribution supporting much of ►*multivariate statistical analysis*.

Beyond modeling chance outcomes, MDs describe probabilistic features of data–analytic operations, to include statistical inference, decision theory (see ►Decision Theory: An Introduction, and ►Decision Theory: An Overview), and other evidentiary analyses. In inference the frequentist seeks joint distributions (1) of multiparameter estimates, and (2) of statistics for testing multiple hypotheses, both parametric and nonparametric. Such distributions derive from observational models. Similarly, multiparameter Bayesian methods require MDs in modeling prior, contemporary, and posterior distributions for the parameters. In addition, MDs serve to capture dependencies owing to repeated measurements on experimental subjects. MDs derive from other distributions through transformations, projections, conditioning, convolutions, extreme values, mixing, compounding, truncating, and censoring. Specifically, experiments modeled conditionally in a random environment yield unconditional distributions as mixtures; see Everitt and Hand (1981), Lindsay (1995), McLachlan and Basford (1988), and Titterington et al. (1985). Random processes, to include such concepts as stationarity, are characterized through MDs as their finite–dimensional projections. Beyond probability, MD–theory occasionally supports probabilistic proofs for purely mathematical theorems. In short, MDs arise throughout statistics, applied probability, and beyond, and their properties are essential to understanding those fields.

In what follows $\mathbb{R}^k$, $\mathbb{R}^k_+$, $\mathbb{F}_{n\times k}$, $\mathbb{S}_k$, and $\mathbb{S}^+_k$ respectively designate Euclidean k–space, its positive orthant, the real $(n \times k)$ matrices, the real symmetric $(k \times k)$ matrices, and their positive definite varieties. Special arrays are $\boldsymbol{I}_k$, the $(k \times k)$ identity, and the diagonal matrix $\mathrm{Diag}(a_1, \ldots, a_k)$. The transpose, inverse, trace, and determinant of $\boldsymbol{A} \in \mathbb{F}_{k\times k}$ are $\boldsymbol{A}'$, $\boldsymbol{A}^{-1}$, $\mathrm{tr}(\boldsymbol{A})$, and $|\boldsymbol{A}|$, with $\boldsymbol{a}' = [a_1, \ldots, a_k]$ as the transpose of $\boldsymbol{a} \in \mathbb{R}^k$. For $\boldsymbol{Y} \in \mathbb{R}^k$ random, its expected vector, dispersion matrix, and law of distribution are $E(\boldsymbol{Y}) \in \mathbb{R}^k$, $V(\boldsymbol{Y}) \in \mathbb{S}^+_k$, and $\mathcal{L}(\boldsymbol{Y})$. Abbreviations include *pdf*, *pmf*, *cdf*, and *chf*, for probability density, probability mass, cumulative distribution, and ►characteristic functions, respectively.

Some Properties

MDs merit scrutiny at several levels. At one extreme are weak assumptions on existence of low-order moments, as in Gauss–Markov theory. At the other extremity are rigidly parametric models, having MDs of specified functional forms to be surveyed subsequently. In between are

Multivariate Statistical Distributions. Table 1 Examples of spherical distributions on $\mathbb{R}^n$ having density $f(\boldsymbol{x})$ or characteristic function $\xi(\boldsymbol{t})$; see Chmielewski (1981)

Density or *chf*	Comments	
Normal	$f(\boldsymbol{x}) = c_1 \exp(-\boldsymbol{x}'\boldsymbol{x}/2)$	$N_n(\boldsymbol{0}, \boldsymbol{I}_n)$
Pearson Type II	$f(\boldsymbol{x}) = c_2(1-\boldsymbol{x}'\boldsymbol{x})^{\gamma-1}$	$\gamma > 1$
Pearson Type VII	$f(\boldsymbol{x}) = c_3(1+\boldsymbol{x}'\boldsymbol{x})^{-\gamma}$	$\gamma > n/2$
Student t	$f(\boldsymbol{x}) = c_4(1+\nu^{-1}\boldsymbol{x}'\boldsymbol{x})^{-(\nu+n)/2}$	ν a positive integer
Cauchy	$f(\boldsymbol{x}) = c_5(1+\boldsymbol{x}'\boldsymbol{x})^{-(n+1)/2}$	Student t $\nu = 1$
Scale mixtures	$f(\boldsymbol{x}) = c_6 \int_0^\infty t^{-n/2} \exp(-\boldsymbol{x}'\boldsymbol{x}/2t) dG(t)$	$G(t)$ a *cdf*
Stable laws	$\xi(\boldsymbol{t}) = c_7 \exp[\gamma(\boldsymbol{t}'\boldsymbol{t})^{\alpha/2}]$	$0 < \alpha < 2; \gamma > 0$

classes of MDs exhibiting such common structural features as symmetry or unimodality, giving rise to *semiparametric* models of note. Of particular relevance are derived distributions that are unique to all members of an underlying class.

Specifically, distributions on $\mathbb{F}_{n\times k}$ in the class $\{L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}); \phi \in \Phi\}$ have *pdf*s as given in Table 3. Here $\boldsymbol{\Theta} \in \mathbb{F}_{n\times k}$ comprise location parameters; $\boldsymbol{\Gamma} \in \mathbb{S}_n^+$ and $\boldsymbol{\Sigma} \in \mathbb{S}_k^+$ are scale parameters; $\phi(\cdot)$ is a function on $\mathbb{S}_k^+$; and $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ is a factor of $\boldsymbol{\Sigma}^{-1}$. These distributions are invariant for $\boldsymbol{\Gamma} = \boldsymbol{I}_n$ in that $\mathcal{L}(\boldsymbol{Y} - \boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{Q}(\boldsymbol{Y} - \boldsymbol{\Theta}))$ for every real orthogonal matrix $\boldsymbol{Q}(n \times n)$. A subclass, taking $\phi(\boldsymbol{A}) = \psi(\mathrm{tr}(\boldsymbol{A}))$, with ψ defined on $[0,\infty)$, is $S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ as in Table 3. Here independence among rows of $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]'$ and multinormality are linked: If $\mathcal{L}(\boldsymbol{Y}) \in S_{n,k}(\boldsymbol{\Theta}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$, then $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ are mutually independent if and only if $\boldsymbol{Y}$ is matrix normal, namely $N_{n,k}(\boldsymbol{\Theta}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$ on $\mathbb{F}_{n\times k}$; see James (1954). A further subclass on $\mathbb{R}^n$, with $k = 1$ and $\boldsymbol{\Sigma}(1\times 1) = 1$, are the *elliptical distributions* on $\mathbb{R}^n$, namely, $\{\mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \psi); \psi \in \Psi\}$, with location-scale parameters $(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ and the typical *pdf* $f(\boldsymbol{y}) = |\boldsymbol{\Gamma}|^{-\frac{1}{2}} \psi((\boldsymbol{y} - \boldsymbol{\theta})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{\theta}))$. The foregoing all contain multivariate normal and heavy–tailed Cauchy models as special cases, and all have served as observational models *in lieu of* multivariate normality. In particular, $\{\mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{I}_n, \psi); \psi \in \Psi\}$ often serve as semiparametric surrogates for $N_n(\boldsymbol{\theta}, \boldsymbol{I}_n)$ in univariate samples, and $\{L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}); \phi \in \Phi\}$ in the analysis of multivariate data. Examples from $\{\mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{I}_n, \psi); \psi \in \Psi\}$ are listed in Table 1, cross-referenced as in Chmielewski (1981) to well-known distributions on $\mathbb{R}^1$.

Inferences built on these models often remain exact as for normal models, certifying their use as semiparametric surrogates. This follows from the invariance of stipulated derived distributions as in Jensen and Good (1981). Further details, for their use as observational models on $\mathbb{R}^k$ and $\mathbb{F}_{n\times k}$, for catalogs of related and derived distributions, and for the robustness of various inferential procedures, are found in Cambanis et al. (1981), Chmielewski (1981), Devlin et al. (1976), Fang and Anderson (1990), Fang et al. (1990), Fang and Zhang (1990), James (1954), and Kariya and Sinha (1989). Regarding $\{L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}); \phi \in \Phi\}$ and its extensions, see Dawid (1977), Dempster (1969), and Jensen and Good (1981). These facts bear heavily on the robustness and validity of normal-theory procedures for use with nonnormal data, including distributions having heavy tails. The cited distributions all exhibit symmetries, including symmetries under reflections. Considerable recent work addresses skewed MDs, often resulting from truncation; see Arnold and Beaver (2000), for example.

Properties of distributions on $\mathbb{R}^1$ often extend nonuniquely to the case of MDs. Concepts of unimodality on $\mathbb{R}^k$ are developed in Dharmadhikari and Joag-Dev (1988), some enabling a sharpening of joint Chebyshev bounds. Stochastic ordering on $\mathbb{R}^1$ likewise admits a multiplicity of extensions. These in turn support useful probability inequalities on $\mathbb{R}^k$ as in Tong (1980), many pertaining to distributions cited here. Let $\mu(\cdot)$ and $\nu(\cdot)$ be probability measures on $\mathbb{R}^k$, and $\mathcal{C}_k$ the compact convex sets in $\mathbb{R}^k$ symmetric under reflection about $\boldsymbol{0} \in \mathbb{R}^k$. The concentration ordering (Birnbaum 1948) on $\mathbb{R}^1$ is extended in Sherman (1904): $\mu(\cdot)$ is said to be *more peaked about* $\boldsymbol{0} \in \mathbb{R}^k$ than $\nu(\cdot)$ if and only if $\mu(A) \geq \nu(A)$ for every $A \in \mathcal{C}_k$. Specifically, let $P_\Sigma(\cdot; \psi)$ and $P_\Omega(\cdot; \psi)$ be probability measures for $\mathcal{S}_n(\boldsymbol{0}, \boldsymbol{\Sigma}, \psi)$ and $\mathcal{S}_n(\boldsymbol{0}, \boldsymbol{\Omega}, \psi)$. Then a necessary and sufficient condition that $P_\Sigma(\cdot; \psi)$ should be more peaked about $\boldsymbol{0}$ than $P_\Omega(\cdot; \psi)$, is that $(\boldsymbol{\Omega} - \boldsymbol{\Sigma}) \in \mathbb{S}_n^+$, sufficiency in Fefferman et al. (1972), necessity in Jensen (1984). Similar orderings apply when both $(\boldsymbol{\Sigma}, \psi)$ are allowed to vary (Jensen 1984), extending directly to include distributions in $\{S_{n,k}(\boldsymbol{0}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}, \psi); \psi \in \Psi\}$. Numerous further notions of stochastic orderings for MDs are treated in Shaked and Shanthikumar (2007).

Interest in MDs often centers on their dependencies. A burgeoning literature surrounds ▶*copulas*, expressing a joint distribution function in terms of its marginals, together with a finite–dimensional parameter quantifying the degree of dependence; see Nelsen (1998) for example. Further concepts of dependence, including notions rooted in the geometry of $\mathbb{R}^k$, are developed in Joe (1997).

The Basic Tools

Let $(\Omega, \mathfrak{B}, P)$ be a probability space, Ω an event set, $\mathfrak{B}$ a field of subsets of Ω, and P a probability measure. Given a set $\mathfrak{X}_0$, an $\mathfrak{X}_0$–valued random element is a measurable mapping $\boldsymbol{X}(\omega)$ from Ω to $\mathfrak{X}_0$, multivariate when $\mathfrak{X}_0$ is finite-dimensional, as $\mathbb{R}^k$, its *cdf* then given by $F(x_1, \ldots, x_k) = P(\omega : X_1(\omega) \le x_1, \ldots, X_k(\omega) \le x_k)$. To each *cdf* corresponds a P_X on $(\mathbb{R}^k, \mathfrak{B}_k, P_X)$ and conversely, with $\mathfrak{B}_k$ as a field of subsets of $\mathbb{R}^k$. Moreover, $\{P_X = a_1P_1 + a_2P_2 + a_3P_3; a_i \ge 0, a_1 + a_2 + a_3 = 1\}$ decomposes as a mixture: P_1 assigns positive probability to the mass points of P_X; P_2 is absolutely continuous with respect to Lebesgue (volume) measure on $(\mathbb{R}^k, \mathfrak{B}_X, \cdot)$; and P_3 is purely singular. Corresponding to $\{P_1, P_2, P_3\}$ are *cdf*s $\{F_1, F_2, F_3\}$: F_1 has a mass function (*pmf*) $p(x_1, \ldots, x_k) = P(X_1 = x_1, \ldots, X_k = x_k)$, giving jumps of F_1 at its mass points; F_2 has a *pdf* $f_2(x_1, \ldots, x_k) = \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F_2(x_1, \ldots, x_k)$ for almost all $\{x_1, \ldots, x_k\}$. The marginal *cdf* of $\boldsymbol{X}_1' = [X_1, \ldots, X_r]$ is $F_{m1}(x_1, \ldots, x_r) = F(x_1, \ldots, x_r, \infty, \ldots, \infty)$. With $\boldsymbol{X}_2' = [X_{r+1}, \ldots, X_k]$ and $\boldsymbol{x}_2' = [x_{r+1}, \ldots, x_k]$, the conditional *pmf* for $\mathcal{L}(\boldsymbol{X}_1 \mid \boldsymbol{x}_2)$, given that $\{\boldsymbol{X}_2 = \boldsymbol{x}_2\}$, is $p_{1\cdot 2}(x_1, \ldots, x_r) = \frac{p(x_1, \ldots, x_k)}{p_2(x_{r+1}, \ldots, x_k)}$ with $p_2(x_{r+1}, \ldots, x_k)$ as the marginal *pmf* for $\boldsymbol{X}_2$. A similar expression holds for P_2 in terms of the joint and marginal *pdf*s $f(x_1, \ldots, x_k)$ and $f_2(x_{r+1}, \ldots, x_k)$. As noted, F_1 is discrete and F_2 absolutely continuous, pure types to warrant their separate cataloging in the literature. On the other hand, P_3 is singular on a set in $\mathbb{R}^k$ having Lebesgue measure zero, often illustrated as a linear subspace. In contrast, P_3 is known to originate in practice through pairs $(\boldsymbol{X}, \boldsymbol{Y})$ as in Olkin and Tate (1961), such that $\boldsymbol{X}$ is multinomial and $\mathcal{L}(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$ is multivariate normal. Related studies are reported in a succession of articles including the recent (Bedrick et al. 2000).

The study of MDs draws heavily on the calculus of $\mathbb{R}^k$. Distributions not expressible in closed form may admit series expansions, asymptotic expansions of Cornish-Fisher and Edgeworth types, or large-sample approximations via central limit theory. Accuracy of the latter is gauged through Berry–Esséen bounds on rates of convergence, as developed extensively in Bhattacharya and Ranga Rao (1976) under moments of order greater than 2. Moreover, the integral transform pairs of Fourier, Laplace, and Mellin, including *chf*s on $\mathbb{R}^k$, are basic. Elementary operations in the space of transforms carry back to the space of distributions through inversion. Affine data transformations are intrinsic to the use of *chf*s of MDs, as treated extensively in Lukacs and Laha (1964). On the other hand, Mellin transforms couple nicely with such nonlinear operations as powers, products, and quotients of random variables, as treated in Epstein (1948) and Subrahmaniam (1970) and subsequently. In addition, functions generating joint moments, cumulants, factorial moments, and probabilities are used routinely. Projection methods determine distributions on $\mathbb{R}^k$ completely, via the one-dimensional distributions of every linear function. To continue, a property is said *to characterize* a distribution if unique to that distribution. A general treatise is Kagan et al. (1973), including reference to some MDs reviewed here.

We next undertake a limited survey of continuous and discrete MDs encountered with varying frequencies in practice. Developments are cited for random vectors and matrices. Continuing to focus on semiparametric models, we identify those distributions derived and unique to underlying classes of models, facts not widely accessible otherwise. The principal reference for continuous MDs is the encyclopedic (Kotz et al. 2000), coupled with monographs on multivariate normal (Tong 1990) and Student t (Kotz and Nadarajah 2004) distributions. For discrete MDs, encyclopedic accounts are archived in Johnson et al. (1997) and Patil and Joshi (1968).

Continuous Distributions

Central to classical **multivariate statistical analysis** are $\{N_{n,k}(\boldsymbol{\Theta}, \boldsymbol{I}_n, \boldsymbol{\Sigma}); n > k\}$ for $\mathcal{L}(\boldsymbol{Y})$, and the essential derived distribution $\mathcal{L}(\boldsymbol{W}) = W_k(n, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, with $\boldsymbol{W} = \boldsymbol{Y}'\boldsymbol{Y}$, as noncentral Wishart having n degrees of freedom, scale matrix $\boldsymbol{\Sigma}$, and noncentrality matrix $\boldsymbol{\Lambda} = \boldsymbol{\Theta}'\boldsymbol{\Theta}$, with central *pdf* as in Table 3.

Student tDistributions

Vector distributions. There are two basic types. Let $[Y_1, \ldots, Y_k]$ be multivariate normal with means $[\mu_1, \ldots, \mu_k]$, unit variances, and correlation matrix $\boldsymbol{R}(k \times k)$. A *Type I t distribution* is that of $\{T_j = Y_j/S; 1 \le j \le k\}$ such that $\mathcal{L}(\nu S^2) = \chi^2(\nu)$ independently of $[Y_1, \ldots, Y_k]$. Its central *pdf* is listed in Table 2. To continue, suppose that $\boldsymbol{S} = [S_{ij}]$ and $\mathcal{L}(\nu \boldsymbol{S}) = W_k(\nu, \boldsymbol{R})$, independently of $[Y_1, \ldots, Y_k]$. A *Type II t distribution* is that of $\{T_j = Y_j/S_{jj}; 1 \le j \le k\}$. Both types are central if and only if $\{\mu_1 = \cdots = \mu_k = 0\}$. These distributions arise in multiple comparisons, in the construction of rectangular confidence sets for means, in the Bayesian analysis of multivariate normal data, and in various multistage procedures. For further details see Kotz et al. (2000) and Tong (1990).

More generally, if $\mathcal{L}(X_1, \ldots, X_k, Z_1, \ldots, Z_\nu) = \mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ with $\boldsymbol{\theta}' = [\mu_1, \ldots, \mu_k, 0, \ldots, 0]$ and $\boldsymbol{\Gamma} = \mathrm{Diag}(\boldsymbol{R}, \boldsymbol{I}_\nu)$, then with $\nu S^2 = \left(Z_1^2 + \ldots + Z_\nu^2\right)$, the central distribution of $\{T_j = X_j/S; 1 \le j \le k\}$ is Type I multivariate t for all distributions in $\{\mathcal{S}_n(\boldsymbol{\theta}, \boldsymbol{\Gamma}, \psi); \psi \in \Psi\}$ as structured. Multiple comparisons using $\{T_1, \ldots, T_k\}$ under normality thus are

Multivariate Statistical Distributions. Table 2 Standard *pdf*s for some continuous distributions on $\mathbb{R}^k$

Type	Density	Comments
Student t	$k_1\left[1+\nu^{-1}(\boldsymbol{t}-\boldsymbol{\mu})'\boldsymbol{R}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-(\nu+k)/2}$	$\boldsymbol{t}\in\mathbb{R}^k$
Dirichlet	$k_2(1-\Sigma_1^k u_j)^{\alpha_0-1}\Pi_1^k u_j^{\alpha_j-1}$	$\{0\le u_j\le 1;\Sigma_1^k u_j\le 1\}$
Inv. Dirichlet	$k_3\Pi_1^k v_j^{\alpha_j-1}/[1+\Sigma_1^k v_j]^{\alpha/2}$	$\{0\le v_j<\infty;\alpha=\Sigma_0^k\alpha_j$
$\lvert\boldsymbol{W}-w\boldsymbol{\Sigma}\rvert=0$	$k_4\Pi_1^k w_i^{(\nu-k-1)/2}\Pi_{i<j}(w_i-w_j)e^{-\frac{1}{2}(\Sigma_1^k w_i)}$	$\{w_1>\cdots>w_k>0\}$
$\lvert\boldsymbol{S}_1-\ell\boldsymbol{S}_0\rvert=0$	$k_5\Pi_1^k \ell_i^{\frac{1}{2}(m-k-1)}\Pi_1^k(\ell_i+1)^{-(m+n)/2}\Pi_{i<j}(\ell_i-\ell_j)$	$\{\ell_1>\cdots>\ell_k>0\}$

Multivariate Statistical Distributions. Table 3 Standard *pdf*s for some continuous distributions on $\mathbb{R}^k$

Type	Density	Comments
$N_{n,k}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\Sigma})$	$\kappa_1\exp[-\frac{1}{2}\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{\Theta})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y}-\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-1}]$	$\boldsymbol{Y}\in\mathbb{F}_{n\times k}$
$L_{n,k}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\Sigma})$	$\kappa_2\lvert\boldsymbol{\Gamma}\rvert^{-\frac{k}{2}}\lvert\boldsymbol{\Sigma}\rvert^{-\frac{n}{2}}\phi(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{Y}-\boldsymbol{\Theta})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y}-\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-\frac{1}{2}})$	$\boldsymbol{Y}\in\mathbb{F}_{n\times k},\phi\in\Phi$
$S_{n,k}(\boldsymbol{\Theta},\boldsymbol{\Gamma},\boldsymbol{\Sigma})$	$\kappa_3\lvert\boldsymbol{\Gamma}\rvert^{-\frac{k}{2}}\lvert\boldsymbol{\Sigma}\rvert^{-\frac{n}{2}}\psi(\mathrm{tr}(\boldsymbol{Y}-\boldsymbol{\Theta})'\boldsymbol{\Gamma}^{-1}(\boldsymbol{Y}-\boldsymbol{\Theta})\boldsymbol{\Sigma}^{-1})$	ψ on $[0,\infty)$
Wishart	$\kappa_4\lvert\boldsymbol{W}\rvert^{(\nu-k-1)/2}\exp(-\frac{1}{2}\mathrm{tr}\boldsymbol{W}\boldsymbol{\Sigma}^{-1})$	$\boldsymbol{W}\in\mathbb{S}_k^+$
Gamma Hsu (l940)	$\kappa_5\lvert\boldsymbol{W}\rvert^{(n-k-1)/2}\phi(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{\Sigma}^{-\frac{1}{2}})$	$\phi\in\Phi,\boldsymbol{W}\in\mathbb{S}_k^+$
Gamma Lukacs and Laha (1964)	$\kappa_6\lvert\boldsymbol{W}\rvert^{\lambda-1}\exp(-\mathrm{tr}\boldsymbol{W}\boldsymbol{\Sigma}^{-1})$	$\lambda>0,\boldsymbol{W}\in\mathbb{S}_k^+$
Matric T	$\kappa_7\lvert\boldsymbol{I}_k-\nu^{-1}\boldsymbol{T}'\boldsymbol{T}\rvert^{-(\nu+r)/2}$	$\boldsymbol{T}\in\mathbb{F}_{r\times k}$
Dirichlet	$\kappa_8\Pi_1^k\lvert\boldsymbol{W}_j\rvert^{(\nu_j-k-1)/2}\lvert\boldsymbol{I}_k-\Sigma_1^k\boldsymbol{W}_j\rvert^{(\nu_0-k-1)/2}$	$f(\boldsymbol{W}_1,\ldots,\boldsymbol{W}_k)$
Inv. Dirichlet	$\kappa_9\Pi_1^k\lvert\boldsymbol{V}_j\rvert^{(\nu_j-k-1)/2}\lvert\boldsymbol{I}_k+\Sigma_1^k\boldsymbol{V}_j\rvert^{(\nu_T-k-1)/2}$	$f(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_k)$

exact in level for linear models having spherical errors (Jensen 1979). Similarly, if $\mathcal{L}(\boldsymbol{Y})=S_{n,k}(\boldsymbol{\Theta},\boldsymbol{I}_n,\boldsymbol{\Sigma})$ with parameters $\boldsymbol{\Theta}=[\boldsymbol{\theta},\ldots,\boldsymbol{\theta}]'$, $\boldsymbol{\theta}\in\mathbb{R}^k$; if $X_j=n^{1/2}\bar{Y}_j$ with $\{\bar{Y}_j=(Y_{1j}+\cdots+Y_{nj})/n;1\le j\le k\}$; and if $\boldsymbol{S}$ is the sample dispersion matrix; then the central distribution of $\left\{T_j=X_j/S_{jj}^{\frac{1}{2}};1\le j\le k\right\}$ is Type II multivariate t for every $\mathcal{L}(\boldsymbol{Y})$ in $\{S_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma},\psi);\psi\in\Psi\}$. Noncentral distributions generally depend on the particular distribution $\mathcal{S}_n(\boldsymbol{\theta},\boldsymbol{\Gamma})$ or $S_{n,k}(\boldsymbol{\Theta},\boldsymbol{I}_n,\boldsymbol{\Sigma})$.

Matric T distributions. Let $\boldsymbol{Y}$ and $\boldsymbol{W}$ be independent, $\mathcal{L}(\boldsymbol{Y})=N_{r,k}(\boldsymbol{0},\boldsymbol{I}_r,\boldsymbol{\Sigma})$ and $\mathcal{L}(\boldsymbol{W})=W_k(\nu,\boldsymbol{\Sigma})$ such that $\nu\ge k$, and let $\boldsymbol{T}=\boldsymbol{Y}\boldsymbol{W}^{-\frac{1}{2}}$ using any factorization $\boldsymbol{W}^{-\frac{1}{2}}$ of $\boldsymbol{W}^{-1}$. Then $\mathcal{L}(\boldsymbol{T})$ is *matric t* with *pdf* as in Table 3. Alternatively, consider $\boldsymbol{X}'=[\boldsymbol{Y}',\boldsymbol{Z}']$ with distribution $S_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma})$ such that $n=r+\nu$ and $\nu\ge k$, and again let $\boldsymbol{T}=\boldsymbol{Y}\boldsymbol{W}^{-\frac{1}{2}}$ but now with $\boldsymbol{W}=\boldsymbol{Z}'\boldsymbol{Z}$. These variables arise from distributions $S_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma})$ in the same manner as for $N_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma})$. Then $\boldsymbol{T}$ has a matric t distribution for every distribution $\mathcal{L}(\boldsymbol{Y})$ in $\{S_{n,k}(\boldsymbol{0},\boldsymbol{I}_n,\boldsymbol{\Sigma},\psi);\psi\in\Psi\}$. This property transfers directly to $\mathcal{L}(\boldsymbol{ATB})$ as in Dickey (1967) with $\boldsymbol{A}$ and $\boldsymbol{B}$ nonsingular.

Gamma Distributions

Vector Distributions. Extract $\mathrm{Diag}(W_{11},\ldots,W_{kk})$ from $\boldsymbol{W}=[W_{ij}]$. Their joint distributions arise in the analysis of nonorthogonal designs, in time-series, in multiple comparisons, in the analysis of multidimensional contingency tables, in extensions of Friedman's χ^2 test in two-way data based on ranks, and elsewhere. There is a gamma distribution on $\mathbb{R}_+^k$ for diagonals of the matrix Gamma (Lukacs and Laha 1964) of Table 3; k–variate χ^2 when $\boldsymbol{W}$ is Wishart; see Kibble (1941) for $k=2$; and a k–variate exponential distribution for the case $n=2$. Rayleigh distributions $\mathcal{L}\left(W_{11}^{\frac{1}{2}},W_{22}^{\frac{1}{2}},\ldots,W_{kk}^{\frac{1}{2}}\right)$ on $\mathbb{R}_+^k$ support the detection of signals from noise (Miller 1975); more general such distributions are known (Jensen 1970a); as are more general χ^2 distributions on $\mathbb{R}^k$ having differing marginal degrees of freedom (Jensen 1970b). Densities here are typically intractable, often admitting multiple series expansions in special functions. Details are given in Kotz et al. (2000). As $n\to\infty$, the χ^2 and Rayleigh distributions on $\mathbb{R}_+^k$ are multinormal in the limit, for central and noncentral cases alike, whereas for fixed n, the limits as noncentrality parameters

grow again are multivariate normal (Jensen 1969). Alternative approximations, through normalizing Wilson-Hilferty transformations, are given in Jensen (1976) and Jensen and Solomon (1994).

Matric distributions. Let $\mathcal{L}(\boldsymbol{Y}) \in L_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \phi)$ with $n \geq k$; the *pdf* of $\boldsymbol{W} = \boldsymbol{Y}'\boldsymbol{Y}$ is given in Table 3 under Gamma (Hsu 1940) as in that reference. The *pdf* under Gamma (Lukacs and Laha 1964), with $\lambda > 0$, reduces to that of a scaled Wishart matrix when 2λ is an integer. The noncentral Wishart *pdf* with $\boldsymbol{\Lambda} \neq \boldsymbol{0}$ admits series expansions in special polynomials. Moreover, as $n \to \infty$, for fixed $\boldsymbol{\Lambda}$ its limit distribution is multinormal, and for fixed n, its ▶asymptotic normality attains as the noncentrality parameters grow in a specified manner (Jensen 1976). Wishart matrices arise in matrix normal samples, e.g., as scaled sample dispersion matrices, and otherwise throughout multivariate distribution theory. Parallel remarks apply for Gamma (Hsu 1940) of Table 3 when the underlying observational model belongs to $\{L_{n,k}(\boldsymbol{\Theta}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \phi); \phi \in \Phi\}$.

Dirichlet Distributions

If X and Y are independent gamma variates having a common scale, then $U = X/(X + Y)$ and $V = X/Y$ have *beta* and *inverted beta* distributions, respectively, the scaled Snedecor-Fisher F specializing from the latter. This section treats vector and matrix versions of these.

Vector distributions. Let $\{Z_0, \ldots, Z_k\}$ be independent gamma variates with common scale and the shape parameters $\{\alpha_0, \ldots, \alpha_k\}$, and let $T = (Z_0 + \cdots + Z_k)$. Then the joint distribution of $\{U_j = Z_j/T; 1 \leq j \leq k\}$ is the k-dimensional *Dirichlet distribution* $D(\alpha_0, \ldots, \alpha_k)$ with *pdf* as given in Table 2. An important special case is that $\{\alpha_j = \nu_j/2; 0 \leq j \leq k\}$ with $\{\nu_0, \ldots, \nu_k\}$ as positive integers and with $\{Z_0, \ldots, Z_k\}$ as independent χ^2 variates. However, in this case neither χ^2 nor independence is required. For if $\boldsymbol{y} = \left[\boldsymbol{y}_0', \boldsymbol{y}_1', \ldots, \boldsymbol{y}_k'\right]' \in \mathbb{R}^n$ with $\{\boldsymbol{y}_j \in \mathbb{R}^{\nu_j}; 0 \leq j \leq k\}$ and $n = \nu_0 + \cdots + \nu_k$ such that $\mathcal{L}(\boldsymbol{y}) = \mathcal{S}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, then $\left\{U_j = \boldsymbol{y}_j'\boldsymbol{y}_j/T; 1 \leq j \leq k\right\}$, but now with $T = \boldsymbol{y}_0'\boldsymbol{y}_0 + \boldsymbol{y}_1'\boldsymbol{y}_1 + \cdots + \boldsymbol{y}_k'\boldsymbol{y}_k$, has the distribution $D(\nu_0/2, \nu_1/2, \ldots, \nu_k/2)$ for all such $\mathcal{L}(\boldsymbol{y}) \in \{\mathcal{S}_n(\boldsymbol{0}, \boldsymbol{I}_n, \psi); \psi \in \Psi\}$.

The *inverted* Dirichlet is that of $\{V_j = Z_j/Z_0; 1 \leq j \leq k\}$, with $\{Z_0, \ldots, Z_k\}$ as before, having *pdf* as listed in Table 2. The scaled $\{V_j = \nu_0 Z_j/\nu_j Z_0; 1 \leq j \leq k\}$ then have a *multivariate F distribution* whenever $\{\alpha_j = \nu_j/2; 0 \leq j \leq k\}$ with $\{\nu_0, \ldots, \nu_k\}$ as positive integers. This arises in the ▶analysis of variance in conjunction with ratios of independent mean squares to a common denominator (Finney 1941). As before, neither χ^2 nor independence is required in the latter; take $\left\{V_j = \nu_0 \boldsymbol{y}_j'\boldsymbol{y}_j/\nu_j \boldsymbol{y}_0'\boldsymbol{y}_0; 1 \leq j \leq k\right\}$ with $\mathcal{L}(\boldsymbol{y}) \in \{\mathcal{S}_n(\boldsymbol{0}, \boldsymbol{I}_n, \psi); \psi \in \Psi\}$ as for Dirichlet distributions.

Matric distributions. Take $\{\boldsymbol{S}_0, \ldots, \boldsymbol{S}_k\}$ in $\mathbb{S}_k^+$ as independent Wishart matrices with $\{\mathcal{L}(\boldsymbol{S}_j) = W_k(\nu_j, \boldsymbol{\Sigma}); \nu_j \geq k; 0 \leq j \leq k\}$. Let $\boldsymbol{T} = \boldsymbol{S}_0 + \cdots + \boldsymbol{S}_k$ and $\left\{\boldsymbol{W}_j = \boldsymbol{T}^{-\frac{1}{2}}\boldsymbol{S}_j\boldsymbol{T}^{-\frac{1}{2}}; 1 \leq j \leq k\right\}$. A matric Dirichlet distribution (Olkin and Rubin 1964), taking the lower triangular square root, has *pdf* as listed in Table 3, such that $\boldsymbol{W}_j$ and $\left(\boldsymbol{I}_k - \Sigma_1^k \boldsymbol{W}_j\right)$ are positive definite, and $\nu_T = \nu_0 + \cdots + \nu_k$. Neither independence nor the Wishart character is required. If instead $\boldsymbol{Y} = \left[\boldsymbol{Y}_0', \boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_k'\right] \in \mathbb{F}_{n \times k}$, $n = \nu_0 + \cdots + \nu_k$, $\nu_j \geq k$, and $\left\{\boldsymbol{S}_j = \boldsymbol{Y}_j'\boldsymbol{Y}_j; j = 0, 1, \ldots, k\right\}$, then for $\mathcal{L}(\boldsymbol{Y}) = S_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$, invariance properties assure that $f(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_k)$ is identical to that given in Table 3, for every distribution $\mathcal{L}(\boldsymbol{Y})$ in $\{S_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \psi); \psi \in \Psi\}$.

An *inverted* matric Dirichlet distribution (Olkin and Rubin 1964) takes $\{\boldsymbol{S}_0, \boldsymbol{S}_1, \ldots, \boldsymbol{S}_k\}$ as before, and defines $\left\{\boldsymbol{V}_j = \boldsymbol{S}_0^{-\frac{1}{2}}\boldsymbol{S}_j\boldsymbol{S}_0^{-\frac{1}{2}}; 1 \leq j \leq k\right\}$ using the symmetric root of $\boldsymbol{S}_0$. Its *pdf* $f(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_k)$ is known allowing $\boldsymbol{S}_0$ to be noncentral. The central *pdf* is given in Table 3. The special case $k = 1$ is sometimes called a *Type II multivariate beta distribution.* Again neither independence nor the Wishart character is required. To see this, again take $\left\{\boldsymbol{S}_j = \boldsymbol{Y}_j'\boldsymbol{Y}_j; 0 \leq j \leq k\right\}$ as for matric Dirichlet distributions, and conclude that $f(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_k)$, as in Table 3, is identical for every $\mathcal{L}(\boldsymbol{Y})$ in $\{S_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \psi); \psi \in \Psi\}$.

Distributions of Latent Roots

Topics in multivariate statistics, to include reduction by invariance, tests for hypotheses regarding dispersion parameters, and the study of energy levels in physical systems, all entail the latent roots of random matrices. Suppose that $\mathcal{L}(\boldsymbol{W}) = W_k(\nu, \boldsymbol{\Sigma})$, and consider the ordered roots $\{w_1 > \ldots > w_k > 0\}$ of $|\ \boldsymbol{W} - w\boldsymbol{\Sigma}\ | = 0$. Their joint *pdf* is listed in Table 2. On occasion ratios of these roots are required, including simultaneous inferences for dispersion parameters, for which invariance in distribution holds. For if $\boldsymbol{W} = \boldsymbol{Y}'\boldsymbol{Y}$, then the joint distributions of ratios of the roots of $|\ \boldsymbol{W} - w\boldsymbol{\Sigma}\ | = 0$ are identical for all $\mathcal{L}(\boldsymbol{Y}) \in \{S_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \psi); \psi \in \Psi\}$ such that $n \geq k$.

To continue, consider $\boldsymbol{S}_0$ and $\boldsymbol{S}_1$ as independent Wishart matrices having $W_k(\nu_0, \boldsymbol{\Sigma})$ and $W_k(\nu_1, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, respectively. Then central ($\boldsymbol{\Lambda} = \boldsymbol{0}$) and noncentral joint distributions of the roots of $|\ \boldsymbol{S}_1 - \ell\boldsymbol{S}_0\ | = 0$ are known, as given in Table 2 for the case $\boldsymbol{\Lambda} = \boldsymbol{0}$. An invariance result holds for the central case. For if $\boldsymbol{Y} = \left[\boldsymbol{Y}_0', \boldsymbol{Y}_1'\right]'$ with $n = \nu_0 + \nu_1$ such that $\nu_0 \geq k$ and $\nu_1 \geq k$, $\boldsymbol{S}_0 = \boldsymbol{Y}_0'\boldsymbol{Y}_0$ and $\boldsymbol{S}_1 = \boldsymbol{Y}_1'\boldsymbol{Y}_1$, then by invariance the latent root *pdf* $f(\ell_1, \ldots, \ell_k)$ is the same for all $\mathcal{L}(\boldsymbol{Y})$ in $\{L_{n,k}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma}, \phi) : \phi \in \Phi\}$, as given in Table 3.

Multivariate Statistical Distributions. Table 4 Some discrete multivariate compound distributions

Basic distribution	Mixing parameters	Compounding distribution	Source	Resulting distribution
Bivariate binomial $(n, \pi_{01}, \pi_{10}, \pi_{11})$	n	Poisson	Papageorgiou (1983)	bivariate Poisson
Multinomial $(n, \pi_1, \ldots, \pi_s)$	$(\pi_1, \ldots, \pi_s)$	Dirichlet	Johnson et al. (1997) and Patil and Joshi (1968)	s–variate negative hypergeometric
Multinomial $(n, \pi_1, \ldots, \pi_s)$	n	Logarithmic series	Patil and Joshi (1968)	s–variate modified logarithmic series
Multinomial $(n, \pi_1, \ldots, \pi_s)$	n	Negative binomial	Patil and Joshi (1968)	s–variate negative multinomial
Multinomial $(n, \pi_1, \ldots, \pi_s)$	n	Poisson	Patil and Joshi (1968)	multiple Poisson
Multiple Poisson $(u\lambda_1, \ldots, u\lambda_s)$	u	Gamma	Patil and Joshi (1968)	s–variate negative multinomial
Multiple Poisson $(\lambda_1, \ldots, \lambda_s)$	$(\lambda_1, \ldots, \lambda_s)$	Multinormal	Steyn (1976)	s–variate Poisson–normal
Multiple Poisson $\{\lambda_i = \alpha + (\beta - \alpha)u$	u	Rectangular on (0,1)	Patil and Joshi (1968)	s–variate Poisson–rectangular
Multivariate Poisson $(u\lambda_1, u\lambda_{12}, \ldots, u\lambda_{12\cdot s})$	u	Gamma	Patil and Joshi (1968)	s–variate negative binomial
Negative multinomial $(k, \pi_1, \ldots, \pi_s)$	$(\pi_1, \ldots, \pi_s)$	Dirichlet	Johnson et al. (1997) Patil and Joshi (1968)	s–variate negative multinomial-Dirichlet
Convolution of multinomials $(\gamma_1, \ldots, \gamma_{2^k}, \theta_1, \ldots, \theta_s)$	$(\gamma_1, \ldots, \gamma_{2^k})$	Multivariate hypergeometric	Kotz and Johnson (1983)	numbers judged defective of k types in lot inspection

Other Distributions

Numerous other continuous multivariate distributions are known; a compendium is offered in Kotz et al. (2000). Multivariate versions of *Burr distributions* arise through gamma mixtures of independent Weibull distributions. Various *multivariate exponential distributions* are known; some properties and examples are found on specializing multivariate Weibull distributions. Various *multivariate stable distributions*, symmetric and asymmetric, are characterized through the structure of their *chf*s, as are types of symmetric MDs surveyed earlier. *Multivariate extreme-value distributions* are treated in Kotz et al. (2000), with emphasis on the bivariate case. The *Beta-Stacy distributions* yield a *multivariate Weibull distribution* as a special case. *Multivariate Pareto distributions* have their origins in econometrics. *Multivariate logistic distributions* model binary data in the analysis of quantal responses. Properties of *chf*s support a bivariate distribution having normal and gamma marginals (Kibble 1941).

Discrete Distributions

A guided tour is given with special reference to Johnson et al. (1997) and Patil and Joshi (1968). Inequalities for selected multivariate discrete distributions are offered in Jogdeo and Patil (1975).

Binomial, Multinomial, and Related

The outcome of a random experiment is classified as having or not having each of s attributes $\{A_1, \ldots, A_s\}$. If $\{X_1, \ldots, X_s\}$ are the numbers having these attributes in n independent trials, then theirs is a *multivariate binomial distribution* with parameters

$$\{\pi_i = \Pr(A_i), \pi_{ij} = \Pr(A_iA_j), \ldots, \pi_{12\cdot s} = \Pr(A_1A_2 \ldots A_s); \iota \in [1, 2, \ldots s]; i \neq j \neq k \neq \ldots\}$$

where ι takes successive values $\{i, j, k, \ldots\}$. The ►*binomial* ►distribution $B(n, \pi)$ obtains at $s = 1$. For bivariate binomial distributions see Hamdan (1972), Hamdan and Al-Bayyati (1971), and Hamdan and Jensen (1976). The limit as $n \to \infty$ and $\pi \to 0$ such that $n\pi \to \lambda$ is Poisson, the distribution of "rare events". More generally, as $n \to \infty$ and $\pi_i \to 0$, such that $\{n\theta_i \to \lambda_i, n\theta_{ij} \to \lambda_{ij}, \ldots, n\pi_{12\cdot s} \to \lambda_{12\cdot s}\}$, where $\{\theta_i, \theta_{ij}, \ldots\}$ are specified functions of $\{\pi_i, \pi_{ij}, \ldots\}$, then the limit of the multivariate binomial distribution is *multivariate Poisson*.

Suppose that independent trials are continued until exactly k trials exhibit none of the s attributes. The joint distribution of the numbers $\{Y_1, \ldots, Y_s\}$ of occurrences of $\{A_1, \ldots, A_s\}$ during these trials is a *multivariate Pascal distribution*.

To continue, let $\{A_0, \ldots, A_s\}$ be exclusive and exhaustive outcomes having probabilities $\{\pi_0, \ldots, \pi_s\}$, with $\{0 < \pi_i < 1; \pi_0 + \cdots + \pi_s = 1\}$. The numbers $\{X_1, \ldots, X_s\}$ of occurrences of $\{A_1, \ldots, A_s\}$ in n independent trials have the ►*multinomial distribution* with parameters $(n, \pi_1, \ldots, \pi_s)$. If independent trials are repeated until A_0 occurs exactly k times, the numbers of occurrences of $\{A_1, \ldots, A_s\}$ during these trials have a *negative multinomial distribution* with parameters $(k, \pi_1, \ldots, \pi_s)$.

In a multiway contingency table an outcome is classified according each of k criteria having the exclusive and exhaustive classes $\{A_{io}, A_{i1}, \ldots A_{is_i}; i = 1, \ldots, k\}$. If in n independent trials $\{X_{i1}, \ldots, X_{is_i}; i = 1, \ldots, k\}$ are the numbers occurring in $\{A_{i1}, \ldots, A_{is_i}; i = 1, \ldots, k\}$, then their joint distribution is called a *multivariate multinomial distribution* (also multivector multinomial). These are the joint distributions of marginal sums of the contingency table, to include the k–variate binomial distribution when $\{s_1 = s_2 = \cdots = s_k = 1\}$.

Hypergeometric and Related

A collection of N items consists of $s + 1$ types: N_0 of type A_0, N_1 of type A_1, …, N_s of type A_s, with $N = N_0 + \cdots + N_s$. Random samples are taken from this collection. If n items are drawn without replacement, the joint distribution of the numbers of items of types $\{A_1, \ldots, A_s\}$ is a *multivariate hypergeometric distribution* with parameters $(n, N, N_1, \ldots, N_s)$. With replacement, their distribution is multinomial with parameters $(n, N_1/N, \ldots, N_s/N)$.

If successive items are drawn without replacement until exactly k items of type A_0 are drawn, then the numbers of types $\{A_1, \ldots, A_s\}$ thus drawn have a *multivariate inverse hypergeometric distribution* with parameters $(k, N, N_1, \ldots, N_s)$.

To continue, sampling proceeds in two stages. First, m items are drawn without replacement, giving $\{x_1, \ldots, x_s\}$ items of types $\{A_1, \ldots, A_s\}$. Without replacing the first sample, n additional items are drawn without replacement at the second stage, giving $\{Y_1, \ldots, Y_s\}$ items of types $\{A_1, \ldots, A_s\}$. The conditional distribution of $(Y_1, \ldots, Y_s)$, given that $\{X_1 = x_1, \ldots, X_s = x_s\}$, is a *multivariate negative hypergeometric distribution.*

Multivariate Series Distributions

Further classes of discrete multivariate distributions are identified by types of their *pmf*s. Some arise through truncation and limits. If $[X_1, \ldots, X_s]$ has the s–variate negative multinomial distribution with parameters $(k, \pi_1, \ldots, \pi_s)$, then the conditional distribution of $[X_1, \ldots, X_s]$, given that $[X_1, \ldots, X_s] \neq [0, \ldots, 0]$, converges as $k \to 0$ to the s–variate *logarithmic series distribution* with parameters $(\theta_1, \ldots, \theta_s)$ where $\{\theta_i = 1 - \pi_i; i = 1, \ldots, s\}$. See Patil and Joshi (1968) for details. A modified multivariate logarithmic series distribution arises as a mixture, on n, of the multinomial distribution with parameters $(n, \pi_1, \ldots, \pi_s)$, where the mixing distribution is a logarithmic series distribution (Patil and Joshi 1968).

A class of distributions with parameters $(\theta_1, \ldots, \theta_s) \in \Theta$, derived from convergent power series, has *pmf*s of the form $p(x_1, \ldots, x_s) = \frac{a(x_1, \ldots, x_s)\theta_1^{x_1} \cdots \theta_s^{x_s}}{f(\theta_1, \ldots, \theta_s)}$ for $\{x_i = 0, 1, 2, \ldots; i = 1, \ldots, s\}$. The class of such distributions, called *multivariate power series distributions,* contains the s–variate multinomial distribution with parameters $(n, \pi_1, \ldots, \pi_s)$; the s-variate logarithmic series distribution with parameters $(\theta_1, \ldots, \theta_s)$; the s-variate negative multinomial distribution with parameters $(k, \pi_1, \ldots, \pi_s)$; and others. See Patil and Joshi (1968) for further properties. Other discrete multivariate distributions are described next.

Other Distributions

A typical *Borel-Tanner* distribution refers to the number of customers served before a queue vanishes for the first time. If service in a single-server queue begins with r customers of type I and s of type II with different arrival rates and service needs for each type, then the joint distribution of the numbers served is the *bivariate Borel-Tanner* distribution as in Shenton and Consul (1973).

In practice *compound distributions* often arise from an experiment undertaken in a random environment; the compounding distribution then describes variation of parameters of the model over environments. Numerous bivariate and multivariate discrete distributions have been obtained through compounding, typically motivated by the structure of the problem at hand. Numerous examples are cataloged in references Johnson et al. (1997) and Patil

and Joshi (1968); examples are listed in Table 4 from those and other sources.

About the Author

Donald Jensen received his Ph.D. from Iowa State University in 1962, and joined Virginia Polytechnic Institute and State University in 1965, attaining the rank of Professor in 1973. He has published over 140 journal articles in distribution theory, multivariate analysis, linear inference, robustness, outlier detection and influence diagnostics, regression design, and quality control. Dr. Jensen served as Associate editor of *The American Statistician* for a decade (1971–1980), and has been a reviewer for Mathematical Reviews for the last 30 years. He is an elected member of the International Statistical Institute. Professor Jensen received an early five-year Research Career Development Award from the US National Institutes of Health.

Cross References

►Binomial Distribution
►Bivariate Distributions
►Gamma Distribution
►Hypergeometric Distribution and Its Application in Statistics
►Multinomial Distribution
►Multivariate Normal Distributions
►Multivariate Statistical Analysis
►Multivariate Statistical Simulation
►Multivariate Technique: Robustness
►Poisson Distribution and Its Application in Statistics
►Statistical Distributions: An Overview
►Student's *t*-Distribution
►Weibull Distribution

References and Further Reading

Adrian R (1808) Research concerning the probabilities of errors which happen in making observations, etc. Analyst Math 1: 93–109
Arnold BC, Beaver RJ (2000) Some skewed multivariate distributions. Am J Math Manage Sci 20:27–38
Bedrick EJ, Lapidus J, Powell JF (2000) Estimating the Mahalanobis distance from mixed continuous and discrete data. Biometrics 56:394–401
Bhattacharya RN, Ranga Rao R (1976) Normal apprximations and asymptotic expansions. Wiley, New York
Birnbaum ZW (1948) On random variables with comparable peakedness. Ann Math Stat 19:76–81
Bravais A (1846) Analyse mathématique sur les probabilites des erreurs de situation d'un point. Mémoires Présentés par Divers Savants a l'Académie Royale des Sciences de l'Institut de France, Paris 9:255–332
Cambanis S, Huang S, Simons G (1981) On the theory of elliptically contoured distributions. J Multivariate Anal 11:368–385
Chmielewski MA (1981) Elliptically symmetric distributions: a review and bibliography. Int Stat Rev 49:67–74 (Excellent survey article on elliptical distributions)
Dawid AP (1977) Spherical matrix distributions and a multivariate model. J Roy Stat Soc B 39:254–261 (Technical source paper on the strucure of distributions)
Dempster AP (1969) Elements of continuous multivariate analysis. Addison-Wesley, London (General reference featuring a geometric approach)
Devlin SJ, Gnanadesikan R, Kettenring JR (1976) Some multivariate applications of elliptical distributions. In: Ikeda S et al (eds) Essays in probability and statistics. Shinko Tsusho, Tokyo, pp 365–394 (Excellent survey article on ellipsoidal distributions)
Dharmadhikari S, Joag-Dev K (1988) Unimodality, convexity, and applications. Academic, New York
Dickey JM (1967) Matrix variate generalizations of the multivariate *t* distribution and the inverted multivariate t distribution. Ann Math Stat 38:511–518 (Source paper on matric *t* distributions and their applications)
Dickson IDH (1886) Appendix to "Family likeness in stature" by F. Galton. Proc Roy Soc Lond 40:63–73
Edgeworth FY (1892) Correlated averages. Philos Mag 5 34:190–204
Epstein B (1948) Some applications of the Mellin transform in statistics. Ann Math Stat 19:370–379
Everitt BS, Hand DJ (1981) Finite mixture distributions. Chapman & Hall, New York
Fang KT, Anderson TW (eds) (1990) Statistical inference in elliptically contoured and related distributions. Allerton, New York
Fang KT, Kotz S, Ng KW (1990) Symmetric multivariate and related distributions. Chapman & Hall, London
Fang KT, Zhang YT (1990) Generalized multivariate analysis. Springer, New York
Fefferman C, Jodeit M, Perlman MD (1972) A spherical surface measure inequality for convex sets. Proc Am Math Soc 33: 114–119
Finney DJ (1941) The joint distribution of variance ratios based on a common error mean square. Ann Eugenic 11:136–140 (Source paper on dependent *F* ratios in the analysis of variance)
Galton F (1889) Natural inheritance. MacMillan, London, pp 134–145
Gauss CF (1823) Theoria Combinationis Observationum Erroribus Minimis Obnoxiae. Muster-Schmidt, Göttingen
Hamdan MA (1972) Canonical expansion of the bivariate binomial distribution with unequal marginal indices. Int Stat Rev 40: 277–280 (Source paper on bivariate binomial distributions)
Hamdan MA, Al-Bayyati HA (1971) Canonical expansion of the compound correlated bivariate Poisson distribution. J Am Stat Assoc 66:390–393 (Source paper on a compound bivariate Poisson distribution)
Hamdan MA, Jensen DR (1976) A bivariate binomial distribution and some applications. Aust J Stat 18:163–169 (Source paper on bivariate binomial distributions)
Helmert FR (1868) Studien über rationelle Vermessungen, im Gebeite der höheren Geodäsie. Zeitschrift für Mathematik und Physik 13:73–129
Hsu PL (1940) An algebraic derivation of the distribution of rectangular coordinates. Proc Edinburgh Math Soc 2 6:185–189 (Source paper on generalizations of Wishart's distribution)
James AT (1954) Normal multivariate analysis and the orthogonal group. Ann Math Stat 25:40–75
Jensen DR (1969) Limit properties of noncentral multivariate Rayleigh and chi-square distributions. SIAM J Appl Math

17:807–814 (Source paper on limits of certain noncentral distributions)
Jensen DR (1970a) A generalization of the multivariate Rayleigh distribution. Sankhya A 32:192–208 (Source paper on generalizations of Rayleigh distributions)
Jensen DR (1970b) The joint distribution of traces of Wishart matrices and some applications. Ann Math Stat 41:133–145 (Source paper on multivariate chi–squared and F distributions)
Jensen DR (l972) The limiting form of the noncentral Wishart distribution. Aust J Stat 14:10–16 (Source paper on limits of noncentral Wishart distributions)
Jensen DR (1976) Gaussian approximation to bivariate Rayleigh distributions. J Stat Comput Sim 4:259–268 (Source paper on normalizing bivariate transformations)
Jensen DR (1979) Linear models without moments. Biometrika 66:611–617 (Source paper on linear models under symmetric errors)
Jensen DR (1984) Ordering ellipsoidal measures: scale and peakedness orderings. SIAM J Appl Math 44:1226–1231
Jensen DR, Good IJ (1981) Invariant distributions associated with matrix laws under structural symmetry. J Roy Stat Soc B 43: 327–332 (Source paper on invariance of derived distributions under symmetry)
Jensen DR, Solomon H (1994) Approximations to joint distributions of definite quadratic forms. J Am Stat Assoc 89:480–486
Joe H (1997) Multivariate models and dependence concepts. Chapman & Hall/CRC, Boca Raton
Jogdeo K, Patil GP (1975) Probability inequalities for certain multivariate discrete distributions. Sankhya B 37:158–164 (Source paper on probability inequalities for discrete multivariate distributions)
Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Wiley, New York (An excellent primary source wlth extensive bibliography)
Kagan AM, Linnik YV, Rao CR (1973) Characterization problems in mathematical statistics. Wiley, New York
Kariya T, Sinha BK (1989) Robustness of statistical tests. Academic, New York
Kibble WF (1941) A two-variate gamma type distribution. Sankhya 5:137–150 (Source paper on expansions of bivariate distributions)
Kotz S, Balakrishnan N, Johnson NL (2000) Continuous multivariate distributions, 2nd edn. Wiley, New York (An excellent primary source with extensive bibliography)
Kotz S, Johnson NL (1983) Some distributions arising from faulty inspection with multitype defectives, and an application to grading. Commun Stat A Theo Meth 12:2809–2821
Kotz S, Nadarajah S (2004) Multivariate t distributions and their applications. Cambridge University Press, Cambridge
Laplace PS (1811) Memoir sur les integrales definies et leur application aux probabilites. Memoires de la classes des Sciences Mathématiques et Physiques l'Institut Impérial de France Année 1810:279–347
Lindsay BG (1995) Mixture models: theory, geometry and applications. NSF–CBMS regional conference series in probability and statistics, vol 5. Institute of Mathematical Statistics, Hayward
Lukacs E, Laha RG (1964) Applications of characteristic functions. Hafner, New York (Excellent reference with emphasis on multivariate distributions)
McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York
Miller KS (1975) Multivariate distributions. Krieger, Huntington (An excellent reference with emphasis on problems in engineering and communications theory)
Nelsen R (1998) An introduction to copulas. Springer, New York
Olkin I, Rubin H (1964) Multivariate beta distributions and independence properties of the Wishart distribution. Ann Math Stat 35:261–269; Correction, 37:297 (Source paper on matric Dirichlet, beta, inverted beta, and related distributions)
Olkin I, Tate RF (1961) Multivariate correlation models with mixed discrete and continuous variables. Ann Math Stat 32:448–465; Correction 36:343–344
Papageorgiou H (1983) On characterizing some bivariate discrete distributions. Aust J Stat 25:136–144
Patil GP, Joshi SW (1968) A dictionary and bibliography of discrete distributions. Hafner, New York (An excellent primary source with extensive bibliography)
Pearson K (1896) Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Philos Trans Roy Soc Lond A 187:253–318
Plana GAA (1813) Mémoire sur divers problémes de probabilitié. Mémoires de l'Académie Impériale de Turin 20:355–408
Schols CM (1875) Over de theorie der fouten in de ruimte en in het platte vlak. Verh Nederland Akademie Wetensch 15:1–75
Shaked M, Shanthikumar JG (2007) Stochastic orders. Springer, New York
Shenton LR, Consul PC (1973) On bivariate Lagrange and Borel–Tanner distributions and their use in queueing theory. Sankhya A 35:229–236 (Source paper on bivariate Lagrange and Borel–Tanner distributions and their applications)
Sherman S (1904) A theorem on convex sets with applications. Ann Math Stat 25:763–766
Spearman C (1904) The proof and measurement of associaton between two things. Am J Psychol 15:72–101
Steyn HS (1976) On the multivariate Poisson normal distribution. J Am Stat Assoc 71:233–236 (Source paper on multivariate Poisson–normal distributions)
Student (1908) The probable error of a mean. Biometrika 6:1–25
Subrahmaniam K (1970) On some applications of Mellin transformations to statistics: dependent random variables. SIAM J Appl Math 19:658–662
Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York
Tong YL (1980) Probability inequalities in multivariate distributions. Academic, New York
Tong YL (1990) The multivariate normal distribution. Springer-Verlag, New York

Multivariate Statistical Process Control

Robert L. Mason[1], John C. Young[2]
[1]Southwest Research Institute, San Antonio, TX, USA
[2]Lake Charles, LA, USA

Statistical process control (SPC) includes the use of statistical techniques and tools, such as ▶control charts, to

monitor change in a process. These are typically applied separately to each process variable of interest. Statistical process control procedures help provide an answer to the question: "Is the process in control?" When an out-of-control event is identified as a signal in a control chart, procedures often are available for locating the specific process variables that are the cause of the problem.

In multivariate statistical process control (MVSPC), multivariate statistical control procedures are used to simultaneously monitor many process variables that are interrelated and form a correlated set that move together (see Mason and Young 2002). The relationships that exist between and among the variables of the multivariate process are used in developing the procedure. Assume that the observation vectors obtained from a process are independent random variables that can be described by a multivariate normal distribution (see ►Multivariate Normal Distributions) with a mean vector and a covariance matrix. Any change in the mean vector and/or the covariance matrix of this distribution is considered an out-of-control situation and should be detectible with an appropriate multivariate control chart.

Implementation of a multivariate control procedure is usually divided into two parts: Phase I and Phase II. Phase I includes the planning, development, and construction phase. In this phase, the practitioner studies the process in great detail. Preliminary data are collected under good operational conditions and examined for statistical control and other potential problems. The major problems include statistical ►outliers, variable collinearities, and autocorrelated observations, i.e., time-dependent observations. After statistical control of the preliminary data is established, the data is used as the process history and referred to as the historical data set (HDS). If the parameters of the process are unknown, parameter estimates of the mean vector and covariance matrix are obtained from the data of the HDS for use in monitoring the process.

Phase II is the monitoring stage. In this phase, new observations are examined in order to determine if the process has deviated from the in-control situation specified by the HDS. Note that, in MVSPC, deviations from the HDS can occur through a mean vector change, a covariance matrix change, or both a mean vector and covariance matrix change in the process. In certain situations a change in one parameter can also induce a change in the other parameter.

Process control is usually determined by examining a control statistic based on the observed value of an individual observation and/or a statistic related to a rational subgroup (i.e., sample) of the observations such as the sample mean. Easy monitoring is accomplished by charting the value of the multivariate control statistic on a univariate chart. Depending on the charted value of this statistic, one can determine if control is being maintained or if the process has moved to an out-of-control situation.

For detecting both large and small shifts in the mean vector, there are three popular multivariate control chart methods. An implicit assumption when using these charts is that the underlying population covariance matrix is constant over the time period of interest. Various forms of ►Hotelling's T^2 statistic are generally chosen when the detection of large mean shifts is of interest (e.g., see Mason and Young 2002). For detecting small shifts in the process mean, the multivariate exponential weighted moving average (MEWMA) statistic (e.g., see Lowry et al. 1992) or the multivariate cumulative sum (MCUSUM) statistic (e.g., Woodall and Ncube 1985) can be utilized. These statistics each have advantages and disadvantages, and they can be used together or separately.

All of the above procedures were developed under the assumption that the data are independent and follow a multivariate normal distribution. Autocorrelated data can present a serious problem for both the MCUSUM and MEWMA statistics, but seems to have lesser influence on the behavior of the T^2 statistic. A main reason for the influence of autocorrelation on the MEWMA and MCUSUM statistics is that both of them are dependent on a subset of past-observed observation vectors, whereas the T^2 statistic depends only on the present observation.

A related problem in MVSPC is monitoring shifts in the covariance matrix for a multivariate normal process when the mean vector is assumed to be stable. A useful review of procedures for monitoring multivariate process variability is contained in Yeh et al. (2006). The methods for detecting large shifts in the covariance matrix include charts based on the determinant of the sample covariance matrix (Djauhari 2005), while the methods for detecting small shifts include charts based on a likelihood-ratio EWMA statistic (Yeh et al. 2004) and on related EWMA-type statistics (Yeh et al. 2003). A recent charting method that is applicable in monitoring the change in covariance matrix for a multivariate normal process is based on a form of Wilks' ratio statistic (Wilks 1963). It consists of taking the ratio of the determinants of two estimators of the process covariance matrix (Mason et al. 2009). One estimator is obtained using the HDS and the other estimator is computed using an augmented data set consisting of the newest observed sample and the HDS. The Wilks' chart statistic is particularly helpful when the number of variables is large relative to the sample size.

Current attention in the MVSPC literature is focused on procedures that simultaneously monitor both the mean vector and the covariance matrix in a multivariate process (e.g., see Reynolds and Cho 2006 or Chen et al. 2005). These charts are based on EWMA procedures and can be very useful in detecting small-to-moderate changes in a process. Several papers also exist that present useful overviews of MVSPC (e.g., see Woodall and Montgomery 1999 and Bersimis et al. 2007). These papers are valuable for their insights on the subject and their extensive reference lists.

About the Authors

Dr. Robert L. Mason is an Institute Analyst at Southwest Research Institute in San Antonio, Texas. He was President of the American Statistical Association in 2003, Vice-President in 1992–1994, and a Member of its Board of Directors in 1987–1989. He is a Fellow of both the American Statistical Association and the American Society for Quality, and an Elected Member of the International Statistical Institute. He has been awarded the Founder's Award and the Don Owen Award from the American Statistical Association and the W.J. Youden Award (twice) from the American Society for Quality. He is on the Editorial Board of the *Journal of Quality Technology*, and is an Associate Editor of *Communications in Statistics*. He has published over 130 research papers and coauthored 6 textbooks including *Statistical Design and Analysis of Experiments with Applications to Engineering and Science* (Wiley, 1989; 2nd ed. 2003). He also is the coauthor (with John C. Young) of *Multivariate Statistical Process Control with Industrial Applications* (ASA-SIAM; 2002).

Prior to his retirement in 2007, Dr. John C. Young was Professor of Statistics for 40 years at McNeese State University in Lake Charles, Louisiana. He has published approximately 100 papers in the statistical, medical, chemical, and environmental literature, and is coauthor of numerous book chapters and three textbooks.

Cross References

►Control Charts
►Hotelling's T^2 Statistic
►Multivariate Normal Distributions
►Outliers
►Statistical Quality Control
►Statistical Quality Control: Recent Advances

References and Further Reading

Bersimis S, Psarakis S, Panaretos J (2007) Multivariate statistical process control charts: an overview. Qual Reliab Eng Int 23:517–543

Chen G, Cheng SW, Xie H (2005) A new multivariate control chart for monitoring both location and dispersion. Commun Stat Simulat 34:203–218

Djauhari MA (2005) Improved monitoring of multivariate process variability. J Qual Technol 37:32–39

Lowry CA, Woodall WH, Champ CW, Rigdon SE (1992) A multivariate exponentially weighted moving average control chart. Technometrics 34:46–53

Mason RL, Young JC (2002) Multivariate statistical process control with industrial applications. ASA-SIAM, Philadelphia, PA

Mason RL, Chou YM, Young JC (2009) Monitoring variation in a multivariate process when the dimension is large relative to the sample size. Commun Stat Theory 38:939–951

Reynolds MR, Cho GY (2006) Multivariate control charts for monitoring the mean vector and covariance matrix. J Qual Technol 38:230–253

Wilks SS (1963) Multivariate statistical outliers. Sankhya A 25: 407–426

Woodall WH, Montgomery DC (1999) Research issues and ideas in statistical process control. J Qual Technol 31:376–386

Woodall WH, Ncube MM (1985) Multivariate CUSUM quality control procedures. Technometrics 27:285–292

Yeh AB, Lin DK, Zhou H, Venkataramani C (2003) A multivariate exponentially weighted moving average control chart for monitoring process variability. J Appl Stat 30: 507–536

Yeh AB, Huwang L, Wu YF (2004) A likelihood-ratio-based EWMA control chart for monitoring variability of multivariate normal processes. IIE Trans 36:865–879

Yeh AB, Lin DK, McGrath RN (2006) Multivariate control charts for monitoring covariance matrix: a review. Qual Technol Quant Manage 3:415–436

Multivariate Statistical Simulation

Mark E. Johnson
Professor
University of Central Florida, Orlando, FL, USA

Multivariate statistical simulation comprises the computer generation of multivariate probability distributions for use in statistical investigations. These investigations may be robustness studies, calibrations of small sample behavior of estimators or confidence intervals, power studies, or other Monte Carlo studies. The distributions to be generated may be continuous, discrete or a combination of both types. Assuming that the n-dimensional distributions have independent components, the problem of variate generation is reduced to simulating from univariate distributions for which, fortunately, there is a vast literature (Devroye 1986; L'Eculer 2010; and international standard ISO 28640, for

example). Thus, the real challenge of multivariate statistical simulation is in addressing the dependence structure of the multivariate distributions.

For a few situations, the dependence structure is readily accommodated from a generation standpoint. Consider the usual n-dimensional multivariate normal distribution (see ▶Multivariate Normal Distributions) with mean vector $\underline{\mu}$ and covariance matrix Σ. For a positive definite covariance matrix, there exists a lower triangular (Cholesky) decomposition $LL' = \Sigma$. Assuming a source of independent univariate normal variates to occupy the vector $\underline{X}$, the random vector $\underline{Y} = L\,\underline{X} + \underline{\mu}$ has the desired multivariate normal distribution. Having been able to generate multivariate normal random vectors, componentwise transformations provide the capability to generate the full Johnson translation system (1949a), of which the log-normal distribution may be the most familiar. In using the multivariate Johnson system, it is possible to specify the covariance matrix of the transformed distribution. Some researchers transform the multivariate normal distribution without noting the severe impact on the covariance matrix of the transformed distribution. This oversight makes it difficult to interpret the results of simulation studies involving the Johnson translation system (see Johnson 1987 for further elaboration).

In expanding to distributions beyond the Johnson translation system, it is natural to consider generalizations of the normal distribution at the core of this system. The exponential power distribution with density function $f(x)$ proportional to $\exp(-|x|^{\tau})$ is a natural starting point since it includes the double exponential distribution ($\tau = 1$), the normal distribution ($\tau = 2$) and the uniform distribution in the limit ($\tau \to \infty$) and is easy to simulate (Johnson 1979). A further generalization of the exponential power distribution amenable to variance reduction simulation designs was developed by Johnson, Beckman and Tietjen (1980) who noted that the normal distribution arises as the product of ZU where Z is distributed as the square root of a chi-squared(3) distribution and is independent of U which is uniform on the interval $(-1, 1)$. Their generalization occurs by considering arbitrary degrees of freedom and powers other than 0.5. Since by Khintchine's unimodality theorem, any unimodal distribution can be represented as such a product there are many possibilities that could be pursued for other constructions ultimately for use in multivariate simulation contexts.

Multivariate distribution *families* are appealing for simulation purposes. A useful extension of the Johnson translation system has been developed by Jones and Pewsey (2009). The family is defined implicitly via the equation

$$Z = \sinh[\delta \sinh^{-1}(X_{\varepsilon,\delta}) - \varepsilon]$$

where Z has the standard normal distribution, $X_{\delta,\varepsilon}$ has a sinh-arcsinh distribution, ε is a skewness parameter and δ relates to the tail weight of the distribution. This family of distributions is attractive for use in Monte Carlo studies, since it includes the normal distribution as a special intermediate (non-limiting) case and covers a variety of skewness and tailweight combinations. Extensions of the Jones-Pewsey family to the multivariate case can follow the approach originally taken by Johnson (1949b), with adaptations by Johnson et al. (1982) to better control impacts of the covariance structure and component distributions.

Variate generation for multivariate distributions is readily accomplished (at least, in principle) for a specific multivariate distribution provided certain conditional distributions are identified. Suppose $\underline{X}$ is a random vector to be generated. A direct algorithm is to first generate X_1 as the marginal distribution of the first component of $\underline{X}$, say x_1. Second, generate from the conditional distribution of X_2 given $X_1 = x_1$ to obtain x_2. Third, generate from the conditional distribution X_3 given, $X_1 = x_1$ and $X_2 = x_2$ and then continue until all n components have been generated. This conditional distribution approach converts the multivariate generation problem into a series of univariate generation problems. For cases in which the conditional distributions are very complicated or not particularly recognizable, there may be alternative formulae for generation, typically involving a transformation to $n+1$ or more independent random variables. Examples include a multivariate Cauchy distribution and the multivariate Burr-Pareto-logistic distributions (see Johnson 1987).

The general challenge in multivariate statistical simulation is to incorporate the dependence structure as it exists in a particular distribution. As noted earlier, the multivariate normal distribution is particularly convenient since dependence is introduced to independent normal components through appropriate linear transformations. Further transformations to the components of the multivariate normal distribution give rise to skewed, light tailed or heavy tailed marginal distributions while retaining some semblance of the dependence structure. An important approach to grappling with the dependence structure is to recognize that marginal distributions are not terribly relevant in that the components can be transformed to the uniform distribution via $U_i = F_i(X_i)$, where F_i is the distribution function of X_i. In other words, in comparing multivariate distributions, the focus can be on the transformed distribution having uniform marginal's. This multivariate distribution is known as a "copula." Examining the

►copulas associated with the Burr, Pareto and logistic distributions led Cook and Johnson to recognize the essential similarity of these three multivariate distributions. A very useful introduction to copulas is Nelsen (2006) while Genest and MacKay (1986) deserve credit for bringing copulas to the attention of the statistical community.

This entry does not cover all possible distributions or families of distributions that could be considered for use in multivariate simulation studies. Additional possibilities (most notably elliptically contoured distributions) are reviewed in Johnson (1987).

About the Author

For biography *see* the entry ►Statistical Aspects of Hurricane Modeling and Forecasting.

Cross References

►Copulas
►Monte Carlo Methods in Statistics
►Multivariate Normal Distributions
►Multivariate Statistical Distributions

References and Further Reading

Cook RD, Johnson ME (1981) A family of distributions for modelling non-elliptically symmetric multivariate data. Technometrics 28:123–131

Devroye L (1986) Non-uniform variate generation. Springer, New York. Available for free pdf download at http://cg.scs.carleton.ca/~luc/mbookindex.html

Genest C, MacKay RJ (1986) The joy of copulas: bivariate distributions with uniform marginals. Am Stat 40:280–283

International Standard 28640 (2010) Random variate generation methods. International Standards Organization (to appear), Geneva

Johnson ME (1987) Multivariate statistical simulation. Wiley, New York

Johnson ME (1979) Computer generation of the exponential power distribution. J Stat Comput Sim 9:239–240

Johnson ME, Beckman RJ, Tietjen GL (1980) A new family of probability distributions with applications to monte carlo studies. JASA 75:276–279

Johnson ME, Ramberg JS, Wang C (1982) The johnson translation system in monte carlo studies. Commun Stat Comput Sim 11:521–525

Johnson NL (1949a) Systems of frequency curves generated by methods of translation. Biometrika 36:149–176

Johnson NL (1949b) Bivariate distributions based on simple translation systems. Biometrika 36:297–304

Jones MC, Pewsey A (2009) Sinh-arcsinh distributions. Biometrika 96:761–780

L'Ecuyer P (2010) Non-uniform random variate generation. Encyclopedia of statistical science. Springer, New York

Nelsen RB (2006) An introduction to copulas, 2nd edn. Springer, New York

Multivariate Techniques: Robustness

Mia Hubert[1], Peter J. Rousseeuw[2]
[1]Associate Professor
Katholieke Universiteit Leuven, Leuven, Belgium
[2]Senior Researcher
Renaissance Technologies, New York, NY, USA

The usual multivariate analysis techniques include location and scatter estimation, ►principal component analysis, factor analysis (see ►Factor Analysis and Latent Variable Modelling), discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems), ►canonical correlation analysis, multiple regression and cluster analysis (see ►Cluster Analysis: An Introduction). These methods all try to describe and discover structure in the data, and thus rely on the correlation structure between the variables. Classical procedures typically assume normality (i.e. gaussianity) and consequently use the sample mean and sample covariance matrix to estimate the true underlying model parameters.

Below are three examples of multivariate settings used to analyze a data set with n objects and p variables, forming an $n \times p$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ the ith observation.

1. ►Hotelling's T^2 statistic for inference about the center of the (normal) underlying distribution is based on the sample mean $\bar{\boldsymbol{x}} = \sum_{i=1}^{n} \boldsymbol{x}_i$ and the sample covariance matrix $\boldsymbol{S}_x = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$.
2. Classical principal component analysis (PCA) uses the eigenvectors and eigenvalues of $\boldsymbol{S}_x$ to construct a smaller set of uncorrelated variables.
3. In the multiple regression setting, also a response variable $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is measured. The goal of linear regression is to estimate the parameter $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta})' = (\beta_0, \beta_1, \ldots, \beta_p)'$ relating the response variable and the predictor variables in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

The least squares slope estimator can be written as $\hat{\boldsymbol{\beta}}_{LS} = \boldsymbol{S}_x^{-1} \boldsymbol{s}_{xy}$ with $\boldsymbol{s}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ the cross-covariance vector. The intercept is given by $\hat{\beta}_0 = \bar{y} - \hat{\boldsymbol{\beta}}_{LS}' \bar{\boldsymbol{x}}$.

These classical estimators often possess optimal properties under the Gaussian model assumptions, but they can be strongly affected by even a few ►outliers. Outliers are data points that deviate from the pattern suggested by

the majority of the data. Outliers are more likely to occur in datasets with many observations and/or variables, and often they do not show up by simple visual inspection. When the data contain nasty outliers, typically two things happen:

- The multivariate estimates differ substantially from the "right" answer, defined here as the estimates we would have obtained without the outliers.
- The resulting fitted model does not allow to detect the outliers by means of their residuals, Mahalanobis distances, or the widely used "leave-one-out" diagnostics.

The first consequence is fairly well-known (although the size of the effect is often underestimated). Unfortunately the second consequence is less well-known, and when stated many people find it hard to believe or paradoxical. Common intuition says that outliers must "stick out" from the classical fitted model, and indeed some of them do so. But the most harmful types of outliers, especially if there are several of them, may affect the estimated model so much "in their direction" that they are now well-fitted by it.

Once this effect is understood, one sees that the following two problems are essentially equivalent:

- Robust estimation: find a "robust" fit, which is similar to the fit we would have found without the outliers.
- Outlier detection: find all the outliers that matter.

Indeed, a solution to the first problem allows us, as a by-product, to identify the outliers by their deviation from the robust fit. Conversely, a solution to the second problem would allow us to remove or downweight the outliers followed by a classical fit, which yields a robust estimate.

It turns out that the more fruitful approach is to solve the first problem and to use its result to answer the second. This is because from a combinatorial viewpoint it is more feasible to search for *sufficiently many* "good" data points than to find *all* the "bad" data points.

Many robust multivariate estimators have been constructed by replacing the empirical mean and covariance matrix with a robust alternative. Currently the most popular estimator for this purpose is the *Minimum Covariance Determinant* (MCD) estimator (Rousseeuw 1984). The MCD method looks for the h observations (out of n) whose classical covariance matrix has the lowest possible determinant. The raw MCD estimate of location is then the average of these h points, whereas the raw MCD estimate of scatter is a multiple of their covariance matrix. Based on these raw estimates one typically carries out a reweighting step, yielding the reweighted MCD estimates (Rousseeuw and Van Driessen 1999).

The MCD location and scatter estimates are affine equivariant, which means that they behave properly under affine transformations of the data. Computation of the MCD is non-trivial, but can be performed efficiently by means of the FAST-MCD algorithm (Rousseeuw and Van Driessen 1999) which is available in standard SAS, S-Plus, and R.

A useful measure of robustness is the *finite-sample breakdown value* (Donoho and Huber 1983; Hampel et al. 1986). The breakdown value is the smallest amount of contamination that can have an arbitrarily large effect on the estimator. The MCD estimates of multivariate location and scatter have breakdown value $\approx (n-h)/n$. The MCD has its highest possible breakdown value of 50% when $h = [(n+p+1)/2]$. Note that no affine equivariant estimator can have a breakdown value above 50%.

Another measure of robustness is the *influence function* (Hampel et al. 1986), which measures the effect on an estimator of adding a small mass of data in a specific place. The MCD has a bounded influence function, which means that a small contamination at any position can only have a small effect on the estimator (Croux and Haesbroeck 1999).

In regression, a popular estimator with high breakdown value is the *Least Trimmed Squares* (LTS) estimator (Rousseeuw 1984; Rousseeuw and Van Driessen 2006). The LTS is the fit that minimizes the sum of the h smallest squared residuals (out of n). Other frequently used robust estimators include S-estimators (Rousseeuw and Yohai 1984) and MM-estimators (Yohai 1987), which can achieve a higher finite-sample efficiency than the LTS.

Robust multivariate estimators have been used to robustify the Hotelling T^2 statistic (Willems et al. 2002), PCA (Croux and Haesbroeck 2000; Salibian-Barrera et al. 2006), multiple regression with one or several response variables (Rousseeuw et al. 2004; Agulló et al. 2008), discriminant analysis (Hawkins and McLachlan 1997; Hubert and Van Driessen 2004; Croux and Dehon 2001), factor analysis (Pison et al. 2003), canonical correlation (Croux and Dehon 2002), and cluster analysis (Hardin and Rocke 2004).

Another important group of robust multivariate methods are based on projection pursuit (PP) techniques. They are especially useful when the dimension p of the data is larger than the sample size n, in which case the MCD is no longer well-defined. Robust PP methods project the data on many univariate directions and apply robust estimators of location and scale (such as the median and the median absolute deviation) to each projection. Examples include the Stahel-Donoho estimator of location and scatter (Maronna and Yohai 1995) and generalizations (Zuo et al. 2004), robust

PCA (Li and Chen 1985; Croux and Ruiz-Gazen 2005; Hubert et al. 2002; Boente et al. 2006), discriminant analysis (Pires 2003), canonical correlation (Branco et al. 2005), and outlier detection in skewed data (Brys et al. 2005; Hubert and Van der Veeken 2008). The hybrid ROBPCA method (Hubert et al. 2005; Debruyne and Hubert 2009) combines PP techniques with the MCD and has led to the construction of robust principal component regression (Hubert and Verboven 2003), partial least squares (Hubert and Vanden Branden 2003), and classification for high-dimensional data (Vanden Branden and Hubert 2005).

A more extensive description of robust multivariate methods and their applications can be found in (Hubert et al. 2008; Hubert and Debruyne 2010).

About the Author

Dr. Peter Rousseeuw was Professor and Head (since 1992) of the Division of Applied Mathematics, Universiteit Antwerpen, Belgium. Currently he is a Senior Researcher at Renaissance Technologies in New York. He has (co-)authored over 160 papers, two edited volumes and three books, including *Robust Regression and Outlier Detection* (with A.M. Leroy, Wiley-Interscience, 1987). In 2003 ISI-Thompson included him in their list of Highly Cited Mathematicians. His paper *Least Median of Squares Regression* (1984), Journal of the American Statistical Association, 79, 871–880) which proposed new robust methods for regression and covariance, has been reprinted in *Breakthroughs in Statistics III* (the three-volume collection consists of the 60 most influential publications in statistics from 1850 to 1990), Kotz and Johnson 1997, Springer-Verlag, New York. He is an Elected Member, International Statistical Institute (1991) and an Elected Fellow of Institute of Mathematical Statistics (elected 1993) and American Statistical Association (elected 1994). He was Associate Editor, *Journal of the American Statistical Association* (1988–1993), and *Computational Statistics and Data Analysis* (1988–1998). He has supervised 20 Ph.D. students.

Cross References

►Eigenvalue, Eigenvector and Eigenspace
►Functional Derivatives in Statistics: Asymptotics and Robustness
►Hotelling's T^2 Statistic
►Multivariate Data Analysis: An Overview
►Multivariate Outliers
►Multivariate Statistical Analysis
►Outliers
►Principal Component Analysis
►Robust Inference
►Robust Statistics

References and Further Reading

Agulló J, Croux C, Van Aelst S (2008) The multivariate least trimmed squares estimator. J Multivariate Anal 99:311–318
Boente G, Pires AM, Rodrigues I (2006) General projection-pursuit estimates for the common principal components model: Influence functions and Monte Carlo study. J Multivariate Anal 97:124–147
Branco JA, Croux C, Filzmoser P, Oliviera MR (2005) Robust canonical correlations: a comparative study. Comput Stat 20:203–229
Brys G, Hubert M, Rousseeuw PJ (2005) A robustification of independent component analysis. J Chemometr 19:364–375
Croux C, Dehon C (2001) Robust linear discriminant analysis using *S*-estimators. Can J Stat 29:473–492
Croux C, Dehon C (2002) Analyse canonique basée sur des estimateurs robustes de la matrice de covariance. La Revue de Statistique Appliquée 2:5–26
Croux C, Haesbroeck G (1999) Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J Multivariate Anal 71:161–190
Croux C, Haesbroeck G (2000) Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika 87:603–618
Croux C, Ruiz-Gazen A (2005) High breakdown estimators for principal components: the projection-pursuit approach revisited. J Multivariate Anal 95:206–226
Debruyne M, Hubert M (2009) The influence function of the Stahel-Donoho covariance estimator of smallest outlyingness. Stat Probab Lett 79:275–282
Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel P, Doksum K, Hodges JL (eds) A Festschrift for Erich Lehmann. Wadsworth, Belmont, pp 157–184
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley-Interscience, New York
Hardin J, Rocke DM (2004) Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Comput Stat Data Anal 44:625–638
Hawkins DM, McLachlan GJ (1997) High-breakdown linear discriminant analysis. J Am Stat Assoc 92:136–143
Hubert M, Debruyne M (2010) Minimum covariance determinant. Wiley Interdisciplinary Rev Comput Stat 2:36–43
Hubert M, Van der Veeken S (2008) Outlier detection for skewed data. J Chemometr 22:235–246
Hubert M, Van Driessen K (2004) Fast and robust discriminant analysis. Comput Stat Data Anal 45:301–320
Hubert M, Vanden Branden K (2003) Robust methods for partial least squares regression. J Chemometr 17:537–549
Hubert M, Verboven S (2003) A robust PCR method for high-dimensional regressors. J Chemometr 17:438–452
Hubert M, Rousseeuw PJ, Verboven S (2002) A fast robust method for principal components with applications to chemometrics. Chemomet Intell Lab 60:101–111
Hubert M, Rousseeuw PJ, Vanden Branden K (2005) ROBPCA: a new approach to robust principal components analysis. Technometrics 47:64–79
Hubert M, Rousseeuw PJ, Van Aelst S (2008) High breakdown robust multivariate methods. Stat Sci 23:92–119

Li G, Chen Z (1985) Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. J Am Stat Assoc 80:759–766

Maronna RA, Yohai VJ (1995) The behavior of the Stahel-Donoho robust multivariate estimator. J Am Stat Assoc 90: 330–341

Pires AM (2003) Robust discriminant analysis and the projection pursuit approach: practical aspects. In: Dutter R, Filzmoser P, Gather U, Rousseeuw PJ (eds) Developments in robust statistics. Physika Verlag, Heidelberg, pp 317–329

Pison G, Rousseeuw PJ, Filzmoser P, Croux C (2003) Robust factor analysis. J Multivariate Anal 84:145–172

Rousseeuw PJ, Yohai V (1984) Robust regression based on S-estimators. In: Franke J, Haerdle W, Martin RD (eds) Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics No. 26, Springer Verlag, New York, pp 256–272

Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880

Rousseeuw PJ, Yohai AM (1987) Robust regression and outlier detection. Wiley-Interscience, New York

Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223

Rousseeuw PJ, Van Driessen K (2006) Computing LTS regression for large data sets. Data Min Knowl Disc 12:29–45

Rousseeuw PJ, Van Aelst S, Van Driessen K, Agulló J (2004) Robust multivariate regression. Technometrics 46:293–305

Salibian-Barrera M, Van Aelst S, Willems G (2006) PCA based on multivariate MM-estimators with fast and robust bootstrap. J Am Stat Assoc 101:1198–1211

Vanden Branden K, Hubert M (2005) Robust classification in high dimensions based on the SIMCA method. Chemometr Intell Lab 79:10–21

Willems G, Pison G, Rousseeuw PJ, Van Aelst S (2002) A robust Hotelling test. Metrika 55:125–138

Yohai VJ (1987) High breakdown point and high efficiency robust estimates for regression. Ann Stat 15:642–656

Zuo Y, Cui H, He X (2004) On the Stahel-Donoho estimator and depth-weighted means of multivariate data. Annals Stat 32: 167–188

N

National Account Statistics

Barbara M. Fraumeni
Associate Executive Director, Chair of the Ph.D. Program in Public Policy, and Professor of Public Policy
University of Southern Maine, Portland, ME, USA

National accounts are the system by which the current level of economic activity is measured, the summary statistic being Gross Domestic Product (GDP). Government officials, policy makers, businessmen, and investors often react to GDP real growth rates and levels as GDP is a primary indicator of the health of the economy. Safeguards are taken in the United States, and probably in many other countries, to ensure that no one has advance knowledge of these numbers before the official release to protect against someone capitalizing from inside knowledge.

National accounts are the economist's version of business accounts. Both national and business accountants follow specific rules and methodologies. The most common set of rules for national accounts is summarized for national accounts in the *System of National Accounts* (SNA); the current published version is dated 2008. (Many countries have not yet changed their national accounts to reflect changes between SNA 1993 and SNA 2008.) The SNA was developed by Sir Richard Stone and is maintained by a group of international organizations. The United States national accounts (National Income and Product Accounts or NIPAs) were developed by Simon Kuznets and use a different system than the SNA. However, the NIPAs are currently largely compatible with the SNA. National accounts and business accounts are similar in that they both use a double-entry system and have a number of account components which relate to a snapshot, i.e., an assets and liabilities balance sheet for a particular date, or a flow account, i.e., a production account or a profit or loss statement for the year. However, national accounts and business accounts use significantly different rules and methodologies.

GDP is a measure of the goods and services currently produced by labor and property located within a specific geographic region. GDP can be measured in three different ways using an expenditure, income or value-added approach. The expenditures approach estimates GDP as:

$$\text{GDP} = \text{Consumption} + \text{Investment} + \text{Government Expenditures} + \text{Exports} - \text{Imports},$$

where in this formula consumption and investment exclude consumption and investment by government, which are included in the Government Expenditures total. Imports are subtracted to ensure that only production within a specific geographic region are included in GDP. The income approach estimates GDP as the sum of income received and costs incurred in production. The value-added approach estimates GDP as:

$$\text{GDP} = \text{Total Sales} - \text{Total Intermediate Inputs}.$$

An intermediate input is a good or service purchased for resale, i.e., by a retailer from a wholesaler, or a good or service purchased for use in producing another good or service (hence the name intermediate "input"), i.e., flour used to produce a loaf of bread. If the price paid for a product by a retailer, i.e., carrots brought by a grocery store from a farm, were included in GDP as well as the price paid for the product by an individual consumer, i.e., carrots bought by an individual consumer from a grocery store, then there would be double-counting in GDP equal to the price paid by the retailer, i.e., the grocery store. At each stage of production normally some inputs are added, such as transportation, packaging, and labor costs. These costs as reflected in the final price to a consumer, investor, or government are included in GDP. Note that GDP, regardless of which approach is used, does not include capital gains as these do not result from current production.

There are three major categories of output in the SNA: Market, produced for your own use, and other non-market output. Market produced for your own use includes food produced and consumed on a farm. Other non-market output consists of goods and or services produced by non-profit institutions or governments that are supplied free, or at prices that are not economically significant. Although market output is generally the easiest to estimate, problems arise in the estimation of all three types of output.

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

In all cases, collection of data is critical to the estimation of GDP. Missing data points are estimated using interpolation or extrapolation with indicators, such as computer programmer for software production, supplementary data sources, and/or best judgment. Other components of GDP are imputed notably because there is no market information. The imputation for services from owner-occupied housing is normally the largest imputation. Rent payments for houses of similar sizes and quality in similar locations are often used to impute these services when possible. Services from intangibles, such as research and development (R & D), represent a particularly challenging measurement problem.

GDP is presented in the prices of the day, frequently called nominal or current dollars, or in units which allow for comparisons over time, called volume in SNA terminology, real in NIPA terminology, or by many others constant dollars or quantity. Creating volume or real GDP is one of the major challenges facing national accountants. Volume or real measures are created by holding prices constant, allowing for changes in the number and the quality of goods and services over time. Volume or real measures are estimated directly or by deflating nominal measures with prices indexes. Typically indexes are used such as the Paasche, Laspeyres, Fisher or Thornqvist index. Fixed price (volume or quantity) indexes commonly are a Paasche or Laspeyres index; chained prices (volume or quantity) indexes frequently are a Fisher or Thornqvist index. Quality changes in goods and services can be particularly difficult to measure, such as quality changes in computers. During the second half of the nineties in the United States, prices of computers were declining at the same time as performance and sales of computers was increasing. If price and volume (or quantity) indexes had not been adjusted for quality changes and chain indexes were not used, GDP would have been misestimated. Aggregate indexes, such as GDP, will change if the composition of goods and services produced changes even if there are no quality changes because indexes typically involve nominal dollar weights.

This short description describes the primary macroeconomic aggregate of national accounts: GDP, but there are many other economic statistics contained in national accounts. To give a sense of the large volume of information available in national accounts, note that there are nine accounts in the SNA and seven accounts in the NIPAs. The SNA accounts include those for production, distribution of income, redistribution of income, use of income, capital, financial, other changes in asset values, balance sheet, and goods and services. The NIPA accounts include those for domestic income and product, private enterprise income, personal income and outlay, government receipts and expenditures, foreign transactions current, domestic capital, and foreign transactions capital. For further information on national accounts, it is recommended that you refer to the references, particularly those by Lequiller and Blades, and Landefeld et al., and the two short documents by the Bureau of Economic Analysis.

About the Author

Dr. Barbara Fraumeni is Professor of Public Policy, Public Policy Ph.D. Chair, Associate Dean for Research and Associate Dean for Academic and Student Affairs, at Muskie School of Public Service, University of Southern Maine. She was Chief Economist of the U.S. Bureau of Economic Analysis (1999–2005). She was awarded the Gold Medal, highest award given by the U.S. Department of Commerce, for creating an R&D Satellite Account that treats R&D as investment and can be used to assess its contribution to economic growth and competitiveness, during her tenure as Chief Economist at the Bureau of Economic Analysis (November 8, 2006). She also received the 2006 American Economic Association's Committee on the Status of Women in the Economics Profession (CSWEP) Carolyn Shaw Bell Award for furthering the status of women in the economics professions through example, mentoring, and achievements. Currently, she is Chair, Committee on the Status of Women in the Economics Profession, American Economic Association (2008–2011). She also serves on the National Academy of Sciences/Committee on National Statistics Panel on Productivity in Higher Education, the United Nations Economic Commission for Europe/Organisation for Economic Cooperation and Development/Eurostat Task Force on Measuring Sustainable Development, as a Senior Fellow of the China Center for Human Capital and Labor Market Research of the Central University of Finance and Economics, and as a Research Fellow of the National Bureau of Economic Research. Her areas of expertise include measurement issues and national income accounting.

Cross References

►Business Surveys
►Comparability of Statistics
►Economic Growth and Well-Being: Statistical Perspective
►Economic Statistics
►Eurostat
►Federal Statistics in the United States, Some Challenges
►Promoting, Fostering and Development of Statistics in Developing Countries

References and Further Reading

Bureau of Economic Analysis, A Guide to the National Income and Product Accounts of the United States. Bureau of Economic Analysis, U.S. Department of Commerce, September 2006, at http://www.bea.gov/national/pdf/nipaguid.pdf

Bureau of Economic Analysis, Measuring the Economy: A Primer on GDP and the National Income and Product Accounts, prepared by Stephanie H. McCulla and Shelly Smith, Bureau of Economic Analysis, U.S. Department of Commerce, September 2007 at http://www.bea.gov/national/pdf/nipa_primer.pdf

Commission of the European Communities-Eurostat, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and World Bank, *System of National Accounts 1993*, Commission of the European Communities-Eurostat, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and World Bank, Brussels/Luxembourg, New York, Paris, Washington, 1993 at http://unstats.un.org/unsd/sna1993/toctop.asp

Commission of the European Communities-Eurostat, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and World Bank, *System of National Accounts 1993*, Commission of the European Communities-Eurostat, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and World Bank, Brussels/Luxembourg, New York, Paris, Washington, 2008 at http://unstats.un.org/unsd/nationalaccount/SNA2008.pdf

Landefeld JS, Seskin EP, Fraumeni BM (2008) Taking the pulse of the economy: measuring GDP. J Econ Perspect 22(2):193–216

Lequiller, François and Derek Blades, Understanding National Accounts, Organization for Economic Co-operation and Development, 2006 at http://www.oecd.org/document/58/0,3343,en_2649_34245_38445370_1_1_1_1,00.html

Network Models in Probability and Statistics

Stephen E. Fienberg[1], Andrew C. Thomas[2]
[1]Professor
Carnegie Mellon University, Pittsburgh, PA, USA
[2]Visiting Assistant Professor
Carnegie Mellon University, Pittsburgh, PA, USA

A network is a representation for a collection of individuals or other units connected in a pairwise fashion by relationships, such as friendship. Networks are typically displayed visually as "graphs," so that individuals correspond to the "nodes" of a graph, with the existence of a relationship indicated by an edge between pairs of nodes. Relationships can be univariate or multivariate, and the connections between individuals can be either directed (from one to the other) or undirected. In terms of statistical science, a network model is one that accounts for the structure of the network ties in terms of the probability that each network tie exists, whether conditional on all other ties, or as considered part of the distribution of the ensemble of ties.

A network with N nodes has $\binom{N}{2}$ unordered pairs of nodes, and hence $2\binom{N}{2}$ possible directed edges. If the labels on edges reflect the nodes they link, as (i,j), Y_{ij} represents the existence of an edge from individual i to j, and $\{\mathbf{Y}\} = \{Y_{12}, Y_{13}, \ldots Y_{(N-1)N}\}$ represents the ties in the graph.

Conditionally Independent Edges and Dyads

For a binary graph with conditionally independent edges, each edge outcome Y_{ij} can be expressed as a Bernoulli binary random variable with probability of existence p_{ij}. The simplest of this class of network models is the Erdős–Rényi–Gilbert random graph model (Erdős and Rényi 1959, 1960; Gilbert 1959) (sometimes referred to as the Erdős-Rényi model, or "the" random graph), in which any given edge exists with probability p. This model extends immediately to directed graphs, where any arc has the same existence probability p as any other. There is a large, and still growing, probabilistic literature on random graph models and their generalizations, that is well summarized in Durrett (2006) and Chung and Lu (2006).

The Erdős–Rényi–Gilbert class of model assumes that there is no differentiation between nodes, and that the two arcs within a dyad are independent of each other. The p_1 model of Holland and Leinhardt (1981) proposes that three factors affect the outcome of a dyad: the "gregariousness" α of an individual, how likely they are to have outgoing ties; the "popularity" β of an individual; how likely they are to have incoming ties; and "reciprocity" ρ, the tendency to which the two arcs in a dyad are identical, taking into account their existing characteristics. Given a parameter for the overall density θ, the form of the joint likelihood is

$$P(X = x) = \exp\left(\rho m + \theta x_{++} + \sum_i \alpha_i x_{i+} + \sum_j \beta_j x_{+j}\right) K(\rho, \theta, \alpha, \beta). \quad (1)$$

where $K(\rho, \theta, \alpha, \beta)$ is a normalizing constant to insure that the total probabilities add to 1. Additionally, Holland and Leinhardt (1981) present an iterative proportional fitting method for maximum likelihood estimation for this model, and discuss the complexities involved in assessing goodness-of-fit.

A natural extension of the p_1 model is the case of tightly linked "blocks" of nodes, within which the α and β parameters are equated, suggesting an equivalence

between members of the same block. The inference for and discovery of "communities" in networks has become especially popular in recent network literature in a variety of different applications; see Newman (2004) for an example.

Ensemble Models and Topological Motivations

Rather than focusing on the dyads as independent units, there are classes of models that consider topological features of interest in the network as the main measure of the model. The best known of these is the Exponential Random Graph Model, or ERGM (Frank and Strauss 1986), also known as p^* (Anderson et al. 1999), which extends directly from the p_1 model, by adding statistical summaries of topological relevance. For example, the number of three-cycles or triangles in a graph is equal to

$$T = \sum_{i,j,k} X_{ij}X_{jk}X_{ki};$$

this can then be added into the likelihood of Eq. (1), as in

$$P(X = x) = \exp\left(\tau T + \rho m + \theta x_{++} + \sum_i \alpha_i x_{i+} + \sum_j \beta_j x_{+j}\right) K(\tau, \rho, \theta, \alpha, \beta), \quad (2)$$

where $K(\tau, \rho, \theta, \alpha, \beta)$ is a new normalizing constant. Due to the computational intractability of this normalizing constant, much of the recent literature on ERGMs uses Markov chain Monte Carlo methods (see ►Markov Chain Monte Carlo) for maximum likelihood estimation of the model parameters. Additionally, these models often have degenerate or near-degenerate solutions, as explained in Rinaldo et al. (2009).

Evolutionary Models and Methods

Two popular models for binary networks come from a network model based charcterized by stepwise changes. The first is the "small world" model Watts and Strogatz (1998), which suggests that many networks in nature exhibit high local clustering – the tendency of a node's neighbours to be connected, a sign of an "ordered" system – with short characteristic path lengths, typical of a more "chaotic" system. The second is the "scale-free" model of Barabási and Albert (1999), in which nodes enter sequentially into a existing system, making connections with previously added nodes. The probability of selecting a node for a new link is proportional to its degree at that time, so that "rich" nodes are more likely to receive new links, a notion that goes back almost a century to the work of Yule (1925). The recent literature on this class of evolutionary models emerged from ideas in statistical physics and has been subject to criticism for its loose assessment of the "scale-free" property Li et al. (2006).

Considerable current interest focuses on continuous-time dynamic models of networks, for example, that describe the changes in edge properties. For example, Wasserman (1980) described Markov models for edge states in which the probability of a friendship being made or broken during a particular time interval is proportional to properties of the individuals and of reciprocity in general. Snijders (2004) among others has extended these ideas.

Further Reading

For in-depth reviews of the probabilistic literature on random graph models, see Durrett (2006) and Chung and Lu (2006). Kolaczyk (2009) provides a detailed examination of a number of different network models and their applications, and Goldenberg et al. (2010) provides a broad statistical review with an extensive bibliography.

Acknowledgments

This research was supported in part by National Science Foundation grant DMS-0631589 and U.S. Army Research Office Contract W911NF-09-1-0360 to Carnegie Mellon University.

About the Author

For biography of Stephen E. Fienberg *see* the entry ►Data Privacy and Confidentiality.

Cross References

►Data Privacy and Confidentiality
►Exponential Family Models
►Markov Chain Monte Carlo
►Probabilistic Network Models
►Social Network Analysis

References and Further Reading

Anderson CJ, Wasserman S, Crouch B (1999) A p^* primer: logit models for social networks. Soc Netw 21:37–66

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Chung F, Lu L (2006) Complex graphs and networks. American Mathematical Society, Providence

Durrett R (2006) Random graph dynamics. Cambridge University Press, New York

Erdös P, Rényi A (1959) The evolution of random graphs. Magyar Tud Akad Mat Kutató Int Közl 5:17–61

Erdös P, Rényi A (1960) On random graphs. Publications Mathematicae 6:290

Frank O, Strauss D (1986) Markov graphs. J Am Stat Assoc 81:832–842

Gilbert EN (1959) Random graphs. Ann Math Stat 30:1141–1144

Goldenberg A, Zhang A, Fienberg SE, Airoldi E (2010) A survey of statistical network models. Foundations and trends in machine learning 2:129–233

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76:33–65 (with discussion)
Kolaczyk ED (2009) Statistical analysis of network data. Springer, New York
Li L, Alderson D, Doyle JC, Willinger W (2006) Towards a theory of scale-free graphs: definition, proper ties, and implications. Internet Math 2(4):431–521
Newman MEJ (2004) Detecting community structure in networks. Eur J Phy B 38:321–330
Rinaldo A, Fienberg SE, Zhou Y (2009) On the geometry of discrete exponential families with application to exponential random graph models. Electron J Stat 3:446–484
Snijders TAB (2005) Models for longitudinal network data. In: Carrington P, Scott J, Wasserman S (eds) Models and methods for social network analysis, Chap 11. Cambridge University Press, New York, pp 215–247
Wasserman S (1980) Analyzing social networks as stochastic processes. J Am Stat Assoc 75:280–294
Watts D, Strogatz S (1998) Collective dynamics of "small-world" networks. Nature 393:440–442
Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos Trans R Soc Lond B 213:21–87

Network Sampling

Ove Frank
Professor Emeritus
Stockholm University, Stockholm, Sweden

Network sampling refers to the observation of a sampled network from some population or family F of possible networks. In particular, F can be a family of subnets obtainable from a fixed graph or network G. In this case, G is usually referred to as the population graph or the population network.

A graph G with vertex set V and edge set E is identified by a binary function $y = \{(u, v, y_{uv}) : (u, v) \in V^2\}$ that for each ordered pair of vertices in V returns a value y_{uv} indicating whether or not $(u, v) \in E$. If $V = \{1, \ldots, N\}$ and the vertices are ordered according to their labeling, y can be displayed as the adjacency matrix (y_{uv}) of G. A more general network G, a valued graph with univariate or multivariate variables defined on the vertex pairs, can be represented by a univariate or multivariate function y on V^2.

A family F of subnets obtainable from a fixed population network specified by a function y can be generated by a vertex sample S selected according to a specified probabilistic sampling design. Let $E(S)$ be the subset of V^2 generated by S, and let $y(S)$ be the restriction of y to $E(S)$. The family F of observable subnets consists of the subnets represented by $y(S)$ for different outcomes of S. In particular, a subnet induced by S consists of y_{uv} for $(u, v) \in S^2$, and a subnet generated from and towards S consists of y_{uv} for all (u, v) with $u \in S$ or $v \in S$.

The subnets represented by $y(S)$ are random due to the randomness of the sampling design of S. Any inference on y based on $y(S)$ is referred to as design based inference. The population network represented by y is usually too complicated to describe in detail, and it might be conveniently summarized by some summary statistics or by some probabilistic model assumptions focusing on important features between the variables involved in y. If the probabilistic sampling design of S is combined with probabilistic model assumptions about y, the observed subnet $y(S)$ is random due to the randomness of both design and model. In this case, model based inference refers to inference on features of the model of y, while model assisted inference is sometimes used as a term for inference on the outcome of y.

As a simple example, consider an undirected population graph on $V = \{1, \ldots, N\}$ given by the adjacency matrix $y = (y_{uv})$ with $y_{vv} = 0$ and $y_{uv} = y_{vu}$ for $u \neq v$. The number of edges R and the number of isolated vertices N_0 should be estimated by using the information from a subgraph induced by a Bernoulli (p)-sample S of vertices. Here vertices in V are independently selected with a common probability p, which for instance could be chosen as $0.1 + N^{-1/2}$ in order to get a small probability for a sample that is too small. Let n_k be the number of vertices of degree k in the sample graph for $k = 0, 1, \ldots$ Then the number of edges in the sample graph is given by $r = (1/2)\Sigma_k k n_k$, and R can be estimated by r/p^2. It can be shown that N_0 has an unbiased estimator given by the alternating series $(1/p)\Sigma_k n_k(-q/p)^k$ where $q = 1 - p$. This estimator has very low precision, and it is desirable to replace it with a model based alternative. Assume that the population graph is a Bernoulli (α)-graph so that y_{uv} for $u < v$ are independent Bernoulli (α)-variables. It follows that R and N_0 have expected values $N(N-1)\alpha/2$ and $N(1-\alpha)^{N-1}$, respectively. Now α can be estimated by the edge density $2r/n(n-1)$ of the sample graph, and it follows that R and N_0 could be predicted to be close to their estimated expected values $N(N-1)r/n(n-1)$ and $N[1 - 2r/n(n-1)]^{N-1}$, respectively.

For a more statistical example, consider a population network with a multivariate y having means and variances that should be estimated by using the information from a subnet generated by a vertex sample S selected according to a specified probabilistic design. Obviously any model assumptions that explain the multivariate structure between the variables in y should be beneficial in order

to find predictors of the estimated expected values of the means and variances of the variables in y.

Further information about the literature on survey sampling in networks can be found in the following references.

About the Author

Professor Frank is one of the pioneers in network sampling, He has published papers on network methodology, snowball sampling, Markov graphs, clustering, and information theory. Jointly with Professor David Strauss he introduced Markov graphs in 1986 and explained how sufficient network statistics can be deduced from explicit assumptions about the dependencies in a network.

Cross References

►Network Models in Probability and Statistics
►Probabilistic Network Models
►Social Network Analysis

References and Further Reading

Frank O (2005) Network sampling and model fitting. In: Carrington P, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York, pp 31–56

Frank O (2009) Estimation and sampling in social network analysis. In: Meyers R (ed) Encyclopedia of complexity and systems science. Springer, New York, pp 8213–8231

Frank O (2011) Survey sampling in networks. In: Scott J, Carrington P (eds) Handbook of social network analysis. Sage, London

Kolaczyk E (2009) Statistical analysis of network data. Springer, New York

Neural Networks

Leslie S. Smith
Professor, Chair of Department of Computing Science and Mathematics, Head of the Computational Intelligence Research Group
University of Stirling, Stirling, UK

Introduction

Neural networks emulate aspects of real neural systems to attempt to achieve some of the remarkable capabilities of real neural systems, or to help understand the working of neural systems. There are many types of neural networks, some concerned with data prediction and classification (which are of interest here), and others with detailed understanding of the operation of real neural systems. What they all share is modelling (on a digital computer) some simplification of the extremely complicated and sophisticated structure of real neural systems: in essence parallel operation of relatively simple but highly interconnected processing elements (model neurons).

Neural networks that are of interest to statisticians use highly simplified model neurons, and interconnections between neurons (synapses). The earliest such networks were based on the seminal work of McCulloch and Pitts (1943) who characterized neurons as threshold logic devices. These were used by Rosenblatt and Minsky to build systems which they called Perceptrons. Such systems are inherently limited (as was shown by Minsky and Papert [1969]), and it was not until the 1980s that modern ideas of using of Neural Networks for statistical purposes came to the fore. (A more detailed history may be found in Haykin [2009], Chap. 1.)

Neural networks consist of a model M, with a set of parameters P. M is one of a generic set of models such as those below. An important concept in neural networks is that the parameters *adapt* as a result of the data applied to the network. The initial issues are choosing the right model M, then choosing an initial set of parameters, and an effective way to adapt the parameters P (a learning rule). In addition, it is necessary to check that the resultant parameter values actually solve the problem. Often a number of models and initial parameter values are used, and then the final results assessed. Neural network adaptation techniques allow them to approximate regression techniques. Because the networks include non-linear elements they can go beyond linear regression. A detailed discussion of the nature of the relationship between neural networks and statistical techniques may be found in Bishop (1995).

Types of Neural Network

Neural networks may be supervised or unsupervised. Supervised neural networks (sometimes characterised as *having a teacher*) are provided with sets of inputs and their corresponding outputs (*training data*) and internally adapt their parameters to attempt *learn* how to map (new) inputs into outputs in a way consistent with the training data. Unsupervised neural networks also receive data which they use to alter their internal structure, but in this case, there are no corresponding outputs: the network needs to approximate some aspects of the internal structure of the training data. Normally, the training data is divided into data used for training, and data used to test the resultant network. Supervised neural networks are used for prediction (function approximation) and classification, and unsupervised networks are normally used for data dimensionality reduction, and classification purposes.

Supervised Neural Networks

Most supervised networks are *feedforward*: that is, the external input is connected to some neurons, and these are then connected to others in turn in such a way that there are no loops. These are in general easier to train than recurrent neural networks, but can have no sensitivity to the history of their inputs. The simplest such network is the simple Delta-unit network (SimpleΔ), and consists of a set of input units and an output unit: however such networks are limited in what they can learn.

In the SimpleΔ network, the output unit calculates a weighted sum of the N inputs X_i, (plus a bias term, written here as $w_0 X_0$, with X_0 always 1).

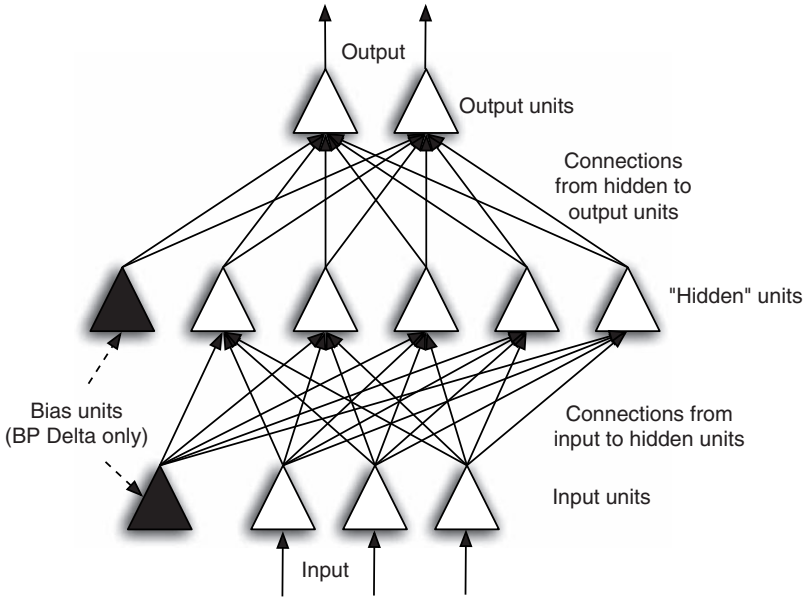
$$A = \Sigma_{(i=0)}^{N} (w_i X_i) \tag{1}$$

where A is the activation of the output unit, and the w_i are the weights between the input units and the output unit. Often this activation is then compressed, for example using a logistic function to produce the output Y so that $Y = 1/(1+\exp(-KA))$. If D is the desired output for a given input vector $\bar{X} = (X_0, X_1, \ldots, X_N)$ then it can be shown that adjusting the weights using

$$\delta w_i = k(D - Y)\frac{dY}{dA}X_i \tag{2}$$

will result in making D closer to Y for small positive k. However, the range of functions (or classifications) that can be represented by such a network is restricted to linearly seperable functions. To enable a more general class of problems to be solved requires that (for example) increasing a value for an input x_i can sometimes increase and sometimes decrease Y, and this requires additional model neurons: these are called *hidden neurons* as they are neither input nor output neurons. They are usually arranged in layers.

The two commonest forms of feedforward supervised neural networks are back-propagated Delta rule networks (BPΔ) and radial basis function (RBF) networks: they (largely) share a common architecture, illustrated in Fig. 1. The primary difference between BPΔ and RBF networks is the nature of the hidden layer units. In BPΔ networks, these calculate a weighted sum of their inputs, then non-linearly squash this output: typically, a logistic function is used for squashing. In this case, if Y is the output,



Neural Networks. Fig. 1 Feedforward neural network. Input arrives at the (three) input units, and is transferred through the connections from input to hidden units to the hidden units. For back-propagated delta-rule networks, there is one other input to each hidden unit, a bias input from a unit (in *black*) whose output is always 1. The output from the hidden units passes through connections to the output units, (again, for back-propogated delta-rule networks there is an extra bias input). The output units provide the output. Note that this is a *fully connected layered feedforward* net because all units in each layer are connected to all units in the next layer

$$Y = 1/\left(1 + \exp(-K\,\Sigma_{(i=0)}^{N}(w_i X_i))\right) \quad (3)$$

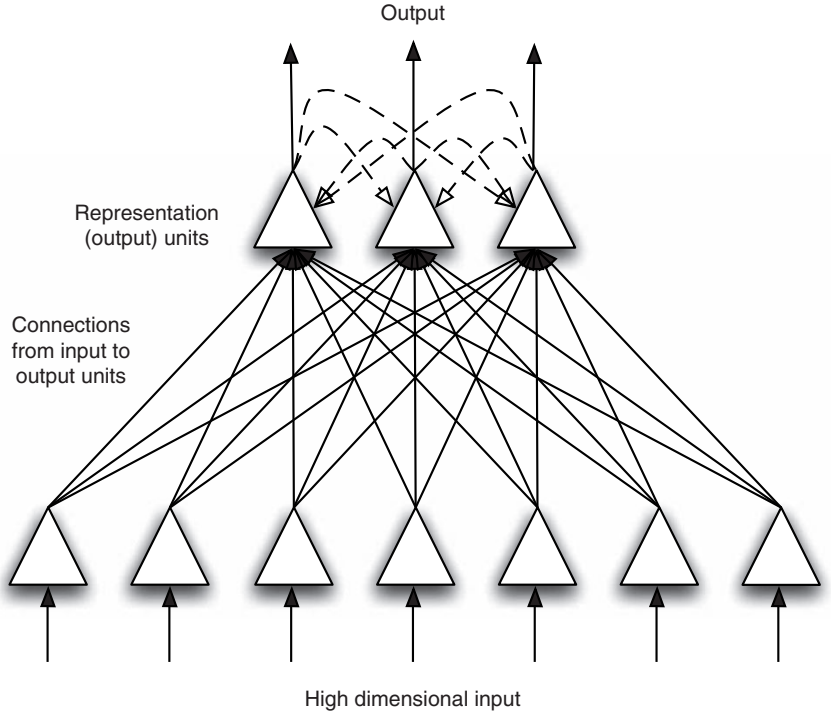
where $w_i X_i$ is weighted input, N is the number of units in the previous layer (including the bias unit), and K is a positive number which determines the steepness of the logistic function. The network adapts by altering all of the w_i. Here M is the network architecture (number of layers and of hidden units in each layer), and P is the w_i. The network shown in Fig. 1 has a single hidden layer with five hidden units: there may be multiple hidden layers, with different numbers of units in each.

In RBF networks, the hidden units compute a high dimensional Gaussian of their input. Thus they have a single maximum, and the region of the input space to which they respond is localized, unlike the BPΔ hidden units. The Gaussian may have different standard deviations (SDs) in different dimensions. Typically, a relatively large number of hidden units is used, and the Gaussian centers and SDs are fixed. (There may be an initial stage in which these centers and SDs are adjusted so that the hidden units together cover the whole of the region of the input space populated by the actual inputs.) In training, only the connections between the hidden units and the output units are adapted, usually using a simple Delta-rule technique. Thus M is the RBF units (and usually their centers and standard deviations), and P is the weights from the RBFs to the output units. Only a single layer of hidden units is used.

For the BPΔ network, the weights are adjusted by propagating the errors back through the feedforward architecture (see Haykin [2009] for details). This uses simple locally calculated gradient descent, although steepest descent is also possible, though the change in weight then requires non-local values for its computation. Because the relationship between the weights and the overall error may be very complex, the BPΔ network can become trapped in local minima (which is why normal practice is to start from different locations in weight space). For the RBF network, the weights to the output units are adjusted using Eq. (2). The RBF network may also be sensitive to the placing and radii of the RBF units: often the centers are initialized to some of the training data, and the radii calculated from initial training data. However, because the weights are only a single layer, gradient descent algorithms are guaranteed to find a global minimum of the error.

Unsupervised Neural Networks

Unsupervised neural networks adapt the connections between neurons as a result of the input to the network.



Neural Networks. Fig. 2 Simple self-organized neural network. Input arrives at the (seven) input units, and is transferred through the connections from these to the representation or output units. These units may be interconnected (*dotted lines*), or they may interact in an algorithmically defined way

Typically, these consist of a single layer of input units providing adjustable weighted inputs to a layer of representation (and also output) units, as shown in Fig. 2.

Generally, the weighted connections from the input to the representation units are adjusted so that the unit that received the greatest activation ("won") will be more likely to win in the future: clearly this leaves open the possibility that some units may never win, or that some units may win too often, so that there may also be adjustment of the sensitivity of output units as well. A more sophisticated self-organized network was designed by Kohonen (1989), and it places a geometry on the representation units so that one can talk of different output units being nearer or further from each other: the learning algorithm used also attempts to ensure that the whole output space is used. This allows the data in the original high-dimensional input space to be projected down to a much lower-dimensional output space: this can be very useful when attempting to understand the structure of, or to classify high dimensional datasets.

Other Architectures

Some network architectures also allow adaptation of the architecture: these often have phases of adding neurons, followed by phases of removing them. The aim is to produce a simple M which can then provide a simple model of the data being analysed. Though not a new idea (Frean 1990; Hassibi et al. 1993), this area has seen recent interest (Franco et al. 2010). Recurrent neural networks (networks with feedback loops) are also used for prediction and classification where the network needs to be sensitive to previous inputs. Training in such networks can present convergence issues: the algorithms used in the BPΔ and RBF networks require the absence of loops. Details of some training techniques may be found in Haykin (2009).

About the Author

Professor Leslie Smith is Chair of the Department of Computing Science and Mathematics at the University of Stirling, Scotland. He was instigator of and first Chair of the IEEE United Kingdom and Republic of Ireland Computational Intelligence Chapter. He has written about 100 peer-reviewed papers. He is a Senior Member of the IEEE.

Cross References

►Bayesian Reliability Modeling
►Business Forecasting Methods
►Data Mining
►Misuse of Statistics
►Simulation Based Bayes Procedures for Model Structures with Non-Elliptical Posteriors
►Statistics: An Overview

References and Further Reading

Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, Oxford
Franco L, Elizondo DA, Jerez JM (eds) (2010) Constructive neural networks. Studies in computational intelligence, vol 258. Springer, Berlin
Frean M (1990) The upstart algorithm: a method for constructing and training feed-forward neural networks. Neural Comput 2(2):189–209
Hassibi B, Stork DG, Wolf G (1993) Optimal brain surgeon and general network pruning. In: Proceedings of the 1993 IEEE international conference on neural networks, IEEE, pp 293–300
Haykin S (2009) Neural networks and learning machines, 3rd edn. Pearson Education, Upper Saddle River
Kohonen T (1989) Self-Organization and associative memory, 3rd edn. Springer, Berlin
McCulloch WS, Pitts W (1943) A logical calculus of ideas immanent in nervous activity. B Math Biophys 5:115–133
Minsky ML, Papert SA (1969) Perceptrons. MIT Press, Cambridge

Neyman-Pearson Lemma

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

Neyman–Pearson lemma (also called fundamental lemma) presented in 1933 is the basic tool in testing statistical hypotheses. Its essence consists of the following mathematical problem.

Given a measure μ on a measurable space $(\mathcal{X}, \mathcal{A})$ and given nonnegative measurable real functions f_0 and f_1 on $(\mathcal{X}, \mathcal{A})$ satisfying the condition $\int_{\mathcal{X}} f_i(x)d\mu(x) = 1$, for $i = 1, 2$, consider the family $\mathcal{S} = \mathcal{S}(\alpha)$ of all $\mathcal{A}$-measurable subsets S of $\mathcal{X}$ such that

$$\int_S f_0(x)d\mu(x) \le \alpha, \text{ where } \alpha \text{ is a given positive constant, non greater than 1.} \quad (1)$$

Find all sets in $\mathcal{S}(\alpha)$ maximizing the integral $\int_S f_1(x)d\mu(x)$.

Neyman-Pearson lemma states that the family of the sets S in $\mathcal{S}(\alpha)$ of the form $\{x \in \mathcal{X} : f_1(x) > kf_0(x)\}$ for some nonnegative constant k is not empty and it includes the desired solution. Moreover, any solution may be presented in this form almost surely w.r.t μ.

A version of this lemma for *randomized tests* is more useful. As known a randomized test is represented by a real function $\phi = \phi(x)$ on $\mathcal{X}$ taking values in the interval $[0, 1]$ which plays the role of a fuzzy set in $\mathcal{X}$. In consequence, the

condition (1) and the integral $\int_{\mathcal{S}} f_1(x)d\mu(x)$ are replaced by $\int_{\mathcal{X}} \phi(x)f_0(x)d\mu(x) \le \alpha$, and $\int_{\mathcal{X}} \phi(x)f_1(x)d\mu(x)$, respectively. In this case the Neyman–Pearson has the following statistical interpretation.

Suppose a random variable X has a density f_X, being one of the functions f_0 or f_1. Then one of the equivalent *most powerful* (MP) tests for the hypothesis $H : f_X = f_0$ under the alternative $K : f_X = f_1$ at the significance level α may be presented in the form

$$\phi(x) = \begin{cases} 1, & \text{if } f_1(x) > kf_1(x) \\ \gamma, & \text{if } f_1(x) = kf_1(x), \\ 0, & \text{if } f_1(x) < kf_1(x) \end{cases} \tag{2}$$

where nonnegative k and γ belonging to the interval $[0,1]$ are determined by the condition

$$\int_{\mathcal{X}} \phi(x)f_0(x)d\mu(x) = \alpha, \tag{3}$$

unless there exists a test ϕ of size less than α with power 1.

The Neyman–Pearson lemma was also generalized for testing a hypothesis of type $H : f_X \in \{f_1, \ldots, f_m\}$ against a simple alternative $K : f_X = f_{m+1}$..

Example Random value X takes values $-1, 0, 1, 2, 3, 4, 5$ according to one of the distributions given in the table.

x	–1	0	1	2	3	4	5
$P_0(X = x)$	0.08	0.15	0.2	0.1	0.45	0.01	0.01
$P_1(X = x)$	0.33	0.1	0.1	0.1	0.05	0.02	0.3

Construct the MP test for the hypothesis H: $P = P_0$ against the alternative K: $P = P_1$ at the significance level $\alpha = 0.05$.

First we compute the likelihood rations $q(x) = \frac{P_1(X=x)}{P_0(X=x)}$ and range them in the descending order. Let $r(x)$ denote the rank of $q(x)$ while x_i – the value x corresponding to the rank i. Further steps may be observed in the table

x	$q(x)$	$r(x)$	i	x_i	$f_i = P_0(X = x_i)$	$s_i = \sum_{j \le i} f_j$
–1	4.13	2	1	5	0.01	0.01
0	0.67	5	2	–1	0.08	0.09
1	0.5	6	3	4	0.01	0.1
2	1.0	4	4	2	0.1	0.2
3	0.11	7	5	0	0.15	0.35
4	2	3	6	1	0.2	0.55
5	30	1	7	3	0.45	1.0

Now we only need to apply the conditions (2) and (3) in discrete form. The last column in the table shows that the MP nonrandomized test for the hypothesis H against K at the significance level 0.05 rejects H, if and only if $X = 5$. Moreover its power is equal to 0.3. The corresponding MP randomized test is defined *by*

$$\phi(x) = \begin{cases} 1, & \text{if} \quad X = 5 \\ 0.5, & \text{if} \quad X = -1 \\ 0, & \text{otherwise.} \end{cases}$$

Its power is equal to 0.47.

About the Author

For biography *see* the entry ►Random Variable.

Cross References

►Bayesian Versus Frequentist Statistical Reasoning
►Frequentist Hypothesis Testing: A Defense
►Fuzzy Logic in Statistical Data Analysis
►Most Powerful Test
►Psychology, Statistics in
►Robust Inference
►Significance Testing: An Overview
►Significance Tests, History and Logic of
►Significance Tests: A Critique
►Statistical Evidence
►Statistical Inference
►Statistical Inference: An Overview
►Statistics: An Overview

References and Further Reading

Lehmann EL, Romano JP (2005) Testing statistical hypotheses, 3rd edn. Springer, New York
Neyman J, Pearson E (1933) On the problem of the most efficient tests of statistical hypotheses, Phil Trans R Stat Soc Lond Ser A 231:289–337
Pfanzagl J (1994) Parametric statistical theory. de Gruyter, Berlin
Zacks S (1981) Parametric statistical inference. Pergamon, Oxford

Nonlinear Mixed Effects Models

Marie Davidian
William Neal Reynolds Professor, Director,
Center for Quantitative Sciences in Biomedicine, North Carolina State University, Raleigh, NC, USA

The *nonlinear mixed effects model* (or *hierarchical nonlinear model*) is a standard framework for analysis of data in the form of continuous repeated measurements over time on each individual from a sample of individuals drawn from a population of interest. It is particularly relevant when the goal is to make inference on the average behavior and variability in the population of individuals of features underlying the individual profiles, where the profiles are well-represented by a function nonlinear in parameters that characterize the features under study.

A key area where this model is appropriate is in pharmacokinetic analysis, where repeated blood samples are collected from each of several subjects at intermittent time points following a dose or doses of a drug, from which continuous drug concentration measurements are ascertained. The objective is to characterize the underlying pharmacological processes within the body that lead to observed concentrations and how these processes vary across subjects in the population. To describe the processes formally, it is routine to represent the body of an individual subject by a simple compartment model, which yields an expression for concentration at any time post-dose as a function nonlinear in parameters that quantify absorption, distribution, and elimination of the drug for the subject. For example, representation of the body by a single "blood compartment," with oral dose D given at time $t = 0$, leads to the standard model for concentration at time t, $C(t)$, given by

$$C(t) = \frac{Dk_a}{V(k_a - Cl/V)}\left\{\exp(-k_a t) - \exp\left(-\frac{Cl}{V}t\right)\right\}, \quad (1)$$

where k_a is the fractional rate of absorption of the drug from the gut into the bloodstream; V is roughly the volume required to account for all drug in the body; and Cl is the clearance rate, the volume of blood from which drug is eliminated per unit time. In (1), the goal is to learn about the parameters (k_a, V, Cl) that summarize pharmacokinetic processes for a given subject and their mean or median values and the extent of variation of them in the population of subjects. This information is critical for designing dosage regimens to maintain drug concentrations in a desired range; if population variation in (k_a, V, Cl) is substantial, designing a regimen that will work well for most individuals may be difficult. If some of the variation is systematically associated with subject characteristics like weight or age, regimens tailored to subpopulations of subjects sharing certain characteristics may be developed. See Giltinan (2006) for an excellent review of this area.

Additional applications where the inferential goals are similar and for which the nonlinear mixed effects model is a suitable framework include growth analysis in agriculture and forestry and analysis of viral dynamics, among others. In general, relevant applications are such that a model for response–time profile at the individual level, derived from theoretical or empirical considerations, like (1), and depending on parameters characterizing directly underlying features of interest, is available and is central to the data-analytic objectives.

For each individual i in a sample of N individuals from a population of interest, $i = 1, \ldots, N$, let Y_{ij} denote a continuous, univariate response (e.g., drug concentration) at time t_{ij}, $j = 1, \ldots, n_i$. Let $\boldsymbol{U}_i$ denote a vector of covariates specifying conditions under which i is observed; for example, in (1), $\boldsymbol{U}_i = D_i$ = oral dose given to individual i at time 0; in the case of multiple doses, $\boldsymbol{U}_i$ would summarize the times and corresponding doses given. The $\boldsymbol{U}_i$ are needed to describe the response–time relationship at the level of individual i, as in (1), and hence are often referred to as "within-individual" covariates. Assume further that a vector of characteristics that do not change over the observation period on i are recorded, such as age, ethnicity, weight, and so on; summarize these in a vector $\boldsymbol{A}_i$, often called "among-individual" covariates because they characterize how individuals may differ but are not required to describe individual response–time relationships. The available data are then $(\boldsymbol{Y}_i, \boldsymbol{U}_i, \boldsymbol{A}_i)$, $i = 1, \ldots, N$, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, which, combining the within- and among-individual covariates as $\boldsymbol{X}_i = \left(\boldsymbol{U}_i', \boldsymbol{A}_i'\right)'$, may be written more succinctly as $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$.

The model may be conceived as a two stage hierarchy. At the first stage, a model for data at the level of a given individual i is specified. Let $m(t, \boldsymbol{U}, \boldsymbol{\theta})$ be a function of time, within-individual conditions of measurement, and a vector of parameters $\boldsymbol{\theta}$ characterizing features underlying the response-time profile; e.g., $m(t, \boldsymbol{U}, \boldsymbol{\theta})$ is $C(t)$, $\boldsymbol{U} = D$, and $\boldsymbol{\theta} = (k_a, V, Cl)'$ in (1). Key to the development is that individual i is assumed to have individual-specific such parameters $\boldsymbol{\theta}_i$ governing his/her trajectory $m(t, \boldsymbol{U}_i, \boldsymbol{\theta}_i)$; in (1), $\boldsymbol{\theta}_i = (k_{ai}, V_i, Cl_i)' = (\theta_{i1}, \theta_{i2}, \theta_{i3})'$. Writing $\boldsymbol{m}_i(\boldsymbol{U}_i, \boldsymbol{\theta}_i) = \{m(t_{i1}, \boldsymbol{U}_i, \boldsymbol{\theta}_i), \ldots, m(t_{in_i}, \boldsymbol{U}_i, \boldsymbol{\theta}_i)\}'$, it is assumed that

$$E(\boldsymbol{Y}_i \mid \boldsymbol{U}_i, \boldsymbol{\theta}_i) = \boldsymbol{m}_i(\boldsymbol{U}_i, \boldsymbol{\theta}_i), \quad \text{so that} \quad E(Y_{ij} \mid \boldsymbol{U}_i, \boldsymbol{\theta}_i) = m(t_{ij}, \boldsymbol{U}_i, \boldsymbol{\theta}_i). \tag{2}$$

A model for $\mathrm{Cov}(\boldsymbol{Y}_i \mid \boldsymbol{U}_i, \boldsymbol{\theta}_i)$ is also specified. The diagonal elements embody assumptions regarding measurement error, sampling variation, and the fact that error-free realizations of the response–time process may not fall directly on the trajectory $m(t, \boldsymbol{U}_i, \boldsymbol{\theta}_i)$ because, for example, it is a simplified representation of a complex biological phenomenon (e.g., as (1) is a gross simplification of the underlying physiology). The off-diagonal elements are dictated by assumptions on possible autocorrelation among the Y_{ij}, $j = 1, \ldots, n_i$, given $(\boldsymbol{U}_i, \boldsymbol{\theta}_i)$. Due to the intermittent nature of the times of measurement $t_{i1}, \ldots, t_{in_i}$, autocorrelation between any pair of measurements is often assumed negligible; however, this need not always be the case. Combining these, it is assumed that

$$\mathrm{Cov}(\boldsymbol{Y}_i \mid \boldsymbol{U}_i, \boldsymbol{\theta}_i) = V_i(\boldsymbol{U}_i, \boldsymbol{\theta}_i, \boldsymbol{\alpha}), \tag{3}$$

where $\boldsymbol{\alpha}$ is the collection of parameters used to describe within-subject variance and correlation. See Davidian (2009) for discussion of considerations involved in specifying model for $\mathrm{Cov}(\boldsymbol{Y}_i \mid \boldsymbol{U}_i, \boldsymbol{\theta}_i)$. The most common model, sometimes adopted by default without adequate consideration of its rather strong assumptions, is $V_i(\boldsymbol{U}_i, \boldsymbol{\theta}_i, \boldsymbol{\alpha}) = \sigma^2 I_{n_i}$, where $\boldsymbol{\alpha} = \sigma^2$; I_n is a $(n \times n)$ identity matrix; and σ^2 represents a constant variance due to, say, measurement error and variation in realizations, that is the same across all subjects.

Along with (2) and (3), the first stage individual level model is ordinarily completed by an assumption on the distribution of $\boldsymbol{Y}_i$ given $(\boldsymbol{U}_i, \boldsymbol{\theta}_i)$. A standard such assumption is that this distribution is normal with mean (2) and covariance matrix (3); this may be reasonable for some types of responses on a transformed scale, in which case Y_{ij} may be taken to be a transformation of the original response in the foregoing discussion.

At the second stage, a model for the population describes possible systematic relationships between individual-specific parameters $\boldsymbol{\theta}_i$ and individual characteristics $\boldsymbol{A}_i$ in the population as well as the "inherent" variation among the $\boldsymbol{\theta}_i$ once a such relationships are taken into account. A general population model is written as

$$\boldsymbol{\theta}_i = \boldsymbol{d}(\boldsymbol{A}_i, \boldsymbol{\beta}, \boldsymbol{b}_i), \tag{4}$$

where $\boldsymbol{d}$ is a r-dimensional function that describes the relationship between $\boldsymbol{\theta}_i$ and $\boldsymbol{A}_i$ in terms of a parameter $\boldsymbol{\beta}$ $(p \times 1)$ and random effects $\boldsymbol{b}_i$ $(q \times 1)$ accounting for the additional inherent variation. The $\boldsymbol{b}_i$ are typically taken to be independent across i, and the $\boldsymbol{b}_i$ are often assumed independent of the $\boldsymbol{A}_i$, $i = 1, \ldots, N$, with $E(\boldsymbol{b}_i) = 0$ and $\mathrm{Cov}(\boldsymbol{b}_i) = G$ for unstructured covariance matrix G; moreover, it is routine to assume that the $\boldsymbol{b}_i$ are normally distributed with these moments. As an example, in the context of (1), supposing $\boldsymbol{A}_i = (w_i, \delta_i)'$, with w_i weight and δ_i an indicator of creatinine clearance, where $\delta_i = 1$ if >50 mL/min, consider $k_{ai} = \theta_{i1} = d_1(\boldsymbol{A}_i, \boldsymbol{\beta}, \boldsymbol{b}_i) = \exp(\beta_1 + b_{i1})$, $V_i = \theta_{i2} = d_2(\boldsymbol{A}_i, \boldsymbol{\beta}, \boldsymbol{b}_i) = \exp(\beta_2 + b_{i2})$, and $Cl_i = \theta_{i3} = d_3(\boldsymbol{A}_i, \boldsymbol{\beta}, \boldsymbol{b}_i) = \exp(\beta_3 + \beta_4 w_i + \beta_5 \delta_i + b_{i3})$, where $\boldsymbol{b}_i = (b_{i1}, b_{i2}, b_{i3})'$ $(q = 3)$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)'$ $(p = 5)$. This model enforces positivity of k_{ai}, V_i, and Cl_i; moreover, if the $\boldsymbol{b}_i$ are normal, then the distributions of their components, and thus those of the $\boldsymbol{\theta}_i$, are lognormal (and hence skewed), which is common in this application. Alternatively, an analogous specification would be to reparameterize (1), taking $\boldsymbol{\theta}_i = \left(k_{ai}^*, V_i^*, Cl_i^*\right)'$, where $k_{ai}^* = \log(k_{ai})$, $V_i^* = \log(V_i)$, and $Cl_i^* = \log(Cl_i)$, and to write $k_{ai}^* = \beta_1 + b_{i1}$, $V_i^* = \beta_2 + b_{i2}$, and $Cl_i^* = \beta_3 + \beta_4 w_i + \beta_5 \delta_i + b_{i3}$. Here, the population model (4) may be written in the linear form $\boldsymbol{\theta}_i = A_i \boldsymbol{\beta} + B_i \boldsymbol{b}_i$, where A_i and B_i are "design matrices" specifying dependence on $\boldsymbol{A}_i$ and whether or not all elements of $\boldsymbol{\theta}_i$ have associated random effects, respectively; see Davidian (2009). A linear population model is the default specification in many papers and software packages.

The nonlinear mixed effects model may be summarized as is conventional by writing the stage 1, individual level model as in (2) and (3), substituting the population model (4), so that conditioning is with respect to $\boldsymbol{X}_i$ and the random effects $\boldsymbol{b}_i$, as follows:

Stage 1: Individual-Level Model

$$E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{b}_i) = \boldsymbol{m}_i(\boldsymbol{U}_i, \boldsymbol{\theta}_i) = \boldsymbol{m}_i(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i),$$

$$\mathrm{Cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{b}_i) = V_i(\boldsymbol{U}_i, \boldsymbol{\theta}_i, \boldsymbol{\alpha}) = V_i(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\alpha}) \tag{5}$$

Stage 2: Population Model

$$\boldsymbol{\theta}_i = \boldsymbol{d}(\boldsymbol{A}_i, \boldsymbol{\beta}, \boldsymbol{b}_i), \qquad \boldsymbol{b}_i \sim (0, G). \tag{6}$$

Standard assumptions are that the distribution of $\boldsymbol{Y}_i$ given $(\boldsymbol{X}_i, \boldsymbol{b}_i)$ is normal with moments (5); that $\boldsymbol{b}_i \sim N(0, G)$, independently of $\boldsymbol{A}_i$ (and $\boldsymbol{U}_i$, and hence $\boldsymbol{X}_i$); and that $\boldsymbol{b}_i$ are independent across i.

The usual objective is inference on $\boldsymbol{\beta}$ and G, corresponding to average behavior of and extent of variation in features underlying individual profiles in the population. The obvious approach to inference in (5), (6) is maximum likelihood. Letting $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$, and writing the conditional density of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$ as $f_i(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\gamma}, G)$, by independence across i, the loglikelihood for (γ, G) based on the

observed data $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$, $i = 1, \ldots, N$, is

$$\ell(\boldsymbol{\gamma}, G) = \log\left\{\prod_{i=1}^{N} f_i(\boldsymbol{y}_i \,|\, \boldsymbol{x}_i;\, \boldsymbol{\gamma}, G)\right\} = \log\left\{\prod_{i=1}^{N} \int f_i(\boldsymbol{y}_i \,|\, \boldsymbol{x}_i, \boldsymbol{b}_i;\, \boldsymbol{\gamma}) f(\boldsymbol{b}_i;\, G)\, d\boldsymbol{b}_i\right\}, \quad (7)$$

where $f_i(\boldsymbol{y}_i \,|\, \boldsymbol{x}_i, \boldsymbol{b}_i;\, \boldsymbol{\gamma})$ is the conditional density of $\boldsymbol{Y}_i$ given $(\boldsymbol{X}_i, \boldsymbol{b}_i)$ assumed in stage 1; and $f(\boldsymbol{b}_i;\, G)$ is the density of the $\boldsymbol{b}_i$ assumed in stage 2, e.g., the $N(0, G)$ density. From the right side of (7), $\ell(\boldsymbol{\gamma}, G)$ involves N almost certainly analytically intractable q-dimensional integrals, so maximization of $\ell(\boldsymbol{\gamma}, G)$ in $(\boldsymbol{\gamma}, G)$ requires a way of numerically evaluating or analytically approximating these integrals.

A natural approach when $f(\boldsymbol{b}_i;\, G)$ is a normal density is Gaussian quadrature, which approximates the integral by a weighted average of the integrand over a q-dimensional grid. Although accuracy increases with the number of grid points, the larger is q, the greater the computational burden, but reducing the number of grid points in each dimension can compromise accuracy. A variant, adaptive Gaussian quadrature, can reduce the number of grid points needed (Pinheiro and Bates 2005); regardless, because evaluation of the N integrals must be carried out at each internal iteration of the optimization algorithm for maximizing (7), computation can be challenging unless q is small (e.g., $\leq$ 3). The SAS procedure `nlmixed` (SAS Institute 2009) offers options for maximization of (7) using Gaussian or adaptive Gaussian quadrature to "do" the integrals.

Alternative methods accordingly seek to avoid the integrations via analytical approximations to $f_i(\boldsymbol{y}_i \,|\, \boldsymbol{x}_i;\, \boldsymbol{\gamma}, G)$ in (7). The two main approaches, which assume $f_i(\boldsymbol{y}_i \,|\, \boldsymbol{x}_i, \boldsymbol{b}_i;\, \boldsymbol{\gamma})$ and $f(\boldsymbol{b}_i;\, G)$ are normal, follow from writing (5), (6) equivalently as $\boldsymbol{Y}_i = \boldsymbol{m}_i(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i) + V^{1/2}(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\alpha})\, \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i$ is $N(0, I_{n_i})$ conditional on $(\boldsymbol{X}_i, \boldsymbol{b}_i)$. Taking a linear Taylor series about $\boldsymbol{b}_i = \boldsymbol{b}_i^*$ "close" to $\boldsymbol{b}_i$ and disregarding negligible terms leads to

$$\boldsymbol{Y}_i \approx \boldsymbol{m}_i\left(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i^*\right) - Z_i\left(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i^*\right) \boldsymbol{b}_i^* + Z_i\left(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i^*\right) \boldsymbol{b}_i + V_i^{1/2}\left(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i^*, \boldsymbol{\alpha}\right) \boldsymbol{\epsilon}_i, \quad (8)$$

where $Z_i\left(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i^*\right) = \partial / \partial \boldsymbol{b}_i \{\boldsymbol{m}_i(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{b}_i)\}|_{\boldsymbol{b}_i = \boldsymbol{b}_i^*}$. Taking $\boldsymbol{b}_i^* = 0$, the mean of $\boldsymbol{b}_i$, (8) implies that the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$ is approximately normal with

$$E(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i) \approx \boldsymbol{m}_i(\boldsymbol{X}_i, \boldsymbol{\beta}, 0), \quad \mathrm{Cov}(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i) \approx Z_i(\boldsymbol{X}_i, \boldsymbol{\beta}, 0)\, G\, Z_i'(\boldsymbol{X}_i, \boldsymbol{\beta}, 0) + V_i(\boldsymbol{X}_i, \boldsymbol{\beta}, 0, \boldsymbol{\alpha}). \quad (9)$$

This suggests replacing $f_i(\boldsymbol{y}_i \,|\, \boldsymbol{x}_i;\, \boldsymbol{\gamma}, G)$ in (7) by the corresponding normal density and maximizing the resulting likelihood, first proposed by Beal and Sheiner (1982) in the pharmacokinetics literature. This and related methods based on (9) with $\boldsymbol{b}_i^* = 0$ are referred to as "first order" methods and are available in SAS `proc nlmixed`, the SAS macro `nlinmix` (Littell et al. 2006), and in the pharmacokinetics package `NONMEM` (2006).

The "first order" methods may yield too crude an approximation to the true $E(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i)$ and $\mathrm{Cov}(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i)$, resulting in inconsistent inferences on $\boldsymbol{\beta}$ and G. The more "refined" "first order conditional" approximation takes $\boldsymbol{b}_i^*$ in (8) to be the mode $\widehat{\boldsymbol{b}}_i$ of the posterior density $f_i(\boldsymbol{b}_i \,|\, \boldsymbol{y}_i, \boldsymbol{x}_i;\, \boldsymbol{\gamma}, G)$ implied by (5), (6), and takes the distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{X}_i$ to be approximately normal with

$$E(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i) \approx \boldsymbol{m}_i(\boldsymbol{X}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i) - Z_i(\boldsymbol{X}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\widehat{\boldsymbol{b}}_i,$$
$$\mathrm{Cov}(\boldsymbol{Y}_i \,|\, \boldsymbol{X}_i) \approx Z_i(\boldsymbol{X}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i)\, G\, Z_i'(\boldsymbol{X}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i) + V_i(\boldsymbol{X}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{b}}_i, \boldsymbol{\alpha}). \quad (10)$$

Approaches based on (10) iterate between update of the $\widehat{\boldsymbol{b}}_i$ holding the current estimates of $(\boldsymbol{\gamma}, G)$ fixed and maximization in $(\boldsymbol{\gamma}, G)$ of the approximate loglikelihood implied by (10) for fixed $\widehat{\boldsymbol{b}}_i$ and are available in the SAS macro `nlinmix`, `NONMEM`, and the `R` package `nlme` (R Development Core Team 2009).

Another tactic, relevant only when the n_i are sufficiently large to permit individual-specific fitting of the stage 1 model is to use the resulting $\widehat{\boldsymbol{\theta}}_i$ as "data" to fit the stage 2 model; see Davidian and Giltinan (1995, Sect. 5.3) and Davidian (2009). It is also possible to place (5), (6) within a Bayesian framework, adding at third, "hyperprior" stage to the hierarchy with an assumed prior density for $(\boldsymbol{\gamma}, G)$. As shown by Wakefield et al. (1994) and Davidian and Giltinan (1995, Chap. 8), Bayesian inference may be implemented via ▶Markov chain Monte Carlo (MCMC) and carried out using software such as `WinBUGS` (Lunn et al. 2000).

Of necessity, this brief review covers only the basic formulation of the model and selected implementation strategies and software. Alternative strategies, model extensions, and further details are given in the aforementioned references and in the vast literature on this topic; see Davidian (2009) for a bibliography.

About the Author

Dr. Marie Davidian is William Neal Reynolds Professor of Statistics and Director of the Center for Quantitative Sciences in Biomedicine at North Carolina State University (NCSU), and Adjunct Professor of Biostatistics and Bioinformatics at Duke University. She is Fellow, American

Statistical Association (1998), Fellow, Institute of Mathematical Statistics (2006), and Fellow, American Association for the Advancement of Science (2006). Dr. Davidian was awarded the American Statistical Association Award for Outstanding Statistical Application (1993), and the Janet L. Norwood Award for Outstanding Achievement by a Woman in the Statistical Sciences, University of Alabama at Birmingham (2007). In 2009, she received the George W. Snedecor Award, Committee of Presidents of Statistical Societies "for her fundamental contributions to the theory and methodology of longitudinal data, especially nonlinear mixed effects models; significant contributions to the analysis of clinical trials and observational studies; and her leadership as president of ENAR (Eastern North American Region/International Biometric Society) and as a member of the International Biometric Society council." She was Associate editor of the following journals: *Journal of the American Statistical Association* (1995–2001), *Biometrics* (1997–2000), and *Statistica Sinica* (2003–2005). She was also Coordinating Editor of *Biometrics* (2000–2002). Currently, she is Executive Editor, *Biometrics* (2006–2011). Professor Davidian has (co-)authored about 85 refereed papers and several books, including *Nonlinear Models for Repeated Measurement Data* (with Giltinan, D.M., London, Chapman & Hall, 1995), and *Applied Longitudinal Data Analysis* (with Fitzmaurice, G., Verbeke, G., Molenberghs, G., New York, Springer, 2009). She has supervised more than 20 Ph.D. students.

Cross References

►Linear Mixed Models
►Nonlinear Models
►Repeated Measures
►Statistical Inference in Ecology

References and Further Reading

Beal SL, Sheiner LB (1982) Estimating population pharmacokinetics. CRC Crit Rev Biomed Eng 8:195–222

Davidian M (2009) Non-linear mixed-effects models. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G (eds) Longitudinal data analysis. Chapman & Hall/CRC Press, Boca Raton, pp 108–141

Davidian M, Giltinan DM (1995) Nonlinear models for repeated measurement data. Chapman & Hall/CRC Press, London

Davidian M, Giltinan DM (2003) Nonlinear models for repeated measurement data: an overview and update. J Agri Biol Environ Stat 8:387–419

Giltinan DM (2006) Pharmacokinetics and pharmacodynamics. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, 2nd edn. Wiley, Hoboken, pp 4049–4062

Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2006) SAS for mixed models, 2nd edn. SAS Institute, Cary

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325–337

NONMEM (2006) http://www.icondevsolutions.com/nonmem.htm

Pinheiro JC, Bates DM (1995) Approximation to the log-likelihood function in the nonlinear mixed effects model. J Comput Graph Stat 4:12–35

R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org

SAS Institute (2009) SAS/STAT User's Guide 9.2, 2nd edn. SAS Institute, Cary

Wakefield JC, Smith AFM, Racine-Poon A, Gelfand AE (1994) Bayesian analysis of linear and nonlinear population models by using the Gibbs sampler. Appl Stat 41:201–221

Nonlinear Models

Christos P. Kitsos
Professor and Head
Technological Educational Institute of Athens, Athens, Greece

Introduction

When an experiment is performed n times, within the experimental region $X \subseteq \mathcal{R}^k$, known as design space, the outcome or response variable, Y, can be considered either as discrete variable or continuous variable. That means the set of all possible response outcomes Ψ, known as response space, can follow one of the two cases:

Case 1: Ψ is finite, i.e., $\Psi = \{0, 1, 2, \ldots, \lambda\}$ with cardinal number $\nu = \lambda + 1$. The most common case, $\nu = 2$, corresponds to binary response problems where $Y \in \Psi = \{0, 1\}$. When $n > 2$ we are referring to polytomous experiments. In some cases, like Poisson experiments, the set Ψ is countable infinite.
Case 2: Ψ has the power of the continuum, i.e., cardinality c, as Ψ can be any interval in $\mathcal{R}$.

In Case 1 and for binary problems the outcome $Y_i = 0$ or 1, $i = 1, 2, \ldots, n$ is linked with the covariate $x \in X$ and the parameter vector θ, from the parameter space $\Theta \subseteq \mathcal{R}^p$, through a probability model T as

$$p(x) = p(Y_i = 1|x) = T(x; \theta) = 1 - P(Y_i = 0|x) \quad (1)$$

Example 1 In bioassay the typical situation is to consider logit, probit, or exponential models as T, i.e.,

$$T_L(x, \theta) = \log\{p(x)(1 - p(x))\},$$
$$T_P(x, \theta) = \Phi^{-1}\{p(x)\},\ T_E(x, \theta) = \exp(-\theta x)$$

respectively, with "log," "exp," and "Φ" having their trivial meaning.

McCullagh and Nelder (1989) discussed extensively such cases and ▶generalized linear models.

Now, the latter (Case 2) situation is faced with the general regression model: A real-valued, continuous, twice differentiable function, from $X \times \Theta$, is considered to define the (assumed correct) deterministic portion $f(x, \theta)$. The error is applied as a stochastic portion additively in the form

$$Y_i = f(x_i, \theta) + e_i, \quad i = 1, 2, \ldots, n, \quad \theta \in \Theta \subseteq \mathcal{R}^p, \quad x \in X \subseteq \mathcal{R}^k. \tag{2}$$

Example 2 Typical, one variable nonlinear problems, let $x = u$, which might provide response curves with no significant difference between them, for particular values of the parameters, are

MODEL	NAME
$f_G(u, \vartheta) = \vartheta_0 \exp\{\vartheta_1 e^{\vartheta_2 u}\}$	Gompertz model
$f_J(u, \vartheta) = \vartheta_0 + \vartheta_1\{\exp(\vartheta_2 u^{\vartheta_3})\}$	Janoscheck model
$f_L(u, \vartheta) = \vartheta_0 / \{1 + \vartheta_1 \exp(\vartheta_2 u)\}$	Logistic model
$f_B(u, \vartheta) = \{\vartheta_0 + \vartheta_1 \exp(\vartheta_2 u)\}^3$	Bertalanffy model
$f_T(u, \vartheta) = \vartheta_0 + \vartheta_1 \tanh(\vartheta_2(u + \vartheta_3))$	Tanh model
$f_3(u, \vartheta) = \frac{1}{2}\vartheta_0\left\{1 + \frac{2}{\pi}\arctan(\vartheta_1(u - \vartheta_2))\right\}$	3 - Tanh model
$f_4(u, \vartheta) = \vartheta_0 + \frac{2}{\pi}\vartheta_1 \arctan(\vartheta_2(u - \vartheta_3))$	4 - Tanh model

When a real function f exists such that $f(x_i, \theta) = g(x_i)^T\theta$, then the problem is reduced to the linear regression problem. The nonlinear optimal design can be defined through two different problems:

Problem 1: The underlying model describing the physical phenomenon is nonlinear, as in (1), (2). The target then is either to fit (with ▶Least Squares) the model or to estimate $\theta \in \Theta$ as well as possible.

Problem 2: A nonlinear function, known as general nonlinear aspect, of the unknown parameter θ, $\varphi(\theta)$ say, is asked to be estimated as well as possible, even when the underlying model is assumed linear.

Robust estimation of a general nonlinear aspect $\varphi(\theta)$, as in Problem 2 above based on one-step-M-estimators with a bounded asymptotic bias was discussed by Kitsos and Muller (1995).

In both problems, interest is focused on the assumption about the errors. In this article it is assumed that e_i are iid from the normal distribution $N(0, \sigma^2)$.

In principle, $\sigma^2 = \sigma^2(x; \theta)$ in nonlinear situation and $\sigma^2 = \sigma^2(x)$ in linear.

The target is to discuss for the nonlinear design problem:

- When a design is optimal, i.e., possible optimality criteria that can be imposed, see Ford et al. (1989) as well as Pukelseim (1993), among others, for such approach.
- How we can fit the nonlinear model, see the early work of Bard (1974) and Seber and Wild (1989) for such approach.

On the Existence of the Least Square Estimators

After collection of the data the question arises as to whether it is possible to get estimates in all problems, that is those of binary response and regression.

For the model (2), we introduce the quantity

$$S_n(\theta) = \sum (y_i - f(u_i, \theta))^2 = \|y - f(u, \theta)\|_2 \tag{3}$$

where $\|.\|_2$ is the 1_2-norm. An estimate $\hat{\theta}$ will be called the *least squares estimate* (LSE) if

$$S_n(\hat{\theta}) = \min\{S_n(\theta);\ \theta \in \Theta\}. \tag{4}$$

Jennrich (1969), in his pioneering work, imposing some assumptions, proved that the model (2) has an LSE, $\hat{\theta}$, as a measurable function $\Psi \to \Theta$, where Ψ is the space of values of $Y's$. Under the usual normality assumption for the errors, it is known that this LSE coincides with the maximum likelihood estimators (MLE).

For the binary response problem the likelihood function L can be evaluated as

$$L \propto \Pi\{T(u_i, \theta)\}^{y_i}\{1 - T(u_i, \theta)\}^{1-y_i} \tag{5}$$

and maximum likelihood estimators can be obtained. Roughly speaking that occurs when the intersection of the sets of values taken by the explanatory variables corresponding to 1s and to 0s is not the null set. This happens to be a necessary and sufficient condition for the logit and probit models.

Now, having ensured that the likelihood equation can provide MLE and denoting by ℓ the log-likelihood we define the matrix

$$S(\hat{\theta}, \xi_n, y) = -\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\right) \tag{6}$$

where ξ_n is the design measure on n observations. The matrix

$$S(\hat{\theta}, \xi_n, y) = -\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} \quad \Big| \theta = \hat{\theta}\right) \tag{7}$$

will be called the "sample information matrix."

Example 3 Maximum likelihood estimates for the logistic can be obtained through the "normal equations":

$$\sum T_i = \sum y_i \text{ and } \sum u_i T_i = \sum y_i u_i$$

with $T_i = T(u_i; \theta)$ as in T_L model, Example 1 above.

The problem in nonlinear model fit is the construction of confidence intervals, as first Beale (1960) discussed, see also Seber and Wild (1989).

Linearization of the Model

The idea of the (design) matrix X being known is essential in dealing with linear models. In nonlinear models we can not define a matrix X in the same fashion. This can be done only approximately through the partial derivatives of θ, with θ taking its "true" value, θ_t. We define the nxp matrix

$$X = (x_{ij}) = \left.\frac{\partial f(u_i, \theta)}{\partial \theta_j}\right|_{\theta=\theta_t}. \tag{8}$$

Then the matrix $X = X(\theta)$ is formed as a function of θ. Function $f(u, \theta)$ can be linearized through a Taylor series expansion in the neighborhood of θ_t (the *true* parameter) as

$$f(u, \theta) = f(u, \theta_t) + \sum(\theta_j - \theta_{tj})(\partial f(u, \theta))/(\partial \theta_j)\big|_{\theta=\theta_t}. \tag{9}$$

Following the pattern of ▶linear regression models in the nonlinear regression case (See ▶Nonlinear Regression), an approximation to the covariance matrix, of the estimates of the parameters, can be defined as

$$C \cong [X^T(\theta_t)X(\theta_t)]^{-1}\sigma^2. \tag{10}$$

Moreover, for all nonlinear problems a useful approximation to the covariance matrix is

$$C^{-1} \cong nM(\theta_t, \varepsilon). \tag{11}$$

With $M(.,.)$ being the "average per observation information matrix."

The index t declares that the parameter takes its *true* value, which of course is asked to be estimated! Here is exactly the difficulty in nonlinear problems. For this one could seek minimum bias experiments. But the linearization of the model, not only creates problems in confidence intervals (as we have now "banana shape" intervals), but as well as on fitting the model, depending on the curvature of the model.

The linearization idea can be applied to the logit model in the following example.

Example 4 Given that $[1 + \exp(-\theta_1(u - \theta_2))]^{-1} \cong 1/2 + 1/6\,\theta_1(u - \theta_2)$, when $|\theta_1(u - \theta_2)| \leq 3$, then the normal equations of Example 3 are approximately

$$n/2 + (\theta_1/6)\sum(u_i - \theta_2) = \sum y_i$$
$$(1/2)\sum u_i + (\theta_1/6)\sum u_i(u_i - \theta_2) = \sum u_i y_i.$$

Optimality Criteria

For both models (1) and (2) we shall denote by $\eta = E(Y)$. Then it is

$$\eta = E(Y) = \begin{cases} g(x, \theta) & \text{models (2.2)} \\ T(x, \theta) & \text{models (2.1).} \end{cases} \tag{12}$$

Let $\nabla\eta$ denote the vector of partial derivatives of η with respect to $\theta \in \Theta \subseteq \mathcal{R}^p$. Then for the exponential family of models Fisher's information matrix is defined to be

$$I(\theta, x) = \sigma^{-2}(\nabla\eta)(\nabla\eta)^T. \tag{13}$$

The concept of the average-per-observation information matrix plays an important role to the nonlinear problems scenario for the definition of the optimality criteria, as well as for the fit of the model (when defining the appropriate approximate confidence intervals). It is defined for ξ, the design measure, Pukelsheim (1993) to be

$$M(\theta, \xi) = \begin{cases} n^{-1}\Sigma I(\theta, x_i), & \text{discrete case} \\ \int_X I(\theta, x)\xi(dx), & \text{continuous case.} \end{cases} \tag{14}$$

On the basis of the experiment the average-per-observation information matrix $M = M(\theta, \xi)$ is obtained which depends, in principle, on θ.

As the nonlinear experimental design problem suffers "on θ dependence", i.e., $M = M(\theta, \xi)$, there is not a unique theoretical framework for every model (1) or (2).

Example 5 For the logit or probit model the $D(\theta)$-optimal design concentrated at two points, i.e., $\xi_1 = \xi_2 = 1/2$ and the optimal design points are

$$x_1 = (x_0 - \theta_1)/\theta_2, \quad x_2 = (-x_0 - \theta_1)/\theta_2.$$

The D-optimal points corresponds to $x_0 = 1.54$ for the logistic and $x_0 = 1.14$ for the probit. If the design space $X = [\alpha, \beta] \subseteq \mathcal{R}$ is symmetric about $-\theta_1/\theta_2$ and $(\pm x_0 - \theta_1)/\theta_2 \notin X$ then $x_1 = \alpha, x_2 = \beta$.

Example 6 In chemical kinetics different models have been developed to describe a chemical process. To develop the sequential design the initial local optimum design is needed, for the particular model. Therefore, Kitsos (1995) provides all the appropriate support points. For the family of generalized linear models the support points have been provided by Sitter and Torsney (1995).

About the Author

For biography *see* the entry ►Calibration.

Cross References

►Least Squares
►Likelihood
►Nonlinear Mixed Effects Models
►Nonlinear Regression
►Nonlinear Time Series Analysis
►Optimum Experimental Design
►Regression Models with Symmetrical Errors

References and Further Reading

Bard V (1974) Nonlinear parameter estimation. Academic, New York

Bates DM, Watts DG (1981) Parameter transformations for improved approximate confidence regions in nonlinear least squares. Ann Stat 9:1152–1167

Beale EML (1960) Confidence regions in non-linear estimation. J R Stat Soc B 22:41–88

Ford I, Kitsos CP, Titterington DM (1989) Recent advances in nonlinear experiment design. Technometrics 13:49–60

Jennrich RJ (1969) Asymptotic properties of non-linear least squares estimators. Ann Math Stat 40:633–643

Kitsos CP, Titterington DM, Torsney B (1988) An optimal design problem in rhythmometry. Biometrics 44:657–671

Kitsos CP (1995) On the support points of *D*-optimal nonlinear experiment designs for chemical kinetics. In: Kitsos C, Muller W (eds) MODA4 (Model Oriented Data Analysis). Physica, Hidelberg, pp 71–76

Kitsos CP, Muller ChH (1995) Robust estimation of non-linear aspects. In: Kitsos C, Muller W (eds) MODA4 (Model Oriented Data Analysis), pp 71–76

McCullagh P, Nelder J (1989) Generalized linear models. Chapman & Hall, Boca Raton

Pukelsheim F (1993) Optimal design of experiments. Wiley, New York

Seber GAF, Wild CJ (1989) Nonlinear regression. Wiley, New York

Sitter RR, Torsney R (1995) *D*-optimal designs for generalized linear models. In: Kitsos C, Muller W (eds) MODA4 (Model Oriented Data Analysis). Physica, Hidelberg, pp 87–102

Nonlinear Regression

Andrej Pázman
Professor, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia

An important task of statistics is to find the relationship that exists between different variables. In *regression problems* typically one random variable Y, often called the *response variable*, is of particular interest. The other dependent variables $x_1, \ldots, x_k$, also called *explanatory variables or regressors*, are usually non-random and are principally used to predict or explain the behavior of Y. More exactly, the aim is to find the dependence of the mean $E_x(Y)$ on the vector $x = (x_1, \ldots, x_k)^T$ in a form

$$E_x(Y) = \eta(x, \theta),$$

where $\eta(.,.)$ is a "sufficiently smooth" known function, and $\theta = (\theta_1, \ldots, \theta_p)^T$ is the vector of *parameters*, which are to be estimated from the observed data. The remaining part of Y, $\varepsilon_x = Y - \eta(x, \theta)$, is the error variable, it cannot be observed directly, and does not depend on θ, neither stochastically. When $\eta(.,.)$ is linear on θ, the regression is called *linear*, if not, the *regression is nonlinear*. The data related to N observations are given by the *design of the experiment* $\left(x^{(1)}, \ldots, x^{(N)}\right)$, where each $x^{(i)} = \left(x_1^{(i)}, \ldots, x_k^{(i)}\right)^T$ is a choice of the values of the regressors, and by the vector of observations of the response variable, $y = (y_1, \ldots, y_N)^T$. The corresponding model of the experiment can be then written in a vector form

$$y = \eta(\theta) + \varepsilon, \tag{1}$$

where $\eta(\theta) = \left(\eta\left(x^{(1)}, \theta\right), \ldots, \eta\left(x^{(N)}, \theta\right)\right)^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)^T$. The possible values of θ are restricted by the assumption $\theta \in \Theta$ where Θ is the *parameter space*. We have $E(\varepsilon) = 0$ and assume $Var(\varepsilon) = \sigma^2 I$, where σ^2 is the unknown variance of Y. A more general set-up with $Var(\varepsilon) = \sigma^2 W$, with W known and not depending on θ, can be reduced to (1) by a linear transformation of y, hence implicitly we are considering it as well.

Typically many models in physics and chemistry and models for engineering problems are nonlinear regression models with parameters having a physical meaning. The form of the function $\eta(.,.)$ is then prescribed by a physical law, but ►nonlinear models are used also in cases when the model is aimed just for data fitting, and the nonlinear modeling requires the estimation of a smaller number of parameters than the fit of the data by a linear regression (cf. Ratkowsky 1983). One advantage of the nonlinear regression with least squares estimation is that a broad range of relationships can be fitted. One disadvantage is that the detection of ►outliers is much more difficult than in linear models. Sophisticated examples of nonlinear regression are obtained when $\eta(x, \theta)$ as a function of x is a solution of some differential equations, for example in compartmental models describing the exchange of some products in a chemical reaction, or the circulation of substances between different parts of a human body (cf. Seber and Wild 1989).

Estimation of θ. Typically one uses in a regression model the least squares estimator (LSE): $\hat{\theta} = \arg\min_{\theta\in\Theta} \|y - \eta(\theta)\|^2$. In linear models, with $\eta(\theta) = F\theta$, the computation is direct, namely $\hat{\theta} = \left(F^T F\right)^{-1} F^T y$, while in nonlinear models iterative methods are required. Different numerical methods of minimization are available, but the most used is the Gauss–Newton method (cf. Seber and Wild 1989), in which the improvement at the nth step is given by $\theta^{(n+1)} = \theta^{(n)} + \lambda^{(n)} v^{(n)}$ with $v^{(n)} = \left[M\left(\theta^{(n)}\right)\right]^{-1} F^T\left(\theta^{(n)}\right)\left(y - \eta\left(\theta^{(n)}\right)\right)$, where $M(\theta) = F^T(\theta) F(\theta)$, $F(\theta) = \frac{\partial \eta(\theta)}{\partial \theta^T}$, and with some $\lambda^{(n)} \in (0,1]$. Numerical problems can arise here as only local minima are computed, and a clever choice of the starting $\theta^{(1)}$ is necessary. Problems with an eventual ill-conditioned $M\left(\theta^{(n)}\right)$ are usually solved by a regularization called the Levenberg–Marquardt method.

Statistical inference based on the ▶asymptotic normality of $\hat{\theta}$. Under the assumption that i) Θ is compact, ii) $\eta\left(x^{(i)}, \theta\right)$ and its first and second order derivatives with respect to θ are continuous, iii) their values do not vary "too quickly" for an increasing $i \in \{1, 2, \ldots\}$, and iv) $\lim_{N\to\infty} \frac{1}{N} \lim \left\|\eta(\theta) - \eta\left(\bar{\theta}\right)\right\|^2 = 0$ implies $\theta = \bar{\theta}$, we find that the estimators $\hat{\theta}$ and $s^2 = \frac{1}{N-p}\left\|y - \eta\left(\hat{\theta}\right)\right\|^2$ converge with $N \to \infty$ strongly to $\bar{\theta}$, the true value of θ, and to σ^2 respectively (strong consistency). Moreover, if $\bar{\theta}$ is in the interior of Θ, and if $M^*\left(\bar{\theta}\right) = \lim_{N\to\infty} \frac{1}{N} F^T\left(\bar{\theta}\right) F\left(\bar{\theta}\right)$ is nonsingular, then $\sqrt{N}\left[\hat{\theta} - \bar{\theta}\right]$ converges in distribution to a random vector distributed normally with zero mean and variance $\sigma^2\left[M^*\left(\bar{\theta}\right)\right]^{-1}$. All this makes the nonlinear model asymptotically very similar to a linear model with normal errors, hence all standard methods of inference known from linear models are at hand (approximately, for large N): the estimator $\hat{\theta}$ is asymptotically unbiased, with $Var\left(\hat{\theta}\right) \doteq s^2\left[M\left(\hat{\theta}\right)\right]^{-1}$, and the approximate confidence ellipsoid for θ is

$$\left\{\theta \in \Theta : \left[\theta - \hat{\theta}\right]^T M\left(\hat{\theta}\right)\left[\theta - \hat{\theta}\right] < p s^2 \mathcal{F}_{p,N-p}\right\},$$

where $\mathcal{F}_{p,N-p}$ is a quantile of the F-distribution, etc. This is usually exploited in statistical packages for least squares estimation. Methods of *optimal design of experiments* (see ▶Optimum Experimental Design) also make much use of this similarity to a linear model. Higher order asymptotic results are at hand as well (cf. Gallant 1987), for example, in obtaining a ▶bias correction.

Inference problems for small-sample properties of $\hat{\theta}$. For a not too large N and under normal errors a better confidence region for θ than the confidence ellipsoid is the likelihood region

$$\left\{\theta \in \Theta : \|y - \eta(\theta)\|^2 - \left\|y - \eta\left(\hat{\theta}\right)\right\|^2 < p s^2 \mathcal{F}_{p,N-p}\right\}$$

(cf. Bates and Watts 1988 for examples), which is related to the asymptotic properties of the likelihood ratio test. Attempts to make further small sample corrections of these regions lead to the introduction of the *intrinsic and parameter measures of nonlinearity* (cf. Bates and Watts 1988), which are in fact geometric measures of curvatures either of the *expectation surface* of the model

$$\{\eta(\theta) : \theta \in \Theta\}$$

in the Euclidean geometry of the sample space, or of the parameter space in the Riemannian geometry, induced by the metric tensor given by $M(\theta)$ (the *Fisher information matrix for* $\sigma = 1$). Models with a large intrinsic measure of nonlinearity tend to give false or unstable LSE $\hat{\theta}$, while models with a large parameter measure of nonlinearity tend to give $\hat{\theta}$ having a large bias and, in general, a distribution of LSE which is far from a normal distribution. Explicit formulae for these measures are

$$\max_{v\in\mathbb{R}^p, v\neq 0} \sigma \frac{\left\|Z(\theta)\left[\sum_{i,j} v_i \frac{\partial^2 \eta(\theta)}{\partial\theta_i \partial\theta_j} v_i\right]\right\|}{v^T M(\theta) v}$$

with either $Z(\theta) = P(\theta) = F(\theta)\left[M(\theta)\right]^{-1} F^T(\theta)$ for the parametric nonlinearity, or $Z(\theta) = I - P(\theta)$ for the intrinsic nonlinearity. Computation is simplified using the QR decomposition of $F(\theta)$ (cf. Bates and Watts [1988] and Ratkowsky [1983] for an algorithm). In general, the geometry is very useful for understanding the nonlinear regression model (cf. Bates and Watts 1988; Pázman 1993), and extending the result to more general situations such as regression models with parameter constraints (Pázman 2002).

The influence of the parameter nonlinearity of the model is fully reflected by the probability distribution of $\hat{\theta}$, given $\bar{\theta}$ (cf. Pázman 1993), namely

$$\frac{\det\left[Q\left(\hat{\theta}, \bar{\theta}\right)\right]}{(2\pi)^{p/2}\sigma^p \det^{1/2}\left[M\left(\hat{\theta}\right)\right]} \exp\left\{-\frac{1}{2\sigma^2}\left\|P\left(\hat{\theta}\right)\left[\eta\left(\hat{\theta}\right) - \eta\left(\bar{\theta}\right)\right]\right\|^2\right\},$$

with $Q_{ij}\left(\hat{\theta}, \bar{\theta}\right) = M_{ij}\left(\hat{\theta}\right) + \left[\eta\left(\hat{\theta}\right) - \eta\left(\bar{\theta}\right)\right]^T\left[I - P\left(\hat{\theta}\right)\right] \frac{\partial^2 \eta(\theta)}{\partial\theta_i \partial\theta_j}$, which is correct under the assumption that the errors are normal, and that the intrinsic curvature is not so large to cause failure of the LSE.

Notice that some basic features of nonlinear regression models can be extended to more general models where the notion of the error variable is not used because the variance of Y depends on its mean. The best known are the ▶generalized linear models, where another parameter instead of $E_x(Y)$ is a linear function of the unknown parameters and the maximum likelihood estimator is used instead of the LSE.

About the Author

Professor Andrej Pázman is a leading personality in Slovak mathematical statistics. He is a member of the Learned Society of the Slovak Academy of Sciences, a member of Royal Statistical Society in UK, of the Institute of Mathematical Statistics in the USA, of the International Statistical Institute, and an honorary member of the Union of Slovak Mathematicians and Physicists. He was awarded the Gold Medal of the Faculty of Mathematics and Physics of Comenius University, both, the Silver and the Gold Juraj Hronec Medals of the Slovak Academy of Sciences for Contributions to Mathematical Sciences, and also the Premium of the Slovak Literary Found, and the WU Best Paper Award of the city of Vienna. (Adapted from the *Winter Workshop on Mathematical Statistics, Bratislava 2008, dedicated to Andrej Pázman's 70th Birthday*.)

Cross References

▶Box–Cox Transformation
▶Least Squares
▶Nonlinear Models
▶Optimum Experimental Design
▶Regression Models with Increasing Numbers of Unknown Parameters
▶Regression Models with Symmetrical Errors

References and Further Reading

Bates DM, Watts DG (1988) Nonlinear regression analysis and its applications. Wiley, New York
Gallant AR (1987) Nonlinear statistical models. Wiley, New York
Pázman A (1993) Nonlinear statistical models. Kluwer, Dordrecht
Pázman A (2002) Results on nonlinear least-squares estimators under nonlinear equality constraints. J Stat Plan Infer 103: 401–420
Ratkowsky DA (1983) Nonlinear regression modeling. Marcel Dekker, New York
Seber GAF, Wild CJ (1989) Nonlinear regression. Willey, New York (Paperback version (2003))

Nonlinear Time Series Analysis

Howell Tong
Professor Emeritus, Saw Swee Hock Professor of Statistics
London School of Economics and Political Science, London, UK
National University of Singapore, Singapore, Singapore

Introduction

A function f from R^p to R is said to be *linear* if for vectors $x, y \in R^p$ and any real scalar α, $f(\alpha x + y) = \alpha f(x) + f(y)$. Any function f that is not linear is said to be *nonlinear*.

In the analysis of stationary time series, the spectral density function, if it exists, is nonlinear under the above definition. However, for reasons to be made clear later, a statistical analysis that is based on it or its equivalents is ordinarily considered a linear analysis. Often, a time series is observed at discrete time intervals. For a discrete-time stationary time series $\{X_t : t = \ldots, -1, 0, 1, \ldots\}$ with finite variance, $corr(X_t, X_{t+s})$ is a function of s only, say $\rho(s)$, and is called the auto-correlatoin function. The spectral density function is the Fourier transform of $\rho(s)$ if $\sum_{s=-\infty}^{\infty} |\rho(s)| < \infty$. Now, Yule (1927) introduced the celebrated autoregressive model in time series. Typically the model takes the form

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \varepsilon_t, \qquad (1)$$

where the α's are parameters and $\{\varepsilon_t\}$ is a sequence of independent and identically distributed random variables with zero mean and finite variance, or a white noise for short. It is commonly denoted as an $AR(p)$ model. Clearly X_t is a linear function of $X_{t-1}, \ldots, X_{t-p}, \varepsilon_t$. Under the assumption of normality, the distribution of the time series is completely specified by its constant mean, constant variance and $\rho(s)$'s. Perhaps for the close connection with the analysis of linear models (of which the autoregressive model is one), an analysis based on the autocorrelation function or equivalently the spectral density function is loosely referred to as a linear analysis of the time series. By the same token, an analysis based on higher order moments or their Fourier transforms is loosely called a nonlinear analysis. Broadly speaking, tools based on the Fourier transforms of moments constitute what is called the frequency-domain approach, while those based on the moments constitute the time-domain approach, which often includes building a time series model of the form (1) or its generalizations.

Similar discussion as the above can be extended to cover $\{X(t) : t \in R\}$

Can We Do Without Nonlinearity?

A general answer is in the negative simply because the dynamical laws governing Nature or human activities are seldom linear. In the real world, we can see the footprints of nonlinearity everywhere we look. Below are a few examples.

(a) *Phase Transition*
The melting of ice of a glacier will alter fundamentally the amount of water flowing in a river near the glacier. Phase transition (from solid to liquid in the above example) is an important signature of nonlinearity. Animals behave differently (e.g., hunting effort) during time of short food supply versus time of abundant food supply.

(b) *Saturation*
In economics, diminishing return is a well-known phenomenon: doubling your effort does not necessarily double your reward.

(c) *Synchronization*
The celebrated Dutch scientist, Christiaan Huygens, observed that clocks placed on the same piece of soft timber were synchronized! Biological systems can also exhibit synchoronization. It has been noted that girls sharing the same dormitory have a higher chance of synchronizing their menstruation. Even female keepers of baboons have been known to have a similar experience.

(d) *Chaos*
When we toss a coin to randomize our choice, we are exploiting nonlinearity, for the dynamical system underlying the tossing is a set of (typically three) nonlinear ordinary differential equations, the solution of which is generally very sensitive to the initial spinning unless we "cheat." The system is said to generate chaos in a technical sense. When statisticians generate pseudo-random numbers, they are also generating chaos. One of the most commonly used pseudo-random generator is the linear congruential generator, which is a piecewise linear (i.e., nonlinear) function that does precisely this. It might surprise you that you are actually using nonlinear devices almost daily because encrypting passwords is closely related to pseudo-random number generation.

In the following sections, we focus on the time-domain approach because at the current state of development, this approach tends to admit simpler interpretations in practical applications.

What Is a Nonlinear Time Series Model?

A short answer is that it is not a linear time series model. This raises the need to define a linear model. A fairly commonly adopted definition is as follows. A stationary time series model is called a linear time series model if it is equivalent (for example in the mean-square sense) to

$$X_t = \sum_{s=-\infty}^{\infty} \beta_s \varepsilon_{t-s}, \tag{2}$$

where $\{\varepsilon_t\}$ is a white noise and the summation is assumed to exist in some sense. An alternative definition due to Hannan (1973) is one that requires that the minimizer of $E|X_t - h(X_{t-1}, X_{t-2}, \ldots)|^2$ with respect to h over the space of all measurable functions is the linear function. Here the mean square is assumed to exist.

Are Linear Time Series Models Fit for Purpose?

Examples abound of the inability of linear time series models to capture essential features of the underlying dynamics.

Yule (1927) introduced the autoregressive model to model the annual sunspot numbers with a view to capturing the observed 11-year sunspot cycle but noted the inadequacy of his model. He noted the asymmetry of the cycle and attempted to model it with an $AR(4)$ model only to discover that it gave statistically a worse fit than a simpler AR(2) model.

Moran (1953) fitted an $AR(2)$ model to the annual lynx data corresponding to the MacKenzie River region in Canada, with a view to capturing the observed 10-year cycle. He was quick to point out that the fitted residuals were heteroscedastic.

Whittle (1954) analyzed a seiche record from Wellington Bay in New Zealand. He noted that, besides the fundamental frequency of oscillations and a frequency due to the reflection of an island at the bay, there were sub-harmonics bearing an interesting arithmetic relation with the above frequencies. Now, sub-harmonics are one of the signatures of nonlinear oscillations, long known to the physicists and engineers.

Examples of Nonlinear Time Series Models

First, we describe parametric models. Due to space limitation, we describe the two most commonly used models. For other models, we refer to Tong (1990). We shall describe (i) the threshold model and (ii) the (generalized) autoregressive conditional heteroscedasticity model, or in short the TAR model and the (G)ARCH model respectively. The former was introduced by Tong in 1977 and developed systematically in Tong and Lim (1980) and Tong

(1983, 1990), and the latter by Engles (1982), later generalized by Bollerslev (1986).

There are several different but equivalent ways to express a TAR model. Here is a simple form. Let $\{Z_t\}$ denote an indicator time series that takes positive integer values, say $\{1, 2, \ldots, K\}$. Let $\{\eta_t\}$ denote a white noise with zero mean and unit variance, $\alpha_0^{(j)}, \alpha_i^{(j)}, \beta^{(j)}$ be real constants for $j = 1, 2, \ldots, K$. Then a time series $\{X_t : t = 0, \pm 1, \pm 2, \ldots\}$ is said to follow a *threshold autoregressive model* if it satisfies, when $Z_t = j$, $j = 1, \ldots, K$,

$$X_t = \alpha_0^{(j)} + \sum_{i=1}^{p} \alpha_i^{(j)} X_{t-i} + \beta^{(j)} \eta_t. \tag{3}$$

For the case in which $Z_t = j$ if and only if $X_{t-d} \in R_j$ for some positive integer d and for some partition of R, i.e., $R = \bigcup_{i=1}^{K} R_i$ say, the TAR model is called a *self-exciting threshold autoregressive model*, or SETAR model for short. In this case, given $X_{t-s}, s > 0$, the conditional mean of X_t is piecewise linear, and the conditional variance of X_t piecewise constant.

For the case in which $Z_t = j$ if and only if $Y_{t-d} \in R_j$ for some covariate time series $\{Y_t\}$, some positive integer d and some partition of R, i.e. $R = \cup_{i=1}^{K} R_i$ say, then we have a TAR model driven by (or excited by) $\{Y_t\}$. Note that the covariate time series $\{Y_t\}$, and thus the indicator time series $\{Z_t\}$, can be observable or hidden. If the indicator time series, whether observable or hidden, forms a Markov chain (see ▶Markov Chains), then we call $\{X_t\}$ a Markov-chain driven TAR; this model was first introduced by Tong (Tong and Lim 1980, p. 285; Tong 1982; p. 62). In the econometric literature, the sub-class with a hidden Markov chain is commonly called a *Markov switching model*.

The TAR model, especially the SETAR model, has many practical applications in diverse areas/disciplines, including earth sciences, ecology, economics, engineering, environmental science, finance, hydraulics, medical science, water resources and many others.

The nonlinear parametric model that is mostly and widely used in econometrics and finance is the (G)ARCH model. The ARCH model is given by

$$X_t = \eta_t \sigma_t, \tag{4}$$

where $\{\eta_t\}$ is as defined previously but sometimes assumed to be Gaussian, and $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2$, $\alpha_0 > 0, \alpha_i \geq 0, i = 1, \ldots, p$. Note that the ARCH model differs from the SETAR model in its σ_t being a continuous function instead of a piecewise constant function as in the latter. The GARCH model generalizes σ_t^2 to $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2$, where the β_is are usually also assumed to be non-negative, although the non-negativity assumption may be relaxed; see Cryer and Chan (2008, Chap. 12).

One of the limitations of any parametric modelling approach is the subjectivity of selecting a family of possible parametric models. We can sometimes mitigate the situation if a certain parametric family is suggested by subject matter considerations. In the absence of the above, mitigation is weaker even if we are assured that the family is dense in some sufficiently large space of models. It is then tempting to allow the data to suggest the form of F where we are contemplating a model of say

$$X_t = F(X_{t-1}, \ldots, X_{t-p}, \varepsilon_t), \tag{5}$$

F being unknown. This is one of the strengths of the nonparametric modelling approach, which is a vast and rapidly expanding area. A word of caution is the so-called curse of dimensionality, meaning that when $p > 3$ the estimated F is unlikely to be reliable unless we have a huge sample size. One way to ameliorate the situation is to replace $X_{t-1}, \ldots, X_{t-p}$ by $\xi_{t-1}, \ldots, \xi_{t-q}$ with q much smaller than p, e.g. $q = 1$ or 2. The ξ's are typically suitably chosen but unknown linear functions of X's, sometimes called indices. This is called the semi-parametric modelling approach, which is also a rapidly expanding field. For comprehensive accounts of the above developments, see, e.g., Fan and Yao (2005) and Gao (2007). Another way is to impose some simplifying structure on (5) such as zero interaction as in Chen and Tsay (1993), who gave

$$X_t = F(X_{t-1}) + \cdots + F(X_{t-p}) + \varepsilon_t. \tag{6}$$

About the Author

Past President of the Hong Kong Statistical Society (1983–1984), Professor Howell Tong, was Founding Chair of Statistics, Chinese University of Hong Kong (1982–1985), Chair of Statistics at the University of Kent at Canterbury, (1986–1998), and Chair of Statistics at the London School of Economics (1999–September 2009). Between 1997 and 2004, he was also Chair of Statistics and Pro-Vice Chancellor and Founding Dean of the Graduate School, University of Hong Kong. He was elected a Fellow of the Institute of Mathematical Statistics (1993) and is a Foreign member of the Norwegian Academy of Science and Letters (2000). Among many awards, Professor Tong won the State Natural Science Prize, China, in 2000 and received the Guy medal in Silver (Royal Statistical Society, UK) in 2007, in recognition of his many important contributions to time series analysis.

Cross References

►Box–Jenkins Time Series Models
►Econometrics
►Heteroscedastic Time Series
►Statistical Modeling of Financial Markets
►Time Series

References and Further Reading

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. J Economet 31:307–327
Chen R, Tsay RS (1993) Nonlinear additive ARX models. J Am Stat Assoc 88:955–967
Cryer JD, Chan KS (2008) Time series analysis: with applications in R. Springer, New York
Engles R (1982) Autoregressive conditional heteroscedasticity with estimates of variance of United Kingdom inflation. Econometrica 50:987–1008
Fan J, Yao Q (2005) Nonlinear time series: nonparametric and parametric methods. Springer, New York
Gao J (2007) Nonlinear time series: semiparametric and ponparametric methods. Chapman and Hall/CRC Press, London
Hannan EJ (1973) The asymptotic theory of linear time-series models. J Appl Probab 10:130–145
Moran PAP (1953) The statistical analysis of the Canadian lynx cycle. Aust J Zool 1:163–173
Tong H (1983) Threshold models in nonlinear time series analysis. Springer Lecture Notes in Statistics, New York
Tong H (1990) Non-linear time series: a dynamical system approach. Oxford University Press, Oxford
Tong H, Lim KS (1980) Threshold autoregression, limit cycles and cyclical data. J R Stat Soc Ser B 42:245–292
Whittle P (1954) The statistical analysis of a seiche record. Sears Foundation J Marine Res 13:76–100
Yule U (1927) On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. Phil Trans R Soc Lond Ser A 226:267–298

Nonparametric Density Estimation

Ricardo Cao
Professor
Universidade da Coruña, A Coruña, Spain

The density function of a continuous random variable, X, is the derivative of its distribution function $f(x) = dF(x)/dx$, and can be represented as

$$f(x) = \lim_{h\to 0} \frac{F(x+h) - F(x-h)}{2h} = \lim_{h\to 0} \frac{1}{2h} \int_{x-h}^{x+h} dF(u) = \lim_{h\to 0} \frac{E\left[1\left(|X-x| \le h\right)\right]}{2h}. \quad (1)$$

Nonparametric density estimation can be performed via estimation of the last term in (1). Given a random sample $(X_1, \ldots, X_n)$, this ratio can be estimated replacing the unknown expectation by its empirical analog, suggesting the naive estimator

$$\hat{f}(x) = \frac{1}{2h} \int_{x-h}^{x+h} dF_n(u) = \frac{1}{n} \sum_{i=1}^{n} \frac{1\left(|X_i - x| \le h\right)}{2h}. \quad (2)$$

This estimator was proposed first by Rosenblatt (1956). It is the relative frequency, per unit of length, of the observations within the interval $[x-h, x+h]$. For a fixed sample size, it does not make sense to define the estimator by the limit, when $h \to 0$, of the previous quantity. In fact this limit is zero since, for h small enough, the numerator in (2) equals zero. However, as $n \to \infty$, one may think of the value h (often called smoothing parameter or bandwidth) as a sequence, h_n, tending to zero.

The naive estimator in (2) can be also written as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad (3)$$

with $K(u) = 1\left(|u| \le 1\right)/2$. The function K is called the kernel. In fact the kernel used in (2) is the uniform density function. Other kernel functions used in practice are the triangular density, the Gaussian or the Epanechnikov kernel. Kernel functions are generally required to hold that $\int_{\mathbb{R}} uK(u)\,du = 0$, $\int_{\mathbb{R}} K(u)\,du = 1$ and $\int_{\mathbb{R}} u^2 K(u)\,du < \infty$. This kernel estimator was proposed by Rosenblatt (1956) and Parzen (1962).

Using the structure of (3) as a sum of iid random variables, its bias and variance can be easily computed. In general, the estimator is biased and, for small values of h, the bias and the variance are approximately $E\left(\hat{f}(x) - f(x)\right) \simeq \frac{h^2}{2} f''(x) \int t^2 K(t)\,dt$ and $Var\left(\hat{f}(x)\right) \simeq \frac{f(x)}{nh} \int K(t)^2\,dt$. Thus, it seems reasonable to require, for consistency, that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. The behavior of these two terms is opposite: while the bias increases with the smoothing parameter, the variance decreases as the bandwidth gets large. Hence, it is intuitive that the choice of the smoothing parameter is very important in practice, since it regulates the balance between the bias (systematic error) and the variance (stochastic error) of the estimator. Undoubtedly any practical choice for h has to be a compromise between both terms.

A well-known nonparametric density estimator is the histogram. It is a predecessor of the kernel estimator in (3). The nearest neighbours method consists in using the distance between a given point and its kth nearest sample value to compute a nonparametric density estimator. This method adapts the "amount of smoothing" to the point where the density is estimated. The spline method can be

thought as estimating some piecewise polynomial approximation of the underlying density. The method of orthogonal series can be viewed as estimating some approximation of the true density function, namely, a finite linear combination of terms of a basis of the functional space where the density is assumed to belong. The list of books devoted to nonparametric density estimation includes Silverman (1986), Devroye and Györfi (1985) and Wand and Jones (1995).

An important issue in nonparametric density estimation is the problem of selection of the smoothing parameter. The most popular error measures for bandwidth selection are the integrated squared error criterion $ISE = \int \left(\hat{f}_h(x) - f(x)\right)^2 dx$, and the mean integrated squared error, given by $MISE = E(ISE)$.

Under regularity conditions on f, an asymptotic representation for $MISE$ can be obtained: $MISE(h) = AMISE(h) + o(h^4) + O(n^{-1})$, where

$$AMISE(h) = \frac{h^4}{4} \int f''(x)^2 dx \int t^2 K(t) dt + \frac{1}{nh} \int K(t)^2 dt.$$

Minimization of $AMISE(h)$ in h gives the asymptotically optimal bandwidth

$$h_{AMISE} = \left(\frac{\int K(t)^2 dt}{n \int f''(x)^2 dx \int t^2 K(t) dt} \right)^{1/5}.$$

However, the value h_{AMISE} is not observable since it depends on the curvature of the underlying density. In practice, most of the bandwidth selectors are defined as minimizers of the error measures *ISE*, *MISE* or *AMISE*.

Among the dozens of proposals for bandwidth selection in this context we mention the least squares cross validation, proposed by Rudemo (1982) and Bowman (1984), the biased cross-validation method, (see Scott and Terrell 1987), the plug-in methods (see Sheather and Jones 1991), the smoothed cross-validation, proposed by Hall et al. (1992), and the bootstrap bandwidth selectors (see Cao 1993).

Nonparametric density estimation methods can be also extended to multivariate settings (see the book by Scott 1992). Similar ideas have been successfully used for the estimation of other curves as the regression function, the distribution function, the conditional distribution and density and the hazard rate, among many other.

About the Author

Ricardo Cao is Professor of Statistics at Universidade da Coruña, Spain. He is President of ECAS (European Courses in Advanced Statistics), Editor of *Test* (The Official Journal of Statistics of the Spanish Society for Statistics and Operations Research), Associate Editor of *Computational Statistics* and *Journal of Nonparametric Statistics*. He is also an elected member of the International Statistical Institute. Professor Cao is President of the Mathematics Committee of ANEP (Spanish Agency for Research Evaluation).

Cross References

►Nonparametric Estimation
►Parametric and Nonparametric Reliability Analysis
►Smoothing Techniques

References and Further Reading

Bowman AW (1984) An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71:353–360
Cao R (1993) Bootstrapping the mean integrated squared error. J Multivariate Anal 45:137–160
Devroye L, Györfi L (1985) Nonparametric density estimation: the L_1-view. Wiley, New York
Hall P, Marron JS, Park B (1992) Smoothed cross-validation. Probab Theory Related Fields 92:1–20
Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33:1065–1076
Rosenblatt M (1956) Remarks on some nonparametric estimators of a density function. Ann Math Stat 27:832–837
Rudemo M (1982) Empirical choice of histograms and kernel density estimation. Scand J Stat 9:65–78
Scott DW (1992) Multivariate density estimation: theory, practice and visualization. Wiley, New York
Scott DW, Terrell GR (1987) Biased and unbiased cross-validation in density estimation. J Am Stat Assoc 82:1131–1146
Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc Ser B 53:683–690
Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall, New York
Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall, London

Nonparametric Estimation

Maarten Jansen[1], Gerda Claeskens[2]
[1]Professor
Université libre de Bruxelles, Brussels, Belgium
[2]Professor
K.U. Leuven, Leuven, Belgium

Meanings of the Word "Nonparametric"

The terminology *nonparametric* was introduced by Wolfowitz in 1942 to encompass a group of statistical techniques for situations where one does not specify the

functional form of the distributions of the random variables that one is dealing with. In its earlier form, this comprised mainly methods working with rank statistics, and was also coined "*distribution free*" methods. Most often these methods are applied to perform hypothesis tests. For an example of such a hypothesis test, see the entry by Jurečková (same volume). Other examples include the ►Kolmogorov–Smirnov test, the runs test, ►sign test, ►Wilcoxon-signed-rank test, the Mann–Whitney U-test (see ►Wilcoxon–Mann–Whitney Test) and ►Fisher Exact Test. For an overview and details, see Hollander and Wolfe (1999). This type of nonparametric method has the advantage that it can be applied to ordinal and rank data; the data may be frequencies or counts, and do not have to be measured on a continuous scale.

In more recent times, nonparametric statistics has evolved to settings where a model for the data is not specified a priori, but is in some form determined from the data. This will be explained in more detail below. In such nonparametric models there are parameters to estimate, even many parameters in most cases, hence the name-giving might be somewhat misleading. Main examples of such estimation methods are kernel estimators, splines and wavelet estimators. These techniques are also known as "*smoothing methods*."

Nonparametric Regression Estimation

In parametric regression models we relate the mean of a response variable Y, conditional on covariates X via a parametric function. For example, in ►linear regression models, we assume that $Y = \beta_0 + X\beta_1 + \varepsilon$, where β_0 and β_1 are unknown parameter vectors and ε is often assumed to be a normal random variable with zero mean and an unknown variance σ^2. In nonparametric regression we do not specify the functional form for the conditional mean of Y and write the model as $Y = f(X) + \varepsilon$, where X may be random, or take fixed values. The terminology smoothing arises from the commonly made assumption for most methods (however, see the "►Wavelet Estimation" section below) that the unspecified function f is smooth.

Nonparametric estimation starts with choosing a basis which defines a space of functions. The function f is then approximated within this space by $\tilde{f}(x) = \sum_{j=1}^{J} \beta_j \psi_j(x)$. The basis functions may also depend on further parameters, specifying for example the location. These may be estimated or specified in advance. Fourier series are one example. Nonparametric estimation of f then proceeds with estimating the unknown parameters. Spaces of functions are often infinite dimensional, hence the number of basis functions to be used, J in the above sum, is a tuning parameter. The more basis functions taken, the better the approximation will be, in general. However, estimating more parameters comes at a cost of increased variance and increased computational effort.

The smoothing methods are used in a similar way for the estimation of density functions. While histograms give rough approximations, the nonparametric density estimators are smooth curves. Likewise, splines, wavelets or kernels may be used. For the latter method, see for example Wand and Jones (1995).

Spline Estimation

The choice of the basis characterizes the estimated function. Often taken choices are *spline* functions. A jth degree polynomial spline is a curve that consists of piecewise jth degree polynomial parts that are continuously joined together at *knots*. The smoothness of the resulting function depends on whether also the higher derivatives of the spline are continuous. When each observation x_i, $i = 1, \ldots, n$ is taken as a knot, this results in a *smoothing spline* (see ►Smoothing Splines). When a set of knots $\kappa_1, \ldots, \kappa_K$ is chosen, with $K < n$, the sample size, the function is a *regression spline*. In cases that $K < n$, estimation of the unknown spline coefficients β_j can be done via ►least squares in case of (approximate) normal errors ε. For smoothing splines, one introduces a penalty term that is related to the derivatives of f. Also for regression splines, penalties on the coefficients may be stated, to reduce the influence of the choice of the knots. This results in *penalized regression splines*. An expanded description of spline regression methods can be found in the entry by Opsomer and Breidt (same volume). Some main references are Eubank (1988), Wahba (1990), Green and Silverman (1994) and Ruppert et al. (2003).

Wavelet Estimation

Thanks to a fast decomposition algorithm (Mallat 1989), wavelet bases have gained considerable success as a representation for data to be smoothed. Wavelet basis functions are short waveforms located at a specific point in time or space and with a specific scale. This locality in time and frequency provides a tool for a multiscale and sparse representation of data. Especially piecewise smooth data, with isolated singularities, are typical objects for which wavelets are well suited. Indeed, the singularities can be captured by a relatively limited number of local wavelet basis functions, with appropriate scales, while the smooth intervals in between the singularities produce many but small contributions in a wavelet decomposition.

While other methods may have difficulties in catching singularities, in a wavelet decomposition they pose no bottleneck, provided that at the position of a singularity the wavelet representation is locally more refined than in

between the singularities. The location of the singularities is done automatically, even in the presence of noise, by the fact that the coefficients corresponding to the basis functions at those positions are large, as they carry the contributions that constitute the singularity. Singularities are thus well captured by selecting the largest coefficients, rather than a predetermined subset. Putting the smallest coefficients to zero therefore removes most of the noise without affecting the noise-free data too much. The usage of this thresholding or any sophisticated variant, which is always a nonlinear processing, is always the main reason for using a wavelet decomposition and it is always linked to the intermittend nature of the data, i.e., the presence of isolated singularities in otherwise smooth behavior. The use of thresholds relies on the sparsity property of a wavelet resprepresentation. The multiscale property, on the other hand, is mostly used for additional across-scale processing, for instance to remove false positives after thresholding (for smoother intervals between singularities) or to correct for false negatives by looking across scales (for sharper reconstruction of singularities). Also scale dependent processing is necessary in the case of correlated noise on the observations (Donoho and Johnstone 1995).

The selection of appropriate thresholds has been a major domain of research. Limiting or even reducing to zero the number of false positives is the objective of an important class of thresholds, including the universal threshold (Donoho and Johnstone 1994) or False Discovery Rate thresholds (Benjamini and Hochberg 1995). Another class of thresholds focusses on the expected, integrated squared loss, i.e., risk, of the result. Stein's Unbiased Risk Estimator and modifications (such as cross-validation) provide practical methods for finding minimum risk thresholds. A third, and wide class of threshold assessment methods is based on Bayesian – mostly empirical Bayes – models, such as EBayesthresh (Johnstone and Silverman 2004, 2005). The prior model for noise-free coefficients reflects the idea of sparsity, mostly through a zero-inflated or otherwise mixture model with heavy tails (where heavy here includes everything heavier than the normal distribution).

Kernel and Local Polynomial Estimation

Kernel estimation of a regression function starts from the idea that the function is locally well approximated by a low order polynomial curve. The Nadaraya–Watson estimator locally approximates the curve f at value x by a constant regression function. Observations X_i close to x get a large weight, and observations further away receive less or zero weight. The kernel function K determines the weighting and is assumed to be a density function. The estimator takes the following form, $\widehat{f}_h(x) = \sum_{i=1}^n K_h(x - X_i) Y_i \big/ \sum_{i=1}^n K_h(x - X_i)$, where h is called the bandwidth. This is a tuning parameter, small values of h imply that only close neighbors get a large weight, this might result in a rather wiggly fit. Large values of h will result in much smoother fitted curves. Several studies have focussed on appropriate bandwidth choices, for example via cross-validation or plug-in methods based on asymptotic properties of the estimator. Variants on this estimator are the Priestley–Chao and Gasser–Müller estimator. Local polynomial estimators are similar in spirit. Instead of taking a local constant approximation of the function f around x, a local polynomial approximation is obtained. More information on kernel regression methods can be found in the entry by Opsomer and Breidt (same volume). For more details, see Fan and Gijbels (1996).

About the Authors

Maarten Jansen is Assistant Professor at the Departments of Mathematics and Computer Science of the Université libre de Bruxelles (Belgium). He is Elected member of the International Statistical Institute and author of two books: *Second generation wavelets and applications* (with P. Oonincx, Springer Verlag, 2005) and *Noise reduction by wavelet thresholding* (Springer Verlag, 2001), and about 20 journal papers. Currently he is Associate editor of *Signal Processing*.

For biography of Gerda Claeskens *see* the entry ►Model Selection.

Cross References

►Bootstrap Methods
►Estimation
►Kaplan-Meier Estimator
►Measurement Error Models
►Modeling Survival Data
►Nonparametric Density Estimation
►Nonparametric Estimation Based on Incomplete Observations
►Nonparametric Regression Using Kernel and Spline Methods
►Nonparametric Statistical Inference
►Smoothing Splines
►Smoothing Techniques

References and Further Reading

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300

Eubank R (1988) Spline smoothing and nonparametric regression. Marcel Dekker, New York

Donoho D, Johnstone I (1994) Ideal spatial adaptation via wavelet shrinkage. Biometrika 81(3):425–455
Donoho D, Johnstone I (1995) Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90(432):1200–1224
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman & Hall, London
Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman & Hall, London
Hollander M, Wolfe DA (1999) Nonparametric statistical methods, 2nd edn. Wiley, New York
Johnstone I, Silverman B (2004) Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. Ann Stat 32(4):1594–1649
Johnstone I, Silverman B (2005) Empirical Bayes selection of wavelet thresholds. Ann Stat 33(4):1700–1752
Mallat S (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 11(7):674–693
Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge
Wahba G (1990) Spline models for observational data. SIAM, Philadelphia
Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall, London
Wolfowitz J (1942) Additive partition functions and a class of statistical hypotheses. Ann Stat 13:247–279

Nonparametric Estimation Based on Incomplete Observations

Abdurahim Abdushukurov
Professor and Head, Vice Rector
National University of Uzbekistan, Tashkent, Uzbekistan

Incomplete observations occur in survival analysis, especially in clinical trials and engineering when we partially observe death in biological organisms or failure in mechanical systems.

From statistical literature one can learn that incomplete observations arise in two ways: by censoring and truncation. Note that truncation is sampling an incomplete population, while censoring occurs when we are able to sample the complete population, but the individual values of observations below and/or above given values are not specified. Therefore, censoring should not be confused with truncation. In this article we deal only with right-censoring model, which is easily described from the methodological point of view.

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables (i.i.d.r.v.-s) (the lifetimes) with common distribution function (d.f.) F. Let X_j be censored on the right by Y_j, so that observations available for us at the nth stage consist of the sample $S^{(n)} = \{(Z_j, \delta_j), 1 \le j \le n\}$, where $Z_j = min(X_j, Y_j)$ and $\delta_j = I(X_j \le Y_j)$ with $I(A)$ meaning the indicator of the event A. Suppose that Y_j are again i.i.d.r.v.-s, the so-called censoring times with common d.f. G, independent of lifetimes X_j.

The main problem consists of nonparametrically estimating F with nuisance G based on censored sample $S^{(n)}$, where r.v.-s of interest X_j-s are observed only when δ_j=1. Kaplan and Meier (1958) were the first to suggest the product-limit (PL) estimator F_n^{PL} defined as

$$F_n^{PL}(t) = \begin{cases} 1 - \prod_{\{j: Z_{(j)} \le t\}} \left[1 - \frac{\delta_{(j)}}{n-j+1}\right], & t \le Z_{(n)}, \\ 1, & t > Z_{(n)}, \delta_{(n)} = 1, \\ \textit{undefined}, & t > Z_{(n)}, \delta_{(n)} = 0, \end{cases}$$

where $Z_{(1)} \le \ldots \le Z_{(n)}$ are the ▶order statistics of Z_j and $\delta_{(1)}, \ldots, \delta_{(n)}$ are the corresponding δ_j. In statistical literature there are different versions of this estimator. However, those do not coincide if the largest Z_j is a censoring time. Gill (1980) redefined the F_n^{PL} setting $F_n^{PL}(t) = F_n^{PL}(Z_{(n)})$ when $t > Z_{(n)}$. Further, we use Gill's modification of the PL-estimator. At present, there is an enormous literature on properties of the PL-estimator (see, e.g., Abdushukurov [2009], Akritas [2000], Csörgő [1996], Gill [1980, 1994]) and most of the work on estimating incomplete observations are concentrated on the PL-estimator. However, F_n^{PL} is not a unique estimator of F.

The second, closely related with the F_n^{PL}, nonparametrical estimator of F is the exponential hazard estimator

$$F_n^E(t) = 1 - exp\left\{-\sum_{j=1}^{n} \frac{\delta_{(j)} I(Z_{(j)} \le t)}{n-j+1}\right\}, -\infty < t < \infty.$$

F_n^E plays an important role in investigating the limiting properties of the estimator F_n^{PL}.

Abdushukurov (1998, 1999) proposed another estimator for F of power type:

$$F_n(t) = 1 - (1 - H_n(t))^{R_n(t)} = \begin{cases} 0, & t < Z_{(1)}, \\ 1 - \left(\frac{n-j}{n}\right)^{R_n(t)}, & Z_{(j)} \le t < Z_{(j+1)}, 1 \le j \le n-1, \\ 1, & t \ge Z_{(n)}, \end{cases}$$

where

$$H_n(t) = \frac{1}{n} \sum_{j=1}^{n} I(Z_j \le t)$$

is an empirical estimator of d.f. $H(t) = P(Z_j \le t) = 1 - (1 - F(t))(1 - G(t))$ and

$$R_n(t) = \frac{-log\left(1 - F_n^E(t)\right)}{\sum_{j=1}^{n} \frac{I(Z_{(j)} \le t)}{n-j+1}}$$

As we see, estimator F_n is defined on whole line. Let

$$a_n(t) = \sum_{j=1}^{n} \frac{I(Z_{(j)} \le t)}{(n-j)(n-j+1)}.$$

Note that $\sup\{a_n(t), t \le T\} \le [n(1 - H_n(T))]^{-1} = \mathbb{O}\left(\frac{1}{n}\right)$ with probability 1, where $T < Z_{(n)}$.

The following inequalities show that all three estimators are closely related (Abdushukurov (1998, 2009)): *For $t < Z_{(n)}$ with probability 1*

$(I)\ 0 < -log\left(1 - F_n^{PL}(t)\right) + log\left(1 - F_n^E(t)\right) < a_n(t);$
$(II)\ 0 \le F_n^{PL}(t) - F_n^E(t) < \frac{1}{2} a_n(t);$
$(III) < -log(1 - F_n(t)) + log\left(1 - F_n^E(t)\right) < a_n(t);$
$(IV)\ \left|-log\left(1 - F_n^{PL}(t)\right) + log(1 - F_n(t))\right| < a_n(t);$
$(V)\ \left|F_n^{PL}(t) - F_n(t)\right| < a_n(t);$
$(VI)\ \left|F_n^E(t) - F_n(t)\right| < a_n(t).$

Thus, one can expect the stochastic equivalences of these estimators in the sense of their weak convergence to the same Gaussian process (Abdushukurov 1998). Let d.f.-s F and G be continuous and $T < T_H = \inf\{t : H(t) = 1\}$. Then one can define the sequence of Wiener processes $\{\mathbb{W}_n(x), 0 \le x < \infty\}_{n=1}^{\infty}$ such that when $n \to \infty$

$$\sup_{t \le T} \left| n^{\frac{1}{2}} \left(F_n^*(t) - F(t)\right) - (1 - F(t))\mathbb{W}_n(d(t))\right| \xrightarrow{P} 0,$$

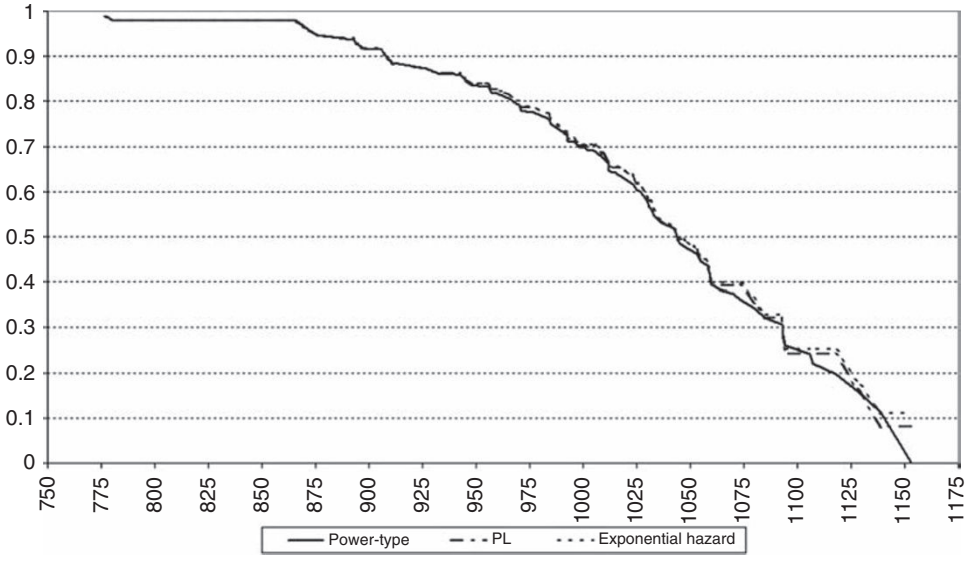
where F_n^* stands for one of estimators F_n^{PL}, F_n^E, F_n and

$$d(t) = \int_{-\infty}^{t} [(1-F)^2(1-G)]^{-1} dF.$$

Here we state the weak convergence result in the form of weak approximation by the sequence of appropriate copies of the limiting Gaussian process. This implies that $n^{\frac{1}{2}}\left(F_n^* - F\right)$ converges weakly in the Skorochod's space $\mathbb{D}(-\infty, T]$ to the mean-zero Gaussian process with covariance function $\sigma(t;s) = (1 - F(t))(1 - F(s))d(min(t,s))$, $t, s \le T$. Thus, we see that all three estimators are equivalent in the asymptotic sense. But as we see in Abdushukurov (1998, 2009) the estimator F_n has some peculiarities and even better properties than F_n^{PL} and F_n^E do for all $n \ge 1$. Let us consider the following exponential representation for any right continuous d.f. (Gill 1980):

$$1 - F(t) = exp\left\{-\int_{-\infty}^{t} \frac{dF(u)}{1 - F(u-)}\right\} \prod_{s \le t} (1 - \Delta\Lambda(s)),$$

where $\Delta\Lambda(s) = (F(s) - F(s-))/(1 - F(s-))$ and $F(s-) = lim_{u \uparrow s} F(u)$. Then we see that F_n^{PL} is a natural estimator for $\prod_{s \le t}(1 - \Delta\Lambda(s))$, that is a discrete d.f. On the other side, F_n^E and F_n are natural estimators for continuous d.f. $F(t) = 1 - exp\left\{-\int_{-\infty}^{t}(1-F)^{-1}dF\right\} = 1 - (1 - H(t))^{R(t)}$, where $R(t) = -log(1 - F(t))/[-log(1 - H(t))]$ - relative risk function. Obviously, the relative risk estimators $F_n(t)$ and $G_n(t) =$



Nonparametric Estimation Based on Incomplete Observations. Fig. 1 Plots of estimators $1 - F_n$, $1 - F_n^{PL}$ and $1 - F_n^E$ of survival function $1 - F$ using Channing House data

$1-(1-H_n(t))^{1-R_n(t)}$ of $F(t)$ and $G(t)$ satisfy the empirical analogy of equality $(1-F(t))(1-G(t)) = 1-H(t)$, $-\infty < t < \infty$, that is $(1-F_n(t))(1-G_n(t)) = 1-H_n(t)$, $-\infty < t < \infty$. But for exponential hazard estimators $F_n^E(t)$ and $G_n^E(t) = 1 - exp\left\{-\sum_{j=1}^{n}(1-\delta_{(j)})I(Z_{(j)} \le t)/(n-j+1)\right\}$ of $F(t)$ and $G(t)$, we have

$$\left(1-F_n^E(t)\right)\left(1-G_n^E(t)\right) = exp\left\{-\sum_{j=1}^{n}\frac{I(Z_{(j)} \le t)}{n-j+1}\right\} \ne 1-H_n(t).$$

Moreover, for $t \ge Z_{(n)}, F_n(t) = 1$, but $F_n^E(t) < 1$. Therefore, F_n is a correct estimator of continuous d.f. F than F_n^{PL} and F_n^E. In the Fig. 1, we demonstrate plots of estimators $1-F_n$, $1-F_n^{PL}$, and $1-F_n^E$ of survival function $1-F$ using well-known Channing House data of size n=97.

About the Author

Dr. Abdurahim Abdushukurov is a Professor and Head, Department of Probability Theory and Mathematical Statistics, National University of Uzbekistan, Uzbekistan. He is also Vice Rector at National University of Uzbekistan. He was the Scientific Researcher at the Institute of Mathematics under the Academy of Sciences of the Republic of Uzbekistan (1981–1990). He has also worked as Associate Professor (1991–2006) and Vice Dean (1992–1995, 2000–2003). He heads the Department of Probability Theory and Mathematical Statistics since 2006. He has authored and co-authored more than 130 papers and 5 books, including *Statistics of Incomplete Data: Asymptotic Results for Nonclassical Models* (NUU, 2009). He is known for proposing a number of semi-parametric and nonparametric estimators in the censored data models including the well-known ACL (Abdushukurov–Cheng–Lin) estimator in Proportional Hazards Model of Koziol–Green.

Cross References

►Censoring Methodology
►Gaussian Processes
►Kaplan-Meier Estimator
►Nonparametric Statistical Inference
►Survival Data

References and Further Reading

Abdushukurov AA (1998) Nonparametric estimation of the distribution function based on relative risk function. Commun Stat Theor Meth 27(8):1991–2012

Abdushukurov AA (1999) On nonparametric estimation of reliability indices by censored samples. Theor Probab Appl 43(N.1): 3–11

Abdushukurov AA (2009) Statistics of incomplete observations: asymptotical theory of estimation for nonclassical models. University Press, Tashkent

Akritas MG (2000) The central limit theorem under censoring. Bernoulli 6:1109–1120

Csörgő S (1996) Universal gaussian approximations under random censorship. Ann Stat 24(6):2744–2778

Gill RD (1980) Censoring and stochastic integrals. Mathematical centre tracts, vol 124. Mathematisch Centrum, Amsterdam

Gill RD (1994) Glivenko-Cantelli for Kaplan–Meier. Math Meth Stat 3(1):76–87

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 58:457–481

Nonparametric Models for ANOVA and ANCOVA Designs

Michael G. Akritas
Professor
Pennsylvania State University, State College, PA, USA

Brief History of Nonparametric Statistics

A statistical procedure is called nonparametric if it is valid under less restrictive assumptions than those required by the classical, or parametric, procedures. The difference between parametric and nonparametric statistics is best illustrated in the context of estimating the population distribution function on the basis of a ►simple random sample from the population. The parametric approach assumes that the population distribution belongs to a particular parametric family of distributions, such as the normal, and estimates the distribution by estimating the unknown parameters. In contrast, the nonparametric approach uses the empirical distribution function. In fact, parametric statistics makes frequent use of a particular branch of nonparametric statistics which consists of diagnostic procedures, including graphics and goodness-of-fit tests, aimed at confirming the approximate validity of the employed assumption.

In its earliest form, the field of nonparametric procedures comprised mainly of *distribution free* test procedures for relatively simple designs. Examples are the ►sign test and signed-rank test for the one-sample location problem and paired data design, the Mann–Whitney–Wilcoxon (MWW) rank sum test for the two-sample problem (see ►Wilcoxon–Mann–Whitney Test), the Kruskal–Wallis test for k-sample problem, the Friedman test for a complete randomized block design, the Wald–Wolfowitz runs test for randomness, Kendall's and Spearman's rank correlation coefficients and tests for independence, and Fisher's exact test for association in 2×2 contingency tables. The popularity of these tests lies in the fact that

they can be applied to ordinal and rank data, have good robustness properties, while some of them, notably the MWW rank sum test and the Kruskal–Wallis test, retain high efficiency relative to corresponding classical procedures under normality. Moreover, the idea of inverting tests to obtain estimators gave rise to the popular Hodges–Lehmann (see ▶Hodges-Lehmann Estimators) and Theil–Sen estimators. Detailed accounts of this branch of nonparametric statistics can be found in Hajek and Sidak (1967), Lehmann (1975), Hettmansperger (1984), Hollander and Wolfe (1999), and Gibbons and Chakraborti (2010).

Following this early activity on distribution-free test procedures and corresponding estimators, nonparametric statistics expanded to include *nonparametric regression* which focuses on estimating an unknown regression function using *smoothing methods* such as kernel, splines and wavelets. See Eubank (1988), Hart (1997), Wahba (1990), Green and Silverman (1994), Fan and Gijbels (1996), Härdle et al. (1998). This remains a very active research area with emphasis now placed in overcoming complications caused by the presence of a large number of covariates. This has led to a number of innovative *semi-parametric* models, variable selection methods, and dimension reduction methods; see Hastie and Tibshirani (1990), Cook and Li (2002), Fan and Li (2001), Fan and Lv (2008) for some representative contributions in these areas.

Another recent direction of nonparametric statistics deals with efforts to extend the Kruskal–Wallis test to more complicated factorial designs. The main obstacle towards a successful extension lies in the fact that in such designs the observations are not iid under the more specialized null hypotheses, such as the hypotheses of no main effects and no interaction, which are of interest in such designs. Two main developments have been the aligned rank tests (cf. Puri and Sen 1971; Mansouri 1999), and distance based methods (cf. Hettmansperger and McKean 1998); see also the recent approach of Gao and Alvo (2005). These procedures, however, are not based on the overall ranks of the observations, and have been developed for homoscedastic models. A different approach to constructing rank tests for factorial designs is the so-called rank transform (RT) method. It consists of substituting ranks instead of the observations in the classical F statistics. This approach was motivated by the fact that asymptotically equivalent versions of the MWW rank sum test or the Kruskal–Wallis test statistics can be obtained by substituting ranks instead of the observations in the pooled variance two sample t test statistic or the one-way ANOVA F statistic, respectively. The validity of the RT in these cases sparked an interest in exploring it further. While the RT procedure was shown to be valid in certain balanced additive designs, Akritas (1990) showed that in general it fails. The main argument for establishing this failure lies in the fact that the hypotheses of no main effects and no interaction are not invariant under monotone transformations and thus cannot be tested by statistics (such as rank statistics) which are invariant under such transformations. In addition, the aforementioned paper pointed to the fact that, in general, ranks are heteroscedastic even if the original observations are homoscedastic. These observations motivated Akritas and Arnold (1994) to introduce nonparametric versions of the linear models for factorial designs and to construct Wald-type weighted rank test statistics for testing nonparametric versions of the common hypotheses of no main effects and no interactions. This approach has been extended to analysis of covariance designs and to high dimensional settings. The rest of this article gives an account of these developments.

Factorial Designs

Consider for simplicity the crossed two-factor design, and let Y_{ijk} denote the kth observation in cell formed from the ith level of the row factor and the jth level of the column factor. The fully nonparametric version of this design specifies only that

$$Y_{ijk} \sim F_{ij}, \tag{1}$$

where F_{ij} is some cumulative distribution function. The model (1) allows discrete and continuous quantitative response variables, and does not assume homoscedasticity. Akritas and Arnold (1994) defined nonparametric hypotheses in terms of the following decomposition

$$F_{ij}(y) = M(y) + A_i(y) + B_j(y) + \Gamma_{ij}(y),\ i = 1, \ldots, a,\ j = 1, \ldots, b, \tag{2}$$

where $M = \overline{F}_{\cdot\cdot}$, $A_i = \overline{F}_{i\cdot} - M$ and $B_j = \overline{F}_{\cdot j} - M$ are the nonparametric main effects for the row and column factors, and $\Gamma_{ij} = F_{ij} - A_i - B_j$ are nonparametric versions of the interaction effects. The nonparametric hypotheses of no main row or column effects, and no interaction specify that the corresponding nonparametric effects are zero: $H_0(A) : A_i = 0$, for all i, $H_0(B) : B_j = 0$, for all j, and $H_0(\Gamma) : \Gamma_{ij} = 0$, for all i and j. Letting $\mu_{ij} = \int y dF_{ij}(y)$ denote the mean of Y_{ijk} it easily follows that

$$\alpha_i = \int y dA_i(y),\ \beta_j = \int y dB_j(y),\ \text{and}\ \gamma_{ij} = \int y d\Gamma_{ij}(y)$$

constitute the unique decomposition of μ_{ij} as $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ so that $\sum \alpha_i = \sum \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$. It follows that the nonparametric versions of the null hypotheses are

stronger than their parametric counterparts (i.e., imply but are not implied by them). It turns out that these stronger versions of the common hypotheses are invariant under monotone transformations and thus can be tested by rank statistics.

For designs with independent observations, Akritas et al. (1997) (abbreviated by AAB97 from now on) constructed a class of Wald-type rank test statistics which adjusts for the heteroscedasticity in the ranks (Akritas 1990). A scored rank version of this statistic was developed in Brunner and Puri (2002). A Box-type approximation which enhances the small sample type I error performance was also proposed in AAB97; for more details on the Box-type approximation procedure see Brunner et al. (1997). The Wald-type statistics in AAB97 were further extended to repeated measures designs in Akritas and Brunner (1997a), to censored independent data in Akritas and Brunner (1997b), and to censored dependent data in O'Gorman and Akritas (2001).

Akritas and Arnold (1994) introduced the nonparametric hypotheses only in the context of the full (or saturated) model, using equal weights in the constraints that define the effects. This practice was followed in the aforementioned subsequent literature on nonparametric hypotheses. An extension of the class of nonparametric hypotheses to non-saturated designs, and to arbitrary weights in defining the effects, is given in Akritas, Stavropoulos and Caroni (2009). Wald statistics, however, are not convenient for non-saturated designs. This is mainly because estimators of the effects, on which Wald statistics are based, are not always available in closed form. This is particularly prevalent with non-orthogonal designs. Moreover, software development, and consequently statistical practice, has favored heavily the likelihood ratio F tests over the Wald-type tests. Akritas (1990) constructed the first rank version of a weighted F statistic but remarked that, because F statistics do not have a closed form expression, a general asymptotic theory of such statistics would require different asymptotic methods. By adopting Wald-type statistics instead of weighted F statistics, AAB97 circumvented this problem, but the methodology is confined to saturated designs. Akritas et al. (2009) derive an asymptotic theory that covers rank versions of weighted F statistics for any hypothesis in any factorial design. In particular, the theory and the statistics apply not only for testing that some effects are zero against general alternatives (adjusted, or type III in SAS, sum of squares) but also for the so-called sequential analysis (type I sum of squares in SAS). Thus, the aforementioned paper extends the application of the robust rank methodology to the entire spectrum of ANOVA applications for factorial designs.

Analysis of Covariance Designs

Consider for simplicity the one-way analysis of covariance design. Thus, we observe (Y_{ij}, X_{ij}), for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, where i enumerates the factor or treatment levels, the covariates X_{ij} are either observed constants or observed random variables, and Y_{ij} is the observed response random variable. The fully nonparametric model of Akritas et al. (2000) assumes only that, given $X_{ij} = x$,

$$Y_{ij} \sim F_{ix},$$

where F_{ix} is a distribution function that depends only on i, x in an unspecified way. The decomposition

$$F_{ix}(y) = M(y) + A_i(y) + D_x(y) + C_{ix}(y),$$

which is extends the decomposition (2) to allowing for a continuous index, helps define effects in this nonparametric context.

Akritas et al. (2000) and Akritas and Van Keilegon (2001) develop test procedures for the hypothesis for no covariate adjusted factor effects $H_0(A) : A_i(y) = 0$, for all i and all y. This procedure was extended to censored data in Du et al. (2003), while Tsangari and Akritas (2003) extended the procedure to ANCOVA designs with up to three covariates. Tsangari and Akritas (2004) considered ANCOVA designs with dependent data.

Testing for the covariate effect, as well as the factor-covariate interaction effect, falls in the category of testing against *high-dimensional alternatives*; see Wang and Akritas (2006). The techniques for doing so are similar to those for analyzing high-dimensional factorial designs, which is discussed next.

High Dimensional Factorial Designs

Advances in data gathering technologies have produced massive data sets giving rise to designs with a large number of factor levels and few replications. Such high dimensional designs have motivated the development of high-dimensional analysis of variance or HANOVA – a term introduced by Fan and Lin (1998). Having a large number of factor levels but possibly few replications calls for a very different asymptotic theory than that involved in the developments outlined in Section ►"Factorial Designs". For example, the asymptotic theory of the F statistic is now obtained by studying the difference $F - 1$.

Dealing with heteroscedasticity when there are only a few replications per cell is particularly challenging. Several approaches for doing so are explored in Akritas and Papadatos (2004), where an up to date literature review is presented. Rank statistics for testing the nonparametric hypotheses (2), and their versions for higher-way layouts, with independent data are presented in Wang and

Akritas (2004, 2009). Bathke and Harrar (2008) considered a version of the nonparametric hypotheses for multivariate designs and constructed corresponding rank test procedures.

The conceptual connection between a factorial design having factors have many levels and ANCOVA designs (the covariate can be thought of as a factor with many levels) was exploited in Wang and Akritas (2006), and in Wang et al. (2008) to develop tests for the covariate effect.

When the factor with many levels is time, the data, known as curves or functional data, are often dependent. The aforementioned procedures have been extended to testing for various aspects of functional dependent data in Wang and Akritas (2010), and in Wang et al. (2010).

Finally, nonparametric models for mixed and random effects designs are presented in Gaugler (2008). The main contribution of the nonparametric modeling in mixed effects designs is the relaxation of the so-called *symmetry assumption*. This is the assumption of independence of the main random effect and the interaction effect. It is shown that violations of this assumption (which likely occur in the majority of real data applications) has devastating effects on the significance level of the traditional *F* test.

About the Author

Michael Akritas is Professor of Statistics, Penn State University. He has served as Director of the Statistical Consulting Center for Astronomy (1994–2000), as Guest Editor for *Sociological Methods and Research*, and for *Statistics & Probability Letters*, and as Associate Editor for *Statistics & Probability Letters*, *Journal of Environmental Statistics*, *Journal of Nonparametric Statistics*, and the *Journal of the American Statistical Association*. He is currently Co-Editor of the *Journal of Nonparametric Statistics*. He is Elected Fellow of the American Statistical Association (2001), and Fellow of the Institute of Mathematical Statistics (2001).

Cross References

►Analysis of Covariance
►Analysis of Variance
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Rank Transformations
►Smoothing Techniques

References and Further Reading

Akritas MG (1990) The rank transform method in some two-factor designs. J Am Stat Assoc 85:73–78
Akritas MG, Arnold SF (1994) Fully nonparametric hypotheses for factorial designs I:multivariate repeated measures designs. J Am Stat Assoc 89:336–343
Akritas MG, Brunner E (1997a) A unified approach to rank tests for mixed models. J Stat Plan Infer 61:249–277
Akritas MG, Brunner E (1997b) Nonparametric methods for factorial designs with censored data. J Am Stat Assoc 92:568–576
Akritas MG, Papadatos N (2004) Heteroscedastic one-way ANOVA and lack-of-fit tests. J Am Stat Assoc 99:368–382
Akritas MG, Van Keilegon I (2001) Nonparametric ANCOVA methods for heteroscedastic nonparametric regression models. J Am Stat Assoc 96:220–232
Akritas MG, Arnold SF, Brunner E (1997) Nonparametric hypotheses and rank statistics for unbalanced factorial designs. J Am Stat Assoc 92:258–265
Akritas MG, Arnold SF, Du Y (2000) Nonparametric models and methods for nonlinear analysis of covariance. Biometrika 87:507–526
Akritas MG, Stavropoulos A, Caroni C (2009) Asymptotic theory of weighted F-statistics based on ranks. J Nonparamer Stat 21: 177–191
Bathke AC, Harrar SW (2008) Nonparametric methods in multivariate factorial designs for large number of factor levels. J Stat Plan Infer 138:588–610
Brunner E, Puri ML (2002) A class of rank-score tests in factorial designs. J Stat Plan Infer 103:331–360
Brunner E, Dette H, Munk A (1997) Box-Type approximations in nonparametric factorial designs. J Am Stat Assoc 92:1494–1502
Cook RD, Li B (2002) Dimension reduction for the conditional mean in regression. Ann Stat 30:455–474
Du Y, Akritas MG, Van Keilegom I (2003) Nonparametric analysis of covariance for censored data. Biometrika 90:269–287
Eubank R (1988) Spline smoothing and nonparametric regression. Marcel Dekker, New York
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman & Hall, London
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360
Fan J, Lin S (1998) Test of significance when data are curves. J Am Stat Assoc 93:1007–1021
Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc B 70:849–911 (with discussions)
Gao X, Alvo M (2005) A Unified nonparametric approach for unbalanced factorial designs. J Am Stat Assoc 100:926–941
Gaugler T (2008) Nonparametric models for crossed mixed effects designs. Doctoral dissertation, The Pennsylvania State University
Gibbons JD, Chakraborti S (2010) Nonparametric statistical inference, 5th edn. Taylor & Francis/CRC Press, Boca Raton
Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models. Chapman & Hall/CRC Press, Boca Raton
Hajek J, Sidak Z (1967) Theory of rank tests. Academic, New York
Hall P, Marron JS, Neeman A (2005) Geometric representation of high dimension, low sample size data. J R Stat Soc B 67:427–444
Härdle W, Kerkyacharian G, Picard D, Tsybakov A (1998) Wavelets, Approximation, and Statistical Applications. Lecture notes in statistics, vol 129. Springer, New York
Hart JD (1997) Nonparametric smoothing and lack-of-fit tests. Springer, New York
Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman & Hall/CRC Press, Boca Raton
Hettmansperger TP (1984) Statistical inference based on ranks. Krieger, Malabar
Hettmansperger TP, McKean JW (1998) Robust nonparametric statistical methods. Hodder Arnold, London

Hollander M, Wolfe DA (1999) Nonparametric statistical methods, 2nd edn. Wiley, New York
Lehmann EL (1975) Nonparametrics: statistical methods based on ranks. Holden-Day, San Francisco
Li H, Lindsay B, Waterman R (2003) Efficiency of projected score methods in rectangular array asymptotics. J R Stat Soc B 65:191–208
Mansouri H (1999) Aligned rank transform tests in linear models. J Stat Plan Infer 79:141–155
O'Gorman J, Akritas MG (2001) Nonparametric models and methods for designs with dependent censored data. Biometrics 57:88–95
Puri ML, Sen PK (1971) Nonparametric methods in multivariate analysis. Wiley, New York
Tsangari H, Akritas MG (2003) Nonparametric ANCOVA with two and three covariates. J Multivariate Anal 88:298–319
Tsangari H, Akritas MG (2004) Nonparametric models and methods for ANCOVA with dependent data. J Nonparametr Stat 16: 403–420
Wahba G (1990) Spline models for observational data. SIAM, Philadelphia
Wang H, Akritas MG (2004) Rank tests for ANOVA with large number of factor levels. J Nonparametr Stat 16:563–590
Wang L, Akritas MG (2006) Testing for covariate effects in the fully nonparametric analysis of covariance model. J Am Stat Assoc 101:722–736
Wang H, Akritas MG (2009) Rank tests in heteroscedastic multi-way HANOVA. J Nonparametr Stat 21:663–681
Wang H, Akritas MG (2010) Inference from heteroscedastic functional data. J Nonparametr Stat (in press)
Wang L, Akritas MG, Van Keilegom I (2008) ANOVA-type nonparametric diagnostic tests for heteroscedastic regression models. J Nonparametr Stat 20:365–382
Wang H, Higgins J, Blasi D (2010) Distribution-free tests for no effect of treatment in heteroscedastic functional data under both weak and long range dependence. Stat Probab Lett (in press)

Nonparametric Predictive Inference

Frank P. A. Coolen
Professor
Durham University, Durham, UK

Overview

Nonparametric predictive inference (NPI) is a statistical method based on Hill's assumption $A_{(n)}$ Hill (1968), which gives a direct conditional probability for a future observable random quantity, conditional on observed values of related random quantities (Augustin and Coolen 2004; Coolen 2006). Suppose that $X_1, \ldots, X_n, X_{n+1}$ are continuous and exchangeable random quantities. Let the ordered observed values of $X_1, \ldots, X_n$ be denoted by $x_{(1)} < x_{(2)} < \ldots < x_{(n)} < \infty$, and let $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$ for ease of notation. For a future observation X_{n+1}, based on n observations, $A_{(n)}$ (Hill 1968) is

$$P(X_{n+1} \in (x_{(j-1)}, x_{(j)})) = \frac{1}{n+1} \quad \text{for } j = 1, 2, \ldots, n+1$$

$A_{(n)}$ does not assume anything else, and is a post-data assumption related to exchangeability. Hill discusses $A_{(n)}$ in detail. Inferences based on $A_{(n)}$ are predictive and nonparametric, and can be considered suitable if there is hardly any knowledge about the random quantity of interest, other than the n observations, or if one does not want to use such information, e.g., to study effects of additional assumptions underlying other statistical methods. $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but it provides optimal bounds for probabilities for all events of interest involving X_{n+1}. These bounds are lower and upper probabilities in the theories of imprecise probability (Walley 1991) and interval probability (Weichselberger 2001), and as such they have strong consistency properties (Augustin and Coolen 2004). NPI is a framework of statistical theory and methods that use these $A_{(n)}$-based lower and upper probabilities, and also considers several variations of $A_{(n)}$ which are suitable for different inferences. For example, NPI has been presented for Bernoulli data, multinomial data and lifetime data with right-censored observations. NPI enables inferences for $m \geq 1$ future observations, with their interdependence explicitly taken into account, and based on sequential assumptions $A_{(n)}, \ldots, A_{(n+m-1)}$. NPI provides a solution to some explicit goals formulated for objective (Bayesian) inference, which cannot be obtained when using precise probabilities (Coolen 2006). NPI is also exactly calibrated (Lawless and Fredette 2005), which is a strong consistency property, and it never leads to results that are in conflict with inferences based on empirical probabilities.

NPI for Bernoulli random quantities (Coolen 1998) is based on a latent variable representation of Bernoulli data as real-valued outcomes of an experiment in which there is a completely unknown threshold value, such that outcomes to one side of the threshold are successes and to the other side failures. The use of $A_{(n)}$ together with lower and upper probabilities enable inference without a prior distribution on the unobservable threshold value as is needed in ▶Bayesian statistics where this threshold value is typically represented by a parameter. Suppose that there is a sequence of $n + m$ exchangeable Bernoulli trials, each with "success" and "failure" as possible outcomes, and data consisting of s successes in n trials. Let Y_1^n denote the random number of successes in trials 1 to n, then a sufficient representation of the data for NPI is $Y_1^n = s$, due to the

assumed exchangeability of all trials. Let Y_{n+1}^{n+m} denote the random number of successes in trials $n+1$ to $n+m$. Let $R_t = \{r_1, \ldots, r_t\}$, with $1 \le t \le m+1$ and $0 \le r_1 < r_2 < \ldots < r_t \le m$, and, for ease of notation, define $\binom{s+r_0}{s} = 0$. Then the NPI upper probability for the event $Y_{n+1}^{n+m} \in R_t$, given data $Y_1^n = s$, for $s \in \{0, \ldots, n\}$, is

$$\overline{P}\left(Y_{n+1}^{n+m} \in R_t \,\middle|\, Y_1^n = s\right) = \binom{n+m}{n}^{-1} \sum_{j=1}^{t} \left[\binom{s+r_j}{s} - \binom{s+r_{j-1}}{s}\right] \binom{n-s+m-r_j}{n-s}$$

The corresponding NPI lower probability is derived via the conjugacy property

$$\underline{P}\left(Y_{n+1}^{n+m} \in R_t \,\middle|\, Y_1^n = s\right) = 1 - \overline{P}\left(Y_{n+1}^{n+m} \in R_t^c \,\middle|\, Y_1^n = s\right)$$

where $R_t^c = \{0, 1, \ldots, m\} \backslash R_t$.

For multinomial data, a latent variable representation via segments of a probability wheel has been presented, together with a corresponding adaptation of $A_{(n)}$ (Coolen and Augustin 2009). For data including right-censored observations, as often occur in lifetime data analysis, NPI is based on a variation of $A_{(n)}$ which effectively uses a similar exchangeability assumption for the future lifetime of a right-censored unit at its moment of censoring (Coolen and Augustin 2004). This method provides an attractive predictive alternative to the well-known Kaplan–Meier estimate (see ▶Kaplan-Meier Estimator) for such data.

Applications

Many applications of NPI have been presented in the literature. These include solutions to problems in Statistics, Reliability and Operational Research. For example, NPI methods for multiple comparisons of groups of real-valued data are attractive for situations where such comparisons are naturally formulated in terms of comparison of future observations from the different groups (Coolen 2001). NPI provides a frequentist solution to such problems which does not depend on counterfactuals, which play a role in hypothesis testing and are often criticized by opponents of frequentist statistics. An important advantage of the use of lower and upper probabilities is that one does not need to add assumptions to data which one feels are not justified. A nice example occurs in precedence testing, where experiments to compare different groups may be terminated early in order to save costs or time (Coolen-Schrijner et al. 2009). In such cases, the NPI lower and upper probabilities are the sharpest bounds corresponding to all possible orderings of the not-fully observed data. NPI provides an attractive framework for decision support in a wide range of problems where the focus is naturally on a future observation. For example, NPI methods for replacement decisions of technical units are powerful and fully adaptive to process data (Coolen-Schrijner and Coolen 2004).

NPI has been applied for comparisons of multiple groups of proportions data (Coolen and Coolen-Schrijner 2007), where the number m of future observations per group plays an interesting role in the inferences. Effectively, if m increases the inferences tend to become more imprecise, while imprecision tends to decrease if the number of observations in the data set increases. NPI for Bernoulli data has also been implemented for system reliability, with particularly attractive algorithms for optimal redundancy allocation (Coolen-Schrijner et al. 2008; MacPhee et al. 2009). NPI for multinomial data enables inference if the number of outcome categories is not known, and explicitly distinguishes between defined and undefined categories for which no observations are available yet (Coolen 2007). Typically, if outcome categories have not occurred yet, the NPI lower probability of the next observation falling in such a category is zero, but the corresponding NPI upper probability is positive and depends on whether or not the category is explicitly defined, on the total number of categories or whether this number is unknown, and on the number of categories observed so far. Such NPI upper probabilities can be used to support cautious decisions, which are often deemed attractive in reliability and ▶risk analysis.

Challenges

Development of NPI is gathering momentum, inferential problems for which NPI solutions have recently been presented or are being developed include aspects of medical diagnosis with the use of ROC curves, robust classification, inference on competing risks, quality control and ▶acceptance sampling. Main research challenges for NPI include its generalization for multidimensional data, which is similarly challenging for NPI as for general nonparametric methods due to the lack of a unique natural ordering of the data. NPI theory and methods that enable information from covariates to be taken into account also provide interesting and challenging research opportunities. A research monograph introducing NPI theory, methods and applications is currently in development, further information is available from www.npi-statistics.com.

About the Author

Frank P.A. Coolen is Professor of Statistics at the Department of Mathematical Sciences, Durham University (UK). He served as editor of the Section on Reliability – Mathematical and Statistical Methods for the Wiley *Encyclopedia of Quantitative Risk Analysis and Assessment* (2008). He currently serves on the editorial

boards of *Journal of Statistical Planning and Inference*, *Journal of Statistical Theory and Practice*, *Journal of Risk and Reliability*, and *Quality and Reliability Engineering International*. His main research is on foundations and methods of Statistics and Reliability, in particular he has been developing Nonparametric Predictive Inference (NPI) over the last decade. Together with collaborators and students, Frank has published about 150 journal and conference papers, a majority of these on NPI or other statistical methods that use lower and upper probabilities.

Cross References

►Acceptance Sampling
►Censoring Methodology
►Nonparametric Statistical Inference
►ROC Curves

References and Further Reading

Augustin T, Coolen FPA (2004) Nonparametric predictive inference and interval probability. J Stat Planning Infer 124:251–272
Coolen FPA (1998) Low structure imprecise predictive inference for Bayes' problem. Stat Probab Lett 36:349–357
Coolen FPA (2006) On nonparametric predictive inference and objective Bayesianism. J Logic Lang Inform 15:21–47
Coolen FPA (2007) Nonparametric prediction of unobserved failure modes. J Risk Reliab 221:207–216
Coolen FPA, Augustin T (2009) A nonparametric predictive alternative to the imprecise Dirichlet model: the case of a known number of categories. Int J Approx Reasoning 50:217–230
Coolen FPA, Coolen-Schrijner P (2007) Nonparametric predictive comparison of proportions. J Stat Planning Infer 137:23–33
Coolen FPA, van der Laan P (2001) Imprecise predictive selection based on low structure assumptions. J Stat Planning Infer 98:259–277
Coolen FPA, Yan KJ (2004) Nonparametric predictive inference with right-censored data. J Stat Planning Infer 126:25–54
Coolen-Schrijner P, Coolen FPA (2004) Adaptive age replacement based on nonparametric predictive inference. J Oper Res Soc 55:1281–1297
Coolen-Schrijner P, Coolen FPA, MacPhee IM (2008) Nonparametric predictive inference for systems reliability with redundancy allocation. J Risk Reliab 222:463–476
Coolen-Schrijner P, Maturi TA, Coolen FPA (2009) Nonparametric predictive precedence testing for two groups. J Stat Theory Pract 3:273–287
Hill BM (1968) Posterior distribution of percentiles: Bayes' theorem for sampling from a population. J Am Stat Assoc 63:677–691
Hill BM (1988) De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In: Bernardo JM et al (eds) Bayesian statistics, vol 3. Oxford University Press, Oxford, pp 211–241
Lawless JF, Fredette M (2005) Frequentist prediction intervals and predictive distributions. Biometrika 92:529–542
MacPhee IM, Coolen FPA, Aboalkhair AM (2009) Nonparametric predictive system reliability with redundancy allocation following component testing. J Risk Reliab 223:181–188
Walley P (1991) Statistical reasoning with imprecise probabilities. Chapman & Hall, London
Weichselberger K (2001) Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervalwahrscheinlichkeit als umfassendes Konzept (in German). Physika, Heidelberg

Nonparametric Rank Tests

Thomas P. Hettmansperger
Professor Emeritus of Statistics
Penn State University, State College, PA, USA

There are three major reasons for considering nonparametric rank tests. They do not require a parametric distribution assumption, such as normality, for the observations in order to determine the p-value of the test (nonparametric or distribution-free property). When the observations come from a population with tails that are heavier than those of a normal population, nonparametric rank tests can be much more powerful and efficient than t and F tests. Finally, they are more robust to ►outliers and data contamination than traditional t and F tests. These three properties (nonparametric, efficient, and robust) have been well documented; see, for example, Hettmansperger and McKean (1998), Hollander and Wolfe (1999), Higgins (2004), Lehmann (2006), and Sprent and Smeeton (2007). Nonparametric methods are implemented in all standard statistical computing packages.

The most widely used nonparametric rank test is the *Mann–Whitney–Wilcoxon rank sum test* (abbreviated MWW) (This test is also commonly called Wilcoxon–Mann–Whitney test and abbreviated as WMW.) (*see* ►Wilcoxon–Mann–Whitney Test). Suppose we have two samples of independent and identically distributed observations, denoted by $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, from two populations represented by continuous cumulative distribution functions $F(x)$ and $G(y)$, respectively. Often, we assume that the population distributions differ only in their locations (typically the mean or median). Then $G(y) = F(y - \Delta)$ where Δ represents the difference in locations. We are interested in testing the null hypothesis $H_0 : \Delta = 0$ versus an alternative hypothesis such as $H_A : \Delta > 0$. This hypothesis indicates that the Y distribution is shifted to the right of the X distribution. The traditional test would be the one-sided two-sample t test.

Example 1 Researchers held the hypothesis that silver content (measured by the percentage of silver in a coin)

decreased from early mintings of Byzantine coins to later mintings in the mid twelfth century. Data (ordered) is contained in Hendy and Charles (1970) and consists in early minting values: 5.9, 6.2, 6.4, 6.6, 6.8, 6.9, 7.0, 7.2, 7.7 and later minting values: 5.1, 5.3, 5.5, 5.6, 5.8, 5.8, 6.2. Based on this data we wish to test $H_0 : \Delta = 0$ versus $H_A : \Delta > 0$ where Δ represents the difference in locations of the early and late mintings. Let X denote the late minting and Y denote the early minting.

The MWW statistic is computed as follows: combine the X and Y data, assign ranks to the combined data. In case of ties assign the average rank. Then the MWW statistic, denoted by W, is the sum of ranks of the Y data. When the null hypothesis is true, the expected value of W is $n(n+m+1)/2$ and we will reject $H_0 : \Delta = 0$ in favor of $H_A : \Delta > 0$ when W is observed sufficiently far above $n(n+m+1)/2$. Sufficiently far is determined by the p-value of the test: p-value $= P(W > \text{obs}W)$ where obs$W$ is the observed value of W from the data.

Example 1, continued Late minting X data has $m = 7$ observations with ranks 1, 2, 3, 4, 5.5, 5.5, 8.5 and the early minting Y data has $n = 9$ with ranks 7, 8.5, 10, 11, 12, 13, 14, 15, 16. The sum of ranks of the early data is $W = 106.5$, the expected value under the null hypothesis is 76.5, and the p-value is 0.001. This indicates that the observed value of $W = 106.5$ is far above the expected value of 76.6, and hence we reject the null hypothesis and conclude that the later minting contained significantly less silver content than the early minting.

Calculating the p-value: The most convenient method is to approximate the p-value using a normal approximation for the null sampling distribution of W. This requires the standard deviation of W under the null hypothesis and is given by $\sqrt{mn(m+n+1)/12}$. Then $[W - n(m+n+1)/2]/\sqrt{mn(m+n+1)/12}$ can be referred to a standard normal table. Accuracy of the approximation can generally be increased by using a ►continuity correction. For the example above, obs$W = 106.5$, $n(m+n+1)/2 = 76.5$, $\sqrt{mn(m+n+1)/12} = 9.45$, and the standardized value of 3.17 determines the p-value = 0.001.

There are tables of the exact distribution of W for restricted values of m and n. Finally, since, under the null hypothesis, all permutations of the combined data are equally likely, a computer can be used to approximate the p-value by sampling the permutations and assigning the first m observations in a permutation to the X sample. In this case the p-value is approximately equal to the number of sampled permutations such that W is greater than or equal to obs$W$ divided by the number of sampled permutations. Note this null permutation distribution of W does not require specification of the underlying population distributions and this determines the nonparametric property of the MWW rank test.

Since we use only the ranks of the data, outliers will have a minimal effect on the value of W. An extreme outlier will only be assigned the maximum or minimum rank of m+n or 1. This is the source of the robustness of W.

In addition to the MWW test above, there is an associated point estimate of the difference in locations Δ. It is given by the median of the mn pairwise differences $Y_j - X_i$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. For the coins data the estimate of Δ is 7.5. Note that this estimate is not generally equal to the difference in sample medians. If we record the range of Δ values for which the MWW test fails to reject the null hypothesis at significance level approximately 0.05, then we have an approximate 95% confidence interval for Δ. For the example, we find a 95.6% confidence interval to be (4.5,11.0). See the references for detailed discussions of the construction of confidence intervals.

In case we wish to carry out an hypothesis test in a one sample setting, we can use the ►Wilcoxon-signed-rank test. Here we have a random sample $X_1, \ldots, X_n$ from a population with continuous distribution function $F(x)$. Let θ denote the median of the population and suppose that $F(x)$ is symmetric about θ. For testing $H_0 : \theta = 0$ versus $H_A : \theta > 0$ let S be the sum of ranks of the positive observations when ranked among the absolute values of the data. The statistic S compares the right side of the sample to the left side much the way the MWW statistic compares the Y-sample to the X-sample. Under the null hypothesis we have the mean and standard deviation of S as $n(n+1)/4$ and $\sqrt{n(n+1)(2n+1)/24}$, respectively. Using the fact that S has an approximate normal sampling distribution, the p-value of the test can be approximated by standardizing S and referring the standardized statistic to a standard normal table. The corresponding estimate of θ is the median of the pairwise averages $(X_i + X_j)/2$ for $i \le j$. Further, a nonparametric confidence interval is available; see the references for details. Traditional methods are based on the one sample t-statistic.

The MWW rank sum test can be extended to the one-way layout for testing the null hypothesis $H_0 : F_1(x) = \ldots = F_c(x)$. In this case we have samples from c populations. The test is constructed by combining all the data and ranking it. The *Kruskal–Wallis statistic* is then based on the c average ranks for the c samples. The formula is:

$$KW = \frac{12}{N(N+1)} \sum_{i=1}^{c} n_i \left[\bar{R}_i - (N+1)/2 \right]^2,$$

where N is the total sample size. Under the null hypothesis KW has an approximate chisquared distribution with $c-1$ degrees of freedom. Hence, it is easy to approximate the p-value of this test. The MWW methods can then be used for multiple comparisons and the estimation of location differences. The Kruskal-Wallis test enjoys the same nonparametric, efficiency, and robustness properties as the MWW.

In summary, nonparametric rank tests for the one, two, and c-sample designs are readily available. Relative to the traditional t and F tests they are robust and efficient. They provide a very attractive alternative to the traditional tests and are easily implemented using standard statistical software. Rank tests and estimates are available for more complex designs such as multi-way analysis of variance and regression. See the references for details.

About the Author

Professor Hettmansperger was Head of the Statistics Department (1988–1990). He is a Fellow of the American Statistical Association, Institute of Mathematical Statistics, and an elected member of the International Statistical Institute. He was the 1986 recipient of the C. I. Noll Award for Teaching in the Eberly College of Science. In 2004 he received the Noether Senior Scholar Award from the American Statistical Association for his outstanding work and contribution to nonparametric statistics.

Cross References

►Asymptotic Relative Efficiency in Testing
►Nonparametric Statistical Inference
►Statistical Fallacies: Misconceptions, and Myths
►Wilcoxon–Mann–Whitney Test
►Wilcoxon-Signed-Rank Test

References and Further Reading

Hendy MF, Charles JA (1970) The production techniques, silver content, and circulation history of the twelfth-century Byzantine. Archaeometry 12:13–21
Hettmansperger TP, McKean JW (1998) Robust nonparametric statistical inference. Arnold, London
Higgins JJ (2004) Introduction to modern nonparametric statistics. Duxbury, Belmond
Hollander M, Wolfe DA (1999) Nonparametric statistical methods, 2nd edn. Wiley, New York
Lehmann EL (2006) Nonparametrics: statistical methods based on ranks, rev edn. Springer, Berlin
Sprent P, Smeeton NC (2007) Applied nonparametric statistical methods, 4th edn. Chapman & Hall/CRC Press, Boca Raton

Nonparametric Regression Based on Ranks

Jana Jurečková
Professor
Charles University in Prague, Prague, Czech Republic

Consider the linear regression model (see ►Linear Regression Models)

$$Y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + e_i, \quad i = 1, \ldots, N$$

where $\beta_0 \in \mathbb{R}_1$, $\boldsymbol{\beta} \in \mathbb{R}_p$ are unknown parameters and $e_1, \ldots, e_N$ are independent errors, identically distributed according to a continuous d.f. F and $\mathbf{x}_i \in \mathbb{R}_p$ are given regressors, $i = 1, \ldots, N$. We look for the rank test of the hypotheses

$$\mathbf{H}_0^{(1)} : \boldsymbol{\beta} = \mathbf{0} \text{ versus } \mathbf{K}^{(1)} : \boldsymbol{\beta} \neq \mathbf{0}, \quad \beta_0 \text{ unspecified},$$

$$\mathbf{H}_0^{(2)} : \boldsymbol{\beta}^* = (\beta_0, \boldsymbol{\beta}^\top)^\top = \mathbf{0} \text{ versus } \mathbf{K}^{(2)} : \boldsymbol{\beta}^* \neq \mathbf{0}.$$

Another situation is that $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$, where $\boldsymbol{\beta}_1 \in \mathbb{R}_{p_1}$, $\boldsymbol{\beta}_2 \in \mathbb{R}_{p_2}$, $p_1 + p_2 = p$, and we want to test the hypothesis

$$\mathbf{H}_1 : \boldsymbol{\beta}_2 = \mathbf{0} \text{ versus } \boldsymbol{\beta}_2 \neq \mathbf{0}, \quad \beta_0, \boldsymbol{\beta}_1 \text{ unspecified}.$$

Rank Tests for $\mathbf{H}_0^{(1)}$

Let $R_{N1}, \ldots, R_{NN}$ be the ranks of $Y_1, \ldots, Y_N$ and let $a_N(1), \ldots, a_N(N)$ be the scores generated by a nondecreasing, square-integrable score function $\varphi : (0,1) \mapsto \mathbb{R}_1$ so that $a_N(i) = \varphi\left(\frac{i}{N+1}\right)$, $i = 1, \ldots, N$. The rank tests are based on the linear rank statistics

$$S_{Nj} = \sum_{i=1}^{N} (x_{ij} - \bar{x}_{Nj}) a_N(R_{Ni}), \quad \bar{x}_{Nj} = \frac{1}{N} \sum_{i=1}^{N} x_{ij}, \quad j = 1, \ldots, N$$

and their vector

$$\mathbf{S}_N = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}}_N) a_N(R_{Ni}) = (S_{N1}, \ldots, S_{Np})^\top.$$

The distribution function of observation Y_i under $\mathbf{H}_0^{(1)}$ is $F(y - \beta_0)$, $i = 1, \ldots, N$. Hence, the vector of ranks $(R_{N1}, \ldots, R_{NN})$ assumes all possible permutations of $(1, 2, \ldots, N)$ with the same probability $\frac{1}{N!}$. Moreover, $\mathbf{S}_N$ depends only on $\mathbf{x}_1, \ldots, \mathbf{x}_N$, on the scores

$a_N(1),\ldots,a_N(N)$ and on the ranks. Our test for $\mathbf{H}_0^{(1)}$ is based on the quadratic form

$$\mathcal{S}_N = A_N^{-2}\left(\mathbf{S}_N^\top \mathbf{Q}_N^{-1}\mathbf{S}_N\right), \tag{1}$$

where

$$A_N^2 = \frac{1}{N-1}\sum_{i=1}^N (a_N(i)-\bar{a}_N)^2, \quad \mathbf{Q}_N = \sum_{i=1}^N (\mathbf{x}_i-\bar{\mathbf{x}}_N)(\mathbf{x}_i-\bar{\mathbf{x}}_N)^\top$$

and $\mathbf{Q}_N^{-1}$ is replaced by the generalized inverse $\mathbf{Q}_N^-$ if $\mathbf{Q}_N$ is singular. We reject $\mathbf{H}_0^{(1)}$ if $\mathcal{S}_N > k_\alpha$ where k_α is a suitable critical value. The distribution of $\mathcal{S}_N$ and hence the critical values under the hypothesis $\mathbf{H}_0^{(1)}$ do not depend on the distribution function F of the errors. For small N, the critical value can be calculated numerically, but it would become laborious with increasing N. Otherwise we must use asymptotic critical values based on the large-sample approximation of the distribution of $\mathcal{S}_N$. The large sample distribution of $\mathcal{S}_N$ under $\mathbf{H}_0^{(1)}$ is χ^2 distribution with p degrees of freedom, under some conditions on the regressors and the scores.

We reject hypothesis $\mathbf{H}_0^{(1)}$ on the significance level α (usually $\alpha = 0.01$ or 0.05) if $\mathcal{S}_N > \chi_p^2(1-\alpha)$, where $\chi_p^2(1-\alpha)$ is the $(1-\alpha)$ quantile of the χ^2 distribution with p degrees of freedom.

Rank Tests for $\mathbf{H}_0^{(2)}$

For the sake of identifiability of parameter β_0, we should assume that the basic distribution of the errors is symmetric around 0. Let $R_{N1}^+,\ldots,R_{NN}^+$ be the ranks of $|Y_1|,\ldots,|Y_N|$. Choose a score-generating function $\varphi^*(u) = \varphi\left(\frac{u+1}{2}\right) : (0,1) \mapsto [0,\infty)$ and the scores $a_N^*(1),\ldots,a_N^*(N)$ generated by φ^*. Put $x_{i0} = 1$, $i = 1,\ldots,N$, and consider the signed-rank statistics

$$\mathbf{S}_N^+ = \left(S_{N,0}^+, S_{N,1}^+,\ldots,S_{N,p}^+\right)^\top,$$

$$S_{N,j}^+ = \sum_{i=1}^N x_{ij}\ \text{sign}\ Y_i\ a_N^*\left(R_{Ni}^+\right), \quad j = 0,1,\ldots,p.$$

The test criterion for $\mathbf{H}_0^{(1)}$ will be the quadratic form

$$\mathcal{S}_N^+ = A_N^{*-2}\left(\mathbf{S}_N^{+\prime}\left(\mathbf{Q}_N^*\right)^{-1}\mathbf{S}_N^+\right)$$

where $A_N^{*2} = \frac{1}{N}\sum_{i=1}^N\left[a_N^*(i)\right]^2$, $\mathbf{Q}_N^* = \sum_{i=1}^N \mathbf{x}_i^*\mathbf{x}_i^{*\top}$, and $\mathbf{x}_i^* = (x_{i0},x_{i1},\ldots,x_{ip})^\top$.

The distribution of $\mathbf{S}_N^+$ (and hence of $\mathcal{S}_N^+$) is generated by $N!2^N$ equally probable realizations of $(\text{sign}\ Y_1,\ldots,\text{sign}\ Y_N)$ and $\left(R_{N1}^+,\ldots,R_{NN}^+\right)$. The asymptotic distribution of $\mathcal{S}_N^+$ under $\mathbf{H}_0^{(2)}$ will be $\chi^2(p+1)$ under some conditions on the regressors and on the scores. Hence, we reject hypothesis $\mathbf{H}_0^{(2)}$ on the significance level α if $\mathcal{S}_N^+ > \chi_{p+1}^2(1-\alpha)$, where $\chi_{p+1}^2(1-\alpha)$ is the $(1-\alpha)$ quantile of the χ^2 distribution with $p+1$ degrees of freedom.

Tests of $\mathbf{H}_1$ Based on Regression Rank Scores

Consider the model

$$Y_i = \beta_0 + x_{ni}^\top\boldsymbol{\beta} + \mathbf{z}_{ni}^\top\boldsymbol{\delta} + e_i, \quad i = 1,\ldots,n$$

with unknown parameters $\beta_0 \in \mathbb{R}^1$, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\delta} \in \mathbb{R}^q$, and the hypothesis

$$\mathbf{H}_1 : \ \boldsymbol{\delta} = \mathbf{0}, \quad \beta_0,\ \boldsymbol{\beta}\ \text{unspecified}.$$

For the sake of the identifiability of the parameters, we rewrite the model in the form

$$Y_i = \beta_0 + \mathbf{x}_{ni}^\top\boldsymbol{\beta} + (\mathbf{z}_{ni} - \widehat{\mathbf{z}}_{ni})^\top\boldsymbol{\delta} + e_i, \quad i = 1,\ldots,n \tag{2}$$

where $\widehat{\mathbf{Z}}_n$ is the projection of matrix $\mathbf{Z}_n$ with the rows $\mathbf{z}_{n1}^\top,\ldots,\mathbf{z}_{nn}^\top$ on the space spanned by the columns of matrix $\mathbf{X}_n$, i.e.,

$$\widehat{\mathbf{Z}}_n = \widehat{\mathbf{H}}_n\mathbf{Z}_n,\ \widehat{\mathbf{H}}_n = \mathbf{X}_n\left(\mathbf{X}_n^\top\mathbf{X}_n\right)^{-1}\mathbf{X}_n^\top.$$

The model under the hypothesis is just $Y_i = \beta_0 + x_{ni}^\top\boldsymbol{\beta} + e_i$, $i = 1,\ldots,n$. The alternative model is (2). The regression rank scores for the hypothetical model are defined as the vector $\widehat{\mathbf{a}}_n(\tau) = (\hat{a}_{n1}(\tau),\ldots,\hat{a}_{nn}(\tau))^\top$ of solutions of the parametric linear programming problem

$$\begin{aligned}
&\sum_{i=1}^n Y_i\,\hat{a}_{ni}(\tau) := \max \quad \text{subject to}\\
&\sum_{i=1}^n \hat{a}_{ni}(\tau) = (1-\tau)n, \ \sum_{i=1}^n x_{ij}\,\hat{a}_{ni}(\tau) = (1-\tau)\sum_{i=1}^n x_{ij},\\
&j = 1\ldots,p,\\
&\widehat{\mathbf{a}}_n(\tau) \in [0,1]^n, \quad 0 \le \tau \le 1.
\end{aligned} \tag{3}$$

The restrictions in (3) imply that $\widehat{\mathbf{a}}_n(\tau)$ is invariant to the transformations $Y_i \mapsto Y_i + b_0 + \mathbf{x}_i^\top\mathbf{b}$, $b_0 \in \mathbb{R}^1$, $\mathbf{b} \in \mathbb{R}^p$.

Gutenbrunner et al. (1993) constructed the tests of $\mathbf{H}_1$ under some conditions on the tails of F (not heavier than the tails of t-distribution with 5 d.f.) and on the $\mathbf{x}_i$. Take a nondecreasing score function $\varphi : (0,1) \mapsto \mathbb{R}^1$ and calculate the scores

$$\widehat{\mathbf{b}}_n = (\hat{b}_{n1},\ldots,\hat{b}_{nn})^\top,\ \hat{b}_{ni} = -\int_0^1 \varphi(t)\,d\hat{a}_{ni}(t),\ i = 1,\ldots,n$$

and the q-dimensional vector of *linear regression rank scores statistics*

$$\mathbf{S}_n = n^{-1/2}(\mathbf{Z}_n - \widehat{\mathbf{Z}}_n)^\top\widehat{\mathbf{b}}_n.$$

The test criterion for the hypothesis $\mathbf{H}_1$ is

$$\mathcal{T}_n^2 = (A(\varphi))^{-2}\, \mathbf{S}_n^\top\, \widetilde{\mathbf{D}}_n^{-1}\, \mathbf{S}_n \text{ where}$$
$$\widetilde{\mathbf{D}}_n = n^{-1}(\mathbf{Z}_n - \widehat{\mathbf{Z}}_n)^\top(\mathbf{Z}_n - \widehat{\mathbf{Z}}_n).$$

The asymptotic distribution of $\mathcal{T}_n^2$ under $\mathbf{H}_1$ is the χ^2 distribution with q degrees of freedom, under some conditions on $\mathbf{x}_{ni}$ and $\mathbf{z}_{ni}$, $i = 1, \ldots, n$. Hence, we reject $\mathbf{H}_1$ on the significance level α provided $\mathcal{T}_n^2 > \chi_q^2(1-\alpha)$.

The usual choices of the *score-generating function* φ in all above tests are

- Wilcoxon scores: $\varphi(u) = 2u - 1, \quad 0 \le u \le 1$
- van der Waerden (normal scores): $\varphi(u) = \Phi^{-1}(u)$, $0 < u < 1$, where Φ is the standard normal distribution function
- Median scores: $\varphi(u) = \frac{1}{2}\,\text{sign}\left(u - \frac{1}{2}\right), \quad 0 \le u \le 1$

The computation aspects for the regression rank scores, as dual to the regression quantiles, are discussed in monograph of Koenker (2005). The tests were extended to the linear autoregression times series by Hallin and Jurečková (1999); there are other related works cited.

Acknowledgements

The research was supported by the Czech Republic Grant 201/09/0133 and by Research Projects MSM 0021620839.

About the Author

For biography *see* the entry ►Adaptive Linear Regression.

Cross References

►Linear Regression Models
►Nonparametric Rank Tests
►Nonparametric Regression Using Kernel and Spline Methods
►Ranks

References and Further Reading

Gutenbrunner C, Jurečková J, Koenker R, Portnoy S (1993) Tests of linear hypotheses based on regression rank scores. J Nonpar Stat 2:307–331

Hájek J, Šidák Z (1967) Theory of rank tests. Academia, Prague & Academic Press, New York

Hájek J, Šidák Z, Sen PK (2000) Theory of rank tests, 2nd edn. Academic, New York

Hallin M, Jurečková J (1999) Optimal tests for autoregressive models based on autoregression rank scores. Ann Stat 27: 1385–1414

Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge, ISBN 0-521-84573-4

Puri ML, Sen PK (1985) Nonparametric methods in general linear models. Wiley, New York

Nonparametric Regression Using Kernel and Spline Methods

Jean D. Opsomer[1], F. Jay Breidt[2]
[1]Professor
Colorado State University, Fort Collins, CO, USA
[2]Professor and Chair
Colorado State University, Fort Collins, CO, USA

The Statistical Model

The entry by Claeskens and Jansen (same volume) provides an overview of ►nonparametric estimation and describes the major classes of nonparametric regression methods in use today. In the current entry, we provide more details on two of these classes, kernel methods and spline methods. When applying these methods, the researcher is interested in estimating the relationship between one dependent variable, Y, and one or several covariates, $X_1, \ldots, X_q$. We discuss here the situation with one covariate, X (the case with multiple covariates is addressed in the references provided below). The relationship between X and Y can be expressed as the conditional expectation

$$\mathrm{E}(Y|X = x) = f(x).$$

Unlike in parametric regression, the shape of the function $f(\cdot)$ is not restricted to belong to a specific parametric family such as polynomials.

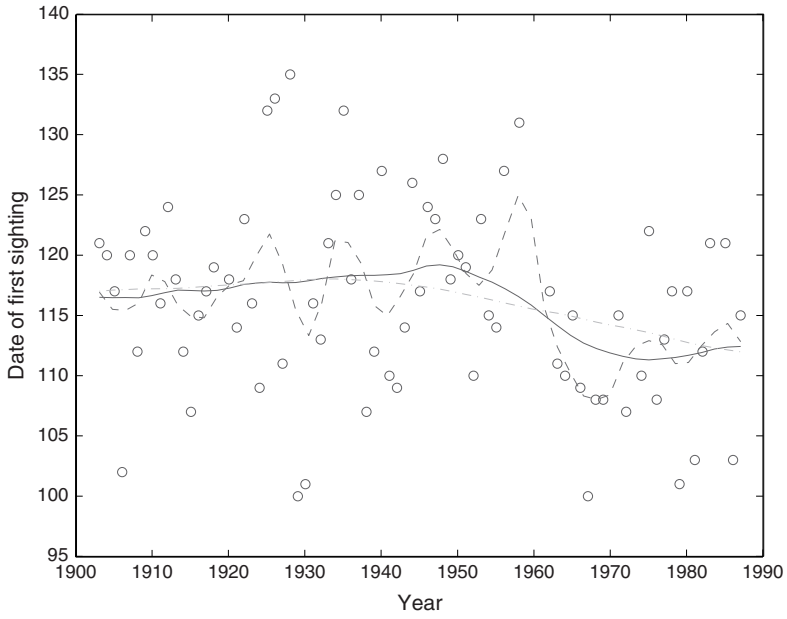
This representation for the mean function is the key difference between parametric and nonparametric regression, and the remaining aspects of the statistical model for (X, Y) are similar between both regression approaches. In particular, the random variable Y is often assumed to have a constant (conditional) variance, $\mathrm{Var}(Y|X) = \sigma^2$, with σ^2 unknown. The constant variance and other common regression model assumptions, such as independence, can be relaxed just as in parametric regression.

Kernel Methods

Suppose that we have a dataset available with observations $(x_1, y_1), \ldots, (x_n, y_n)$. A simple kernel-based estimator of $f(x)$ is the *Nadaraya–Watson kernel regression* estimator, defined as

$$\hat{f}_h(x) = \frac{\sum_{i=1}^n K_h(x_i - x) y_i}{\sum_{i=1}^n K_h(x_i - x)}, \tag{1}$$

with $K_h(\cdot) = K(\cdot/h)/h$ for some kernel function $K(\cdot)$ and bandwidth parameter $h > 0$. The function $K(\cdot)$ is usually a symmetric probability density and examples of commonly used kernel functions are the Gaussian kernel $K(t) = (\sqrt{2\pi})^{-1}\exp(-t^2/2)$ and the *Epanechnikov* kernel $K(t) = \max\left\{\frac{3}{4}(1-t^2), 0\right\}$.



Nonparametric Regression Using Kernel and Spline Methods. Fig. 1 Dates (Julian days) of first sightings of bank swallows in Cayuga Lake basin, with three kernel regressions using bandwidth values h calculated as the range of years multiplied by 0.05 (−−), 0.2 (–) and 0.4 (–·)

Generally, the researcher is not interested in estimating the value of $f(\cdot)$ at a single location x, but in estimating the curve over a range of values, say for all $x \in [a_x, b_x]$. In principle, kernel regression requires computing (1) for any value of interest. In practice, $\hat{f}_h(x)$ is calculated on a sufficiently fine grid of x-values and the curve is obtained by interpolation.

We used the subscript h in $\hat{f}_h(x)$ in (1) to emphasize the fact that the bandwidth h is the main determinant of the shape of the estimated regression, as demonstrated in Fig. 1. When h is small relative to the range of the data, the resulting fit can be highly variable and look "wiggly." When h is chosen to be larger, this results in a less variable, more smooth fit, but it makes the estimator less responsive to local features in the data and introduces the possibility of bias in the estimator. Selecting a value for the bandwidth in such a way that it balances the variance with the potential bias is therefore a crucial decision for researchers who want to apply nonparametric regression on their data. Data-driven bandwidth selection methods are available in the literature, including in the references provided below.

A class of kernel-based estimators that generalizes the Nadaraya–Watson estimator in (1) is referred to as *local polynomial regression* estimators. At each location x, the estimator $\hat{f}_h(x)$ is obtained as the estimated intercept, $\hat{\beta}_0$, in the weighted least squares fit of a polynomial of degree p,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 + \beta_1(x_i - x) + \cdots + \beta_p(x_i - x)^p\right) K_h(x_i - x).$$

This estimator can be written explicitly in matrix notation as

$$\hat{f}_h(x) = (1, 0, \ldots, 0)\left(\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x\right)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y}, \qquad (2)$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{W}_x = \text{diag}\{K_h(x_1 - x), \ldots, K_h(x_n - x)\}$ and

$$\boldsymbol{X}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x & \cdots & (x_n - x)^p \end{bmatrix}.$$

It should be noted that the Nadaraya–Watson estimator (1) is a special case of the local polynomial regression estimator with $p = 0$. In practice, the local linear ($p = 1$) and local quadratic estimators ($p = 2$) are frequently used.

An extensive literature on kernel regression and local polynomial regression exists, and their theoretical properties are well understood. Both kernel regression and local polynomial regression estimators are biased but consistent estimators of the unknown mean function, when that function is continuous and sufficiently smooth. For further information on these methods, we refer to reader to

the monographs by Wand and Jones (1995) and Fan and Gijbels (1996).

Spline Methods

In the previous section, the unknown mean function was assumed to be *locally* well approximated by a polynomial, which led to local polynomial regression. An alternative approach is to represent the fit as a *piecewise* polynomial, with the pieces connecting at points called *knots*. Once the knots are selected, such an estimator can be computed globally in a manner similar to that for a parametrically specified mean function, as will be explained below. A fitted mean function represented by a piecewise continuous curve only rarely provides a satisfactory fit, however, so that usually the function and at least its first derivative are constrained to be continuous everywhere, with only the second or higher derivatives allowed to be discontinuous at the knots. For historical reasons, these constrained piecewise polynomials are referred to as *splines*, leading to the name *spline regression* or *spline smoothing* for this type of nonparametric regression.

Consider the following simple type of polynomial spline of degree p:

$$\beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k}(x - \kappa_k)_+^p, \qquad (3)$$

where $p \geq 1$, $\kappa_1, \ldots, \kappa_K$ are the knots and $(\cdot)_+^p = \max\{(\cdot)^p, 0\}$. Clearly, (3) has continuous derivatives up to degree $(p-1)$, but the pth derivative can be discontinuous at the knots. Model (3) is constructed as a linear combination of *basis functions* $1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p$. This basis is referred to as the *truncated power basis*. A popular set of basis functions are the so-called *B-splines*. Unlike the truncated power splines, the B-splines have compact support and are numerically more stable, but they span the same function space. In what follows, we will write $\psi_j(x), j = 1, \ldots, J$ for a set of (generic) basis functions used in fitting regression splines, and replace (3) by $\beta_1\psi_1(x) + \cdots + \beta_J\psi_J(x)$.

For fixed knots, a regression spline is linear in the unknown parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T$ and can be fitted parametrically using least squares techniques. Under the homoskedastic model described in section "▶The Statistical Model", the *regression spline* estimator for $f(x)$ is obtained by solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \beta_j \psi_j(x_i) \right)^2 \qquad (4)$$

and setting $\hat{f}(x) = \sum_{j=1}^{J} \hat{\beta}_j \psi_j(x)$. Since deviations from the parametric shape can only occur at the knots, the amount of smoothing is determined by the degree of the basis and the location and number of knots. In practice, the degree is fixed (with $p = 1, 2$ or 3 as common choices) and the knot locations are usually chosen to be equally-spaced over the range of the data or placed at regularly spaced data quantiles. Hence, the number of knots K is the only remaining smoothing parameter for the spline regression estimator. As K (and therefore J) is chosen to be larger, increasingly flexible estimators for $f(\cdot)$ are produced. This reduces the potential bias due to approximating the unknown mean function by a spline function, but increases the variability of the estimators.

The *smoothing spline* estimator is an important extension of the regression spline estimator. The smoothing spline estimator for $f(\cdot)$ for a set of data generated by the statistical model described in section "▶The Statistical Model" is defined as the minimizer of

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_{a_x}^{b_x} (f^{(p)}(t))^2 dt, \qquad (5)$$

over the set of all functions $f(\cdot)$ with continuous $(p-1)$th derivative and square integrable pth derivative, and $\lambda > 0$ is a constant determining the degree of smoothness of the estimator. Larger values of λ correspond to smoother fits. The choice $p = 2$ leads to the popular *cubic smoothing splines*. While not immediately obvious from the definition, the function minimizing (5) is exactly equal to a special type of regression spline with knots at each of the observation points $x_1, \ldots, x_n$ (assuming each of the locations x_i is unique). For further information on smoothing splines, see the entry by Wahba (same volume).

Traditional regression spline fitting as in (4) is usually done using a relatively small number of knots. By construction, smoothing splines use a large number of knots (typically, n knots), but the smoothness of the function is controlled by a penalty term and the smoothing parameter λ. The *penalized spline* estimator represents a compromise between these two approaches. It uses a moderate number of knots and puts a penalty on the coefficients of the basis functions. Specifically, a simple type of penalized spline estimator for $m(\cdot)$ is obtained by solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \beta_j \psi_j(x_i) \right)^2 + \lambda \sum_{j=1}^{J} \beta_j^2 \qquad (6)$$

and setting $\hat{f}_\lambda(x) = \sum_{j=1}^{J} \hat{\beta}_j \psi_j(x)$ as for regression splines. Penalized splines combine the advantage of a parametric fitting method, as for regression splines, with the flexible adjustment of the degree of smoothness as in smoothing splines. Both the basis function and the exact form

of the penalization of the coefficients can be varied to accommodate a large range of regression settings.

Spline-based regression methods are extensively described in the statistical literature. While the theoretical properties of (unpenalized) regression splines and smoothing splines are well established, results for penalized regression splines have only recently become available. The monographs by Wahba (1990), Eubank (1999) and Ruppert et al. (2003) are good sources of information on spline-based methods.

About the Authors

Jean Opsomer is Professor of Statistics at Colorado State University. Previously, he was Professor of Statistics and Director of the Center for Survey Statistics and Methodology at Iowa State University. He received his Ph.D. in Operations Research in 1995 from Cornell University. He is a Fellow of the American Statistical Association and an Elected Member of the International Statistical Institute. His research interests include survey statistics, nonparametric regression methods and environmental statistics.

Jay Breidt is Professor and Chair of Statistics at Colorado State University. He received his Ph.D. in 1991 from Colorado State University and was on the faculty at Iowa State University from 1991–2000. He is a Fellow of the American Statistical Association, an Elected Member of the International Statistical Institute, and a member of the Federal Economic Statistics Advisory Committee. His research interests include time series modeling, survey sampling, and environmental statistics.

Cross References

►Jump Regression Analysis
►Nonparametric Estimation
►Nonparametric Regression Based on Ranks
►Smoothing Splines
►Smoothing Techniques

References and Further Reading

Eubank RL (1999) Nonparametric regression and spline smoothing, 2nd edn. Marcel Dekker, New York

Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman & Hall, London

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge

Wahba G (1990) Spline models for observational data. SIAM [Society for Industrial and Applied Mathematics], Philadelphia

Wand MP, Jones MC (1995) Kernel Smoothing. Chapman & Hall, London

Nonparametric Statistical Inference

Jean Dickinson Gibbons[1],
Subhabrata Chakraborti[2]
[1]Russell Professor Emerita of Statistics
The University of Alabama, Tuscaloosa, AL, USA
[2]Professor of Statistics
The University of Alabama, Tuscaloosa, AL, USA

Nonparametric statistical inference is a collective term given to inferences that are valid under less restrictive assumptions than with classical (parametric) statistical inference. The assumptions that can be relaxed include specifying the probability distribution of the population from which the sample was drawn and the level of measurement required of the sample data. For example, we may have to assume that the population is symmetric, which is much less restrictive than assuming the population is the normal distribution. The data may be ratings or ranks, i.e., measurements on an ordinal scale, instead of precise measurements on an interval or ratio scale. Or the data may be counts. In nonparametric inference, the null distribution of the statistic on which the inference is based does not depend on the probability distribution of the population from which the sample was drawn. In other words, the statistic has the same sampling distribution under the null hypothesis, irrespective of the form of the parent population. This statistic is therefore called distribution-free, and, in fact, the field of nonparametric statistics is sometimes called distribution-free statistics. Nonparametric methods are often based on ranks, scores, or counts. This allows us to make less restrictive assumptions and still carry out an inference such as calculate a P-value or find a confidence interval. Strictly speaking, the term nonparametric implies an inference that may or may not be concerned with the value of a parameter of the population. This allows more flexibility, in that the inference may be concerned only with the form of the population, as in goodness-of-fit tests, or with some other characteristic of the data distribution, as in tests for randomness, trend or autocorrelation, or it may be that the inference is concerned with the median of the distribution. It has become customary to include all such tests and estimation procedures under the umbrella of nonparametric statistical inference.

The earliest example of nonparametric statistical inference has been attributed to John Arbuthnot in the early 1700s when the sign test statistic was introduced with the analysis of birth data to show that males have a higher birth rate than females. The sign test statistic (see ►Sign Test) is

the number of plus signs among, say, n differences within a set of n paired sample observations that have no ties. The number of plus signs follows the ▶binomial distribution with parameter p representing the probability of a plus sign. This is true regardless of the form of the parent population of pairs, and hence the sign test statistic is distribution-free. The next example of nonparametric statistical inference was goodness-of-fit tests and contingency table analysis due to Karl Pearson in the early 1900s. Also in the early 1900s Charles Spearman, a psychologist, recommended substituting ranks in calculating the correlation coefficient; this statistic is called Spearman's rho, a measure of rank correlation. Sir Maurice Kendall's landmark book *Rank Correlation Methods* appeared in 1948; this treatise covers Spearman's rho and ▶Kendall's tau in great detail. The fifth edition of this book appeared in 1990, co-authored with Jean D. Gibbons. John Walsh published a three-volume handbook of nonparametric statistical methods in 1962, 1965 and 1968. I.R. Savage compiled a bibliography of nonparametric methods published in 1962.

Gottfried Noether published a slim volume titled *Elements of Nonparametric Statistics* in 1967; this book addressed some of the theoretical bases of nonparametric statistics. The first two comprehensive books suitable for classroom use as a course in nonparametric statistics were *Distribution-free Statistical Tests* by James V. Bradley in 1958 and *Nonparametric Statistical Inference* by Jean D. Gibbons in 1971. The latter book is now in its fifth edition, co-authored with Subhabrata Chakraborti. At the present time there are several other books on the subject. The American Statistical Association organized a separate Section on Nonparametric Statistics in the 1990s and now publishes a separate periodical called the *Journal of Nonparametric Statistics*.

The best known and most widely used nonparametric inference procedures are the ▶Wilcoxon-signed-rank test and confidence interval estimate for the median of one population or the median difference in a population of differences, the Wilcoxon rank sum test (or equivalently the Mann-Whitney test, see ▶Wilcoxon–Mann–Whitney Test) and confidence interval estimate for the difference of the medians of two populations, the Kruskal–Wallis test for equal treatment effects in a one-way ▶analysis-of-variance situation and multiple comparisons, the Friedman test for equal treatment effects in a two-way analysis-of-variance situation and multiple comparisons, the Spearman's rho and Kendall's tau coefficients of correlation for samples from bivariate populations. Nonparametric methods are also available for multivariate data as well as for regression settings although some of these procedures are not as well-known. Bootstrapping and ▶permutation tests, which are applicable under very weak assumptions, take advantage of modern computing power and have brought renewed focus and interest into nonparametric methodology.

The primary advantage of nonparametric procedures over classical (parametric) procedures is that they are inherently robust and valid under very weak assumptions. With classical statistical inference, any conclusions reached should be tempered by qualifying statements like, "If the population from which the sample was drawn is the normal distribution, then …." In practice, one seldom has any reliable information about the population. With a small sample size, one cannot even make an informed guess about the shape of the population. Nonparametric methods are easy to use and understand. Their properties can frequently be derived by combinatorial theory rather than calculus. Kendall's tau as a measure of relative concordance of pairs is much easier to interpret than the classical product-moment correlation coefficient. Data used in classical statistics are implicitly assumed to be measured on at least an interval scale. But many studies, particularly in social science research, collect data as opinions or ratings, as in Likert scale data. Such data cannot possibly be assumed to have come from a normal distribution.

Many nonparametric techniques have a classical or parametric counterpart. The classical techniques are derived under a specific set of assumptions (mainly about the distribution, such as normality) and are frequently the most powerful when those assumptions are fully satisfied. But if those assumptions are not met or cannot be verified (which is usually the case in practice), or are disregarded, or are not even known, the researcher can have little or no faith in any inferences drawn. In fact, the inference may be less reliable than a judicious opinion, or even a guess. In this context a typical question then is "How much is lost by using a nonparametric type of inference if the classical assumptions were indeed met?" Studies of the relative power of nonparametric tests have shown that under the classical assumptions, the nonparametric test is frequently almost as powerful as the corresponding classical test, and little power is lost while considerable confidence in the conclusions is gained. And of course the nonparametric methods are still applicable when many of the assumptions behind the classical methods are not satisfied or can not be fully justified.

About the Authors

Jean Dickinson Gibbons is the Thomas D. Russell Professor Emerita of Applied Statistics at the University of Alabama. She earned her Ph.D. in statistics from Virginia Tech in 1962. Professor Gibbons was the Chair of the Applied Statistics Program at the University of Alabama for 20 years. She was the first Chair of ASA Committee on Women in Statistic (1972–1974) and served three terms on

the ASA Board of Directors. She was elected a Fellow of the American Statistical Association at age 34 and was a Senior Fulbright-Hays Scholar at the Indian Statistical Institute in Calcutta. She has published numerous articles on nonparametric statistics, both theoretical and applied. She is author or co-author of many highly regarded textbooks, including the well-known text *Nonparametric Statistical Inference* (first published in 1970 and is now in its fifth and revised edition (with Professor Chakraborti). Professor Gibbons has been widely considered as one of the leading experts in nonparametric statistics and has enjoyed being a mentor and role model for many young women studying statistics.

"Statistics is my love," Gibbons said. "Its my vocation, as well as my avocation. I was so delighted when I discovered statistics and I think that it is a field that will always be of utmost importance." (Virgina Tech College of Science Magazine, Fall Issue 2009).

Subhabrata Chakraborti is Professor of Statistics and a Robert C. and Rosa P. Morrow Faculty Excellence Fellow at the University of Alabama. He earned his Master's and Ph.D. in statistics from the State University of New York at Buffalo in 1982 and 1984, respectively. He is a Fellow of the American Statistical Association, an elected member of the International Statistical Institute, member of the Institute of Mathematical Statistics and has been a Fulbright Senior Scholar. Professor Chakraborti has authored and co-authored over 75 publications in a variety of international journals. He is the co-author of the new editions of *Nonparametric Statistical Inference*, (2003 and 2010) with Jean D. Gibbons. Professor Chakraborti has been the winner of the Burlington Northern Faculty Achievement Award at the University of Alabama for excellence in teaching. He has been cited for his contributions in mentoring and collaborative work with students and colleagues from around the world. Dr. Chakraborti has served (1997–2002) as an Associate Editor of *Computational Statistics and Data Analysis* and is currently serving (1996–) as an Associate editor of *Communications in Statistics*.

Cross References

►Bayesian Nonparametric Statistics
►Chi-Square Test: Analysis of Contingency Tables
►Fisher Exact Test
►Kendall's Tau
►Kolmogorov-Smirnov Test
►Mood Test
►Multivariate Rank Procedures: Perspectives and Prospectives
►Multivariate Rank Procedures: Perspectives and Prospectives
►Nonparametric Density Estimation
►Nonparametric Estimation
►Nonparametric Estimation Based on Incomplete Observations
►Nonparametric Models for ANOVA and ANCOVA Designs
►Nonparametric Predictive Inference
►Nonparametric Rank Tests
►Parametric Versus Nonparametric Tests
►Permutation Tests
►Randomization Tests
►Robust Statistics
►Sign Test
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Inference
►Statistical Inference: An Overview
►Statistics on Ranked Lists
►Wilcoxon–Mann–Whitney Test
►Wilcoxon-Signed-Rank Test

References and Further Reading

Bradley JV (1968) Distribution-free statistical tests. Prentice-Hall, Englewood Cliffs
Gibbons JD, Chakraborti S (2010) Nonparametric statistical inference, 5th edn. Taylor & Francis/CRC Press, Boca Raton
Kendall MG, Gibbons JD (1990) Rank correlation methods. 5th edn. Edward Arnold, London
Noether GE (1967) Elements of noparametric statistics. Wiley, New York
Pratt JW, Gibbons JD (1981) Concepts of nonparametric theory. Springer, New York
Savage IR (1962) Bibliography of nonparametric statistics. Harvard University Press, Cambridge
Walsh JE (1962) Handbook of nonparametric statistics, I: Investigation of randomness, moments, percentiles, and distribution. Van Nostrand, New York
Walsh JE (1965) Handbook of nonparametric statistics, II: Results for two and several sample problems, symmetry and extremes. Van Nostrand, New York
Walsh JE (1968) Handbook of nonparametric statistics, III: Analysis of variance. Van Nostrand, New York

Non-probability Sampling Survey Methods

H. Öztaş Ayhan
Professor and Head
Middle East Technical University, Ankara, Turkey

Non-probability sampling is generally used in experimental or trial research and does not represent the target population. Non-probability sampling uses subjective judgement and utilizes convenient selection of units from the

population. Non-probability sampling methods produce cost savings for personal interview surveys; the resulting samples often look rather similar to probability sample data (Fowler 2002). There are several non-probability selection methods that are used in practice. We will briefly overview these methods in the following sections.

Convenience (Haphazard) Sampling

The sample is composed of conveniently accessible persons who will contribute to the survey. Samples of volunteer subjects should be included here. It is used in many sciences that display little care about the representativeness of their specimens (Kish 1995).

Purposive (Judgement) Sampling

In purposive sampling, the participants are selected by the researcher subjectively. The selection is based on the judgement of the researcher. Respondents are not selected randomly but by using the judgement of the interviewers. Consequently, there is an unknown probability of inclusion for any selected sample unit.

Expert Choice

Expert choice is a form of purposive or judgment sampling, which is used by experts to pick "typical" or "representative" specimens, units, or portions. Experts often hold differing views on the best way to choose representative specimens or to decide which are the most representative units. Sometimes the researcher asks that, instead of a real population, a hypothetical universe be postulated as the parent of the sample (Kish 1995).

Snowball Sampling

A more specialized type of purposive sample is a *snowball sample*, in which respondents are asked to suggest more respondents. Snowball samples are non-probability samples, since it is impossible to know the probability with which any person in the larger population ends up in the sample (Weisberg 2005). When an incomplete list of a special population is available, enumerators evaluate these units. Later, the enumerator enquires about a possible nearby address that has the same characteristic as the respondent. Then, this address is also visited and enumerated as a sample unit. The accumulation of the additional addresses to the existing list resembles the expansion of a snowball, which is rolling downhill.

Quota Sampling

Quota sampling is a form of purposive sampling, where the enumerators are instructed to obtain specific quotas from which to build a sample roughly proportional to the population on a few demographic variables (Kalton 1983; Kish 1995). Quota sampling is a method of stratified sampling in which the selection within strata is non-random. The sample selection method employs controls (*independent* or *interrelated*) and fixed quotas as alternatives. The statisticians have criticized the method for its theoretical weakness, while market and opinion researchers have defended it for its cheapness and administrative convenience (Moser and Kalton 1979).

Mobile Population Sampling

Mobile population sampling is one of the very specific areas in which special non-probability sampling techniques are widely used. One of the main reasons for this is that it is not possible to obtain a list of the mobile target population, for probability sampling. The second important reason is the mobility of the population elements, which makes it difficult to access as a sample unit. The use of list sampling is not a possibility, while the use of specified area sampling is also difficult for some types of mobile units (e.g., birds, fishes, etc.) due to unrecognized boundary concept of the defined population areas. The mobile population sampling can be examined for several different types of populations. Basically, we can summarize these as *mobile animal populations* and *mobile human populations*.

Sampling from Mobile Animal Populations

The mobile animals can be examined as types of *wild animals, birds*, *insects*, or *fishes*. Sometimes it is possible to create protected boundaries for some of these animals. In this case, it will be possible to examine these animals by using a non-probability selection method. In other cases, there may be no defined boundary for their evaluation. In both cases, it may not be possible to use a probability sampling. Therefore one of the common methods for this enumeration is the use of *capture-tag-recapture method*. The animal has to be caught first, and measured, then labeled with a tag (*mechanical* or *electronic*) and then let go. The second stage is based on the possibility of catching the same labeled animal for another periodic measurement. The method continues until the desired round of planned measurements on the selected units are achieved. The total population is estimated from the proportion in the capture of individuals that have been previously captured and tagged (Kish 1995).

Sampling from Mobile Human Populations within Small Areas

Determining the number of people in a crowded street by ordinary methods would require the demarcation of

a number of small areas in the street and counting the number of people on each of these areas, known as *grid sampling*. Equally it is no use stationing observers at fixed points with instructions to count passers-by is another approach to enumerate the mobile human population, which the method is known as the use of *stationary observers*. These difficulties can be overcome by using *moving observers* instead of stationary observers. To obtain an estimate of the number of people in a street, the observer traverses the street in one direction, counting all the people he passes, in whichever direction they are moving, and deducing all the people who overtake him. Crowded streets were dealt with by teams of two or more observers, moving down the street in a transverse line, with each observer counting the people between him and the next observer (Yates 1960).

Sampling of Nomadic Tribes

Nomadic tribes usually live in unpopulated rural areas. Generally, they deal with animal husbandry and are travelling with their animals through a specified route over several months (Ayhan and Ekni 2003). Due to their mobility, it is not possible to observe and enumerate them by stationary observers. Therefore, moving observers are recommended to be used to enumerate this special population for censuses (see ►Census) as well as surveys.

Sampling of Homeless People

Most sample-based studies of the homeless are basically designed to measure characteristics of the currently homeless. The main challenge for designing such samples is that there typically is no fixed reference on which to base a sampling frame (Sumner et al. 2001). Due to several difficulties related to listing of homeless people within a perfect sampling frame, many homeless studies in the past were based on the non-probability samples. Recently, multistage probability samples have been used for alternative estimation.

Non-probability Sampling from Internet Users

Couper (2000) differentiates three types of non-probability Web surveys that are basically availability samples: *Web surveys as entertainment, self-selected Web surveys*, and *volunteer panels of internet users*. In addition, Ayhan (2005) had proposed a sample weighting adjustment methodology for the non-probability sample that is selected from voluntary participation Web surveys. The future of Web surveys is heading towards voluntary panels of internet surveys, which requires generalizations from the selected non-probability samples.

The use of non-probability sampling methods in practice has also been reviewed by Weisberg (2005), and the following points have been emphasized. First, the usual statistical advice is to avoid non-probability samples. Second, many surveys either explicitly use non-probability sampling or do so implicitly. Third, the balancing of survey errors and survey costs can sometimes justify non-probability sampling, regardless of the usual textbook injunctions against it.

About the Author

Dr. Öztaş Ayhan is a Professor and Head of the Department of Statistics at Middle East Technical University, Ankara, Turkey. He is also the Director of the Statistics Program of the Graduate School. During 1990–1992 period, he served as the Director of Technical Services Department of the State Institute of Statistics. In 1993, he has received the distinguished service medal of the State Institute of Statistics. Professor Ayhan has jointly written and also edited 12 books and research monographs in the fields of social and health statistics. He has published over 40 research articles in the fields of survey sampling and survey methodology. He is an Elected Member of the International Statistical Institute, since 1995. He has been the Country Representative for Turkey of the International Association of Survey Statisticians. From 1997 to 2001, he has also served as a Council Member of the International Association of Survey Statisticians. He has been the Vice President of the Turkish Statistical Association (2007–2009).

Cross References

►Internet Survey Methodology: Recent Trends and Developments
►Representative Samples
►Sample Survey Methods
►Simple Random Sample
►Small Area Estimation
►Sociology, Statistics in
►Statistical Ecology
►Statistical Inference in Ecology

References and Further Reading

Ayhan HÖ (2005) Sample adjustment weights for internet surveys: restricted access versus voluntary participation. In: Dijkum C, Blasius J, Durand C (eds) Recent developments and applications in social research methodology. Barbara Budrich Publishers, Leverkusen-Opladen, pp 1–8

Ayhan HÖ, Ekni S (2003) Coverage error in population censuses: the case of turkey. Surv Methodol 29(2):155–165

Couper MP (2000) Web surveys: a review of issues and approaches. Public Opin Q 64:464–494

Fowler FJ (2002) Survey research methods, 3rd edn. Applied Social Research Methods Series, Vol. 1. Sage Publications, Thousand Oaks
Kalton G (1983) Introduction to survey sampling. Quantitative Applications in the Social Sciences Series No. 35. Sage Publications, Thousand Oaks
Kish L (1995) Survey sampling. Wiley, New York
Moser CA, Kalton G (1979) Survey methods in social investigation, 2nd edn. Heinemann Educational Books, London
Sumner GC, Andersen RM, Wenzel SL, Gelberg L (2001) Weighting for period perspective in samples of the homeless. Am Behav Sci 45(1):80–104
Weisberg HF (2005) The total survey error approach: a guide to the new science of survey research. University of Chicago Press, Chicago
Yates F (1960) Sampling methods for censuses and surveys, 3rd edn. Hafner Publishing Company, New York

Nonresponse in Surveys

Jelke Bethlehem
Professor, Senior Survey Methodologist at Statistics Netherlands
University of Amsterdam, Amsterdam, Netherlands

Our complex society experiences an ever growing demand for statistical information relating to social, demographic, industrial, economic, financial, political, and cultural situation of the country. Such information enables policy makers and others to take informed decisions for a better future. Sometimes, statistical information can be retrieved from administrative sources. More often there is a lack of such sources. In this case, the sample survey is a powerful instrument to collect new statistical information.

A sample survey collects information on a small part of the population. In principle, this sample only provides information about the selected elements of the population. Nevertheless, if the sample is selected using a proper sampling design, it is also possible to make inference about the population as a whole.

Many things can go wrong when carrying out a survey. There are all kinds of phenomena that may have a negative impact on the quality of the survey outcomes. Nonresponse is one such problem.

Nonresponse is the phenomenon that elements in the selected sample do not provide the requested information, or that the provided information is unusable. Nonresponse comes in two forms. *Unit nonresponse* denotes the situation in which all requested information on a sampled population element is missing. If information is missing on some items in the ▶questionnaire only, it is called *item nonresponse*.

Due to non-response the amount of collected data is smaller than expected. This problem can be taken care of by selecting a larger sample. A more serious problem is that estimates of population characteristics may be biased. This situation occurs if, due to non-response, some groups in the population are over- or underrepresented, and these groups behave differently with respect to the characteristics to be investigated. Consequently, wrong conclusions may be drawn from the survey results. Therefore, it is vital to reduce the amount of nonresponse in the fieldwork of the survey as much as possible.

Nonresponse rates have been rising over the years, see Bethlehem (2009). Currently, these rates often exceed 50%. So no information is obtained from more than halve of the sampled elements. Some case studies have shown that nonresponse may result in serious biases. In practice, it is very difficult to asses the possible negative effects of nonresponse, due to lack of information on nonrespondents. And even if such effects can be detected, it is no simple matter to correct for them.

To be able to do something, it is very important to have auxiliary information. For example, if there is an auxiliary variable that has been measured in the sample, and for which population characteristics are known, then this variable can be used to check whether the available data show unbalancedness, i.e., they are not representative for the population.

There are several techniques available to correct for non-response bias. The usual treatment for unit nonresponse is adjustment weighting, and for item nonresponse imputation techniques (see ▶Imputation) can be applied.

The basic principle of *adjustment weighting* is that every observed element is assigned a specific weight. Estimates for population characteristics are obtained by processing the weighted values instead of the values themselves. The easiest and most straightforward method used to compute weights is post-stratification. The population is divided into strata after selection of the sample. If each stratum is homogeneous with respect to the target variable of the survey, then the observed elements resemble the unobserved elements. Therefore, estimates of stratum characteristics will not be very biased, so they can be used to construct population estimates.

To take as much advantage as possible of the available auxiliary information also more advanced weighting methods have been developed, like *linear weighting* (see

Bethlehem and Keller 1987) and *calibration estimation* (see Deville and Särndal 1992).

To carry out adjustment weighting auxiliary variables are needed and preferable auxiliary variables having a strong relationship with the target variables of the survey. In practical situations, the available information may be limited, and also the relationships may not be as strong as preferred. Therefore, sometimes adjustment may not be capable of completely removing the bias.

A fortunate development is that a tendency that more and more data from registers can be used. Such registers contains more useful auxiliary information, and therefore correction procedures based on these variables will be more effective.

Item non-response requires a different approach. A great deal of additional information is available for the elements involved. All available responses to other questions can be used to predict the answer to the missing questions. This computation of a "synthetic" answer to a question is called *imputation*. Predictions are usually based on models describing relationships between the variable with missing values and other variables for which the values are available.

Imputation techniques can be classified as random or deterministic, depending on whether residuals are set to zero or not. Deterministic techniques have the disadvantage that they distort the properties of the distribution of the values of the variable. These techniques tend to predict values in the middle part of the distribution. Therefore, standard errors computed from the imputed data set are generally too small. They create a too optimistic view of the precision of estimates. Random imputation methods do not have this nasty property. They much better able to preserve the original distribution. However, they introduce an extra source of variability.

Another point to take into consideration is the effect of imputation on relationship between variables. Several imputation techniques cause estimates of covariances and correlations to be biased. The estimates produce too low values. For more information, see e.g., Kalton and Kasprzyk (1986).

Research for new imputation techniques is in progress. An example is the multiple imputation technique proposed by Rubin (1979). This technique computes a set of imputed values for each missing value. This results in a number of imputed data sets. Inference is based on the distribution of estimates obtained by computing the estimate for each data set. This approach has attractive theoretical properties, but may not be so easy to implement in practical situations.

An overview of the state of the art in nonresponse research can be found in Groves et al. (2001).

About the Author

Dr. Jelke Bethlehem joined Statistics Netherlands in 1978, first as Research Worker and later as Senior Statistician. His main topics were the treatment of nonresponse in sample surveys, in which he obtained his Ph.D., and disclosure control of published survey data. From 1987 to 1996 he was head of the Statistical Informatics Department, which concentrated on the development of standard software for processing survey data. Currently, he is Senior Advisor of the Department of Statistical Methods of the Central Bureau of Statistics (CBS) Netherlands. He has been the principal investigator of a number of research projects in the area of survey methodology and nonresponse. Dr. Bethlehem is also part-time professor in Statistical Information Processing at the University of Amsterdam. He was Vice-President of the International Association of Survey Statisticians (IASS), from 2005 to 2007. Professor Bethlehem has published widely in a number of journals and books in survey methodology.

Cross References

►Federal Statistics in the United States, Some Challenges
►Imputation
►Maximum Entropy Method for Estimation of Missing Data
►Multi-Party Inference and Uncongeniality
►Multiple Imputation
►Nonresponse in Web Surveys
►Nonsampling Errors in Surveys
►Public Opinion Polls
►Representative Samples
►Sampling From Finite Populations
►Total Survey Error

References and Further Reading

Bethlehem JG (2009) Applied survey methods – a statistical perspective. Wiley, Hoboken, NJ, USA
Bethlehem JG, Keller WJ (1987) Linear weighting of sample survey data. J Offici Stat 3:141–153
Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. J Am Stat Assoc 87:376–382
Groves RM, Dillman DA, Eltinge JL, Little RJA (eds) (2001) Survey Nonresponse. Wiley, New York, USA
Kalton G, Kasprzyk D (1986) The Treatment of Missing Survey Data. Surv Methodol 12:1–16
Rubin DB (1979) Illustrating the use of multiple imputations to handle non-response in sample surveys. Bulletin of the International Statistical Institute. 48, Book 2, pp 517–532

Nonresponse in Web Surveys

Katja Lozar Manfreda, Nejc Berzelak, Vasja Vehovar
Faculty of Social Sciences
University of Ljubljana, Ljubljana, Slovenia

As in other types of surveys, we can distinguish between *unit nonresponse*, where the expected (invited) eligible units do not participate in the survey, and *item nonresponse*, where the units participate, but certain responses are missing to some questions.

In addition, due to the availability of paradata – i.e., data from the process of answering a Web survey ►questionnaire that are stored in the log files together with the responses (Couper 2005) – a variety of *additional nonresponse measures and patterns* may be detected in Web surveys (Bosnjak 2000). In particular, *partial nonresponse*, where respondents answer only part of the questionnaire and then drop out is an issue requiring further discussion.

Here we predominantly focus on unit and partial nonresponse, while leaving the item nonresponse aside because it is not so much a problem of participation in Web surveys, but is more a problem related to Web questionnaire design issues, which is not our focus here.

When discussing the amount of unit and partial nonresponse in Web surveys, it is important to distinguish between different types of Web surveys (Couper 2001). The level of nonresponse can be precisely assessed only for list-based Web surveys, that is, surveys where a list of sample members (either probability or non-probability) are invited individually, giving them a unique URL and/or username/password to answer the questionnaire. Only in this case is it possible to distinguish between crucial response outcomes and calculate required response rates. For example, the AAPOR (AAPOR 2009) lists the following disposition codes for Internet surveys of specifically named persons: complete and partially complete questionnaires, eligible but with nonresponse (either due to explicit or implicit refusals and break-offs, non-contacts, or some other reason), cases with unknown eligibility, and not eligible. This coding is rather difficult, especially if e-mail invitations to the survey are used. For those units that do not respond (do not even log on to the first questionnaire page) it can rarely be established whether they are eligible or whether they actually received the invitation (whether they were actually contacted). The cases with unknown eligibility and those where it cannot be determined if contact has been made occur much more frequently in Web surveys than in any other survey modes.

In contrast to the list-based Web surveys, there are also unrestricted self-selected Web surveys where participants become aware of the questionnaire via word of mouth, advertisements, or serendipity. In this case, any response rate is difficult or impossible to calculate due to the unknown target population (thus an unknown denominator).

Another special aspect of Web surveys are surveys where pre-recruited lists of Internet users or of the general population (usually panels) are invited to the survey. In this case, two types of responses are of interest here: one is cooperation in the pre-recruitment stage and inclusion in the sampling frame, the other is cooperation for a particular Web survey project.

Despite the different types of Web surveys and difficulties in distinguishing between some disposition codes, we can define the following response outcome rates for Web surveys:

- *Response rate* and *completion rate*: The response rate as a ratio of completed questionnaires to the number of eligible units (AAPOR 2009) can be calculated for list-based Web surveys. However, since we often do not have information about the eligibility of the units (unless we somehow estimate their share or simply assume that all invited units are eligible), a completion rate may be a better measure. It is defined as the number of completed questionnaires among all invitations sent. In addition, if a distinction between partial (questionnaire answered only partially, then a drop-out) and complete respondents is made, overall (referring to partial and complete respondents) and full (referring to complete respondents only) response and completion rates can be defined.
- *Absorption rate* and *failure rate* (Vehovar et al. 2002): The absorption rate can be defined for list-based Web surveys as the number of absorbed or delivered (email, mail, fax) invitations among all invitations sent (delivered and undelivered). Its opposite, the failure rate (sometimes also called the attrition rate), is defined as the number of undelivered invitations (known as non-contacts) among all invitations sent. These rates are indicators of the quality of the sampling frame. The absorption rate is similar to the contact rate, and is defined as the ratio of contacted units to the number of eligible units (AAPOR 2009). However, the absorption rate is a more robust measure when the exact number of eligible units cannot be established, which is often the case in Web surveys.

- *Cooperation rate*. The cooperation rate is defined as the ratio of interviews obtained to the number of units contacted (AAPOR 2009). For list-based Web surveys, an approximate measure of the cooperation rate can be calculated if we assume that all units with absorbed (e-mail, mail, fax) invitations are also contacted units. Again, we can distinguish between the overall and full cooperation rates, depending on complete and partial respondents.
- *Refusal rate*. For list-based Web surveys, the refusal rate, as a ratio of the units that explicitly refuse to participate to the number of eligible units (AAPOR 2009), can also be calculated. In this case, those who explicitly state that they do not want to participate and/or want to be excluded from the list are counted as explicit refusals, while eligibility is again usually only estimated.
- *Click-through rate*. This is a special measure that is specific to Web surveys and can be reported for any type of Web surveys, including unrestricted self-selected ones. Namely, it refers to respondents who click on the URL link and access the Web questionnaire. For example, when banner ads or some other publication of the URL is used, the click-through rate is defined as the percentage of those clicking on the banner ad/invitation and going to the questionnaire page among those exposed to the ad/invitation. For list-based Web surveys, the click-through rate is defined as the percentage of invited units who access the first questionnaire page. In general, we can define the click-through rate for all Web surveys where the number of respondents exposed to the invitation can be established (regardless of whether or not they were actually contacted, and regardless of whether or not they are eligible). In special cases, the cooperation click-through rate can be defined as the number of units accessing the Web questionnaire among all contacted, eligible, and available during data collection.
- *Drop-out* and *complete rate*. The drop-out rate (sometimes called the attrition rate) refers to respondents who started to answer the Web questionnaire, but abandoned it prematurely. It refers to the percentage of respondents who only partially complete the questionnaire among all respondents. The opposite – complete rate – is defined as the percentage of complete responses among all respondents. The manner in which partial respondents are distinguished from complete respondents depends on the particular survey. In general, partial respondents can be answering drop-outs, item nonresponding drop-outs with low item nonresponse rate for the displayed questions, and/or item nonrespondents with item nonresponse large enough not to include them among complete respondents.

The above rates for Web surveys can be very variable. The articles published in journals in 2009 included in the Web of Science database and reporting the results of some Web surveys, show the response rates vary from less than 20% to close to 100%. (The Web of Science database returns 1,718 unique records (articles) for 2009 when entering the following keywords: Web surveys, Web questionnaire, Internet survey, Internet questionnaire, online survey, and online questionnaire. After reading abstracts of a ▶simple random sample of 100 of these records, we estimate that 25% of these articles actually do not report on Web surveys. Among the remaining ones, for 48% of the articles not enough data was provided to estimate a response rate or they report on an unrestrictred, self-selected Web survey.) Unfortunately, no systematic study was recently done to give an overview of response rates. However, a meta-analysis of experimental studies comparing response rates of Web vs. some other survey mode (Lozar Manfreda et al. 2008) showed that Web surveys on average yield an 11% lower response rate than other survey modes, when comparable samples and procedures are used. This is alarming if we assume that the level of nonresponse is also an indicator of nonresponse bias. However, since this is not necessarily the case (Groves and Peytcheva 2008), further investigations into how Web respondents differ from nonrespondents are needed and some studies on this topic have already been done (for example Biffignandi and Pratesi 2000; Taylor et al. 2009; Sax et al. 2008).

In general, the *response rate* in Web surveys *depends on different factors*, such as the sociological and technological environment, survey design, and characteristics of the respondent (Vehovar et al. 2002).

The social environment indirectly affects participation in Web surveys through general economic development, telecommunication policy, educational system, technological tradition, etc. In addition, the general survey climate, perception of direct marketing, legitimacy of surveys and their sponsors, data protection scandals, and opinion leaders also influence participation. As regards Web surveys with email invitations, the increasing problem of spam and viruses negatively influence participation to a large degree (Fan and Yan 2009). Email invitations may either be considered spam, and as such ignored or not even delivered (in some email systems they are automatically intercepted by spam filters), or people do not want to click on the URL address in the invitation due to the fear of viruses. The critical issue regarding privacy on the Internet may also

negatively affect participation in Web surveys since Internet users are reluctant to reveal personal information if conditions of its use are not clearly specified.

The social environment interacts closely with the technological environment. General telecommunication and information infrastructure is especially critical for successful implementation of Web surveys. Considerable differences can be observed with respect to Internet penetration in various countries. By the year 2009, Internet penetration surpassed 70% of the active population only in some developed countries such as the US, Canada, some European countries (e.g., Faroe Islands, Switzerland, Luxemburg, Andorra, UK, Spain), or certain Asian countries (e.g., Japan, South Korea), and 80% only in a few countries (e.g., Greenland, Iceland, Norway, Finland, Netherlands, Sweden, Australia, New Zeeland). In general, while the developed world has reached a high penetration rate, in some of the largest countries (like China, Russia, and India) as well as in the remaining developing countries, penetration is still below 25% (Internet World Stats 2009). In addition to the problem of non-coverage, the quality of Internet infrastructure (speed of Internet connections) and cost for Internet connections additionally affects participation in Web surveys.

Regarding the respondents' characteristics, response rates to surveys usually vary across social-demographic categories, survey experience, interest in the survey topic, and other attitudes. In addition, for Web surveys, computer literacy and the respondent's orientation towards computer use also become extremely important. Actually, the intensity of computer and Internet usage is the most important predictor of cooperation in a Web survey, even when observed within the social-demographic categories defined by age, gender, education, and income (Batagelj and Vehovar 1998; Kwak and Radler 2002; Vehovar et al. 1999). Of course, when computer orientation is not controlled, it appears that the usual characteristics of Internet usage also determine the participation in Web surveys: respondents are younger, more educated, richer, and male (Bandilla et al. 2003; Batagelj and Vehovar 1998; Braungberger et al. 2007; McCabea et al. 2006). In addition, the respondent's technical equipment may also affect Web survey participation with technologically advanced users being more willing (or able) to participate.

Although the above influences on web survey participation are of extreme importance, researchers cannot do much to change them. On the other hand, the design of a Web survey is becoming a crucial factor by which survey researchers can influence the readiness of people to participate. Several studies have been done to develop implementation procedures that would result in high quality data from Web surveys. It has also been shown that response rates in Web surveys can be increased particularly by incentives (e.g., Göritz 2006), as well as by mixed-modes (e.g., Dillman et al. 2009; Greene et al. 2008; Kroth et al. 2009), and increased number of contacts with respondents.

In addition to *developing techniques for increasing response rates*, another stream of measures address the *post-survey adjustments*. These techniques not only compensate for nonresponse, but also the non-coverage, frame problems and non-probability nature of the sample. Besides standard weighting and imputation strategies, propensity score adjustment is frequently used in the context of Web surveys (e.g., Lee 2006; Lee and Valliant 2009; Schonlau et al. 2009).

As survey response rates have been declining over the years in general (de Leeuw and de Heer 2002; Grapentine 2008; Stoop 2005) and as Web surveys gain even lower response rates than other survey modes (Lozar Manfreda et al. 2008), the predictions for the future are rather negative. We can expect even less willingness to participate in Web surveys. Therefore, the development of measures to increase response rates (particularly incentives) seems to be crucial. In addition, the increased use of probability-based Internet panels (e.g., Knowledge Network in U.S.A., Centerdata panel in the Netherlands), as well as numerous non-probability access panels, seem to be gaining momentum. This will be emphasized even more due to lower costs of Web survey data collection in comparison to other survey modes. This forces survey organizations to consider this mode of data collection, regardless of the problem of non-coverage and higher levels of nonresponse.

About the Authors

Katja Lozar Manfreda, PhD, is an assistant professor of statistics and Head of the Department of Informatics and Methodology (since 2009), Faculty of Social Sciences, University of Ljubljana, Slovenia. She is the General Secretary of the RC33 committee on Logic and Methodology of the International Sociological Association. She has (co-)authored over 15 scientific papers and book chapters. Dr. Lozar Manfreda has received the AAPOR Warren J. Mitofski Innovators Award in 2009 for the WebSM site. She was a guest editor for a special issue of Journal of Official Statistics in 2006 and she has been a member of the editorial board of the Survey Research Methods since 2007.

Nejc Berzelak Nejc Berzelak is a teaching assistant of social science methodology and researcher, University of

Ljubljana, Slovenia. He is an editor of the WebSM portal, website devoted to web survey methodology.

Dr. Vasja Vehovar is a full professor of Statistics at the Faculty of Social Sciences, University of Ljubljana, Slovenia. From 1996 he has been the principal investigator of the national project Research on Internet in Slovenia (RIS), the leading source for information society research in Slovenia. He is also involved in various EU information society research projects. He is responsible for development of the WebSM portal devoted to web survey methodology and was the coordinator of the corresponding EU Framework project. In 2009, he has received the AAPOR Warren J. Mitofski Innovators Award for the WebSM website, together with Katja Lozar Manfreda. His research interests span from survey methodology to information society issues.

Cross References

►Nonresponse in Surveys
►Nonsampling Errors in Surveys
►Representative Samples
►Sampling From Finite Populations

References and Further Reading

Bandilla W, Bosnjak M, Altdorfer P (2003) Survey administration effects? A comparison of Web-based and traditional written self-administered surveys using the ISSP Environment module. Soc Sci Comput Rev 21(2):235–243

Batagelj Z, Vehovar V (1998) WWW Surveys. In: Ferligoj A (ed) Advances in methodology, data analysis, and statistics. Faculty of Social Sciences, Ljubljana, Slovenia, pp 209–224

Biffignandi S, Pratesi M (2000) The respondents profile in a Web survey on firms in Italy, Paper presented at the International Conference on Methodology and Statistics, Preddvor, Slovenia

Bosnjak M (2000) Participation in non-restricted web surveys: A typology and explanatory model for item non-response. Paper presented at the 55th American Association for Public Opinion Research Annual Conference, Portland, OR

Braungberger K, Wybenga H, Gates R (2007) A comparison of reliability between telephone and web-based surveys. J Business Res 60:758–764

Couper MP (2001) Web surveys: A review of issues and approaches. Public Opin Q 64(4):464–494

Couper MP (2005) Technology trends in survey data collection. Soc Sci Comput Rev 23(4):486–501

de Leeuw ED, de Heer W (2002) Trends in household survey nonresponse: A longitudinal and international comparison. In: Groves RM, Dillman DA, Eltinge JL, Little RJA (eds) survey nonresponse. Wiley, New York, NY, pp 41–54

Dillman DA, Phelps G, Tortora R, Swift K, Kohrell J, Berck J et al (2009) Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. Soc Sci Res 38(1):1–18

Fan W, Yan Z (2009) Factors affecting response rates of the web survey: A systematic review. Comput Hum Behav 26(2):132–139

Göritz AS (2006) Incentives in Web studies: Methodological issues and a review. Int J Internet Sci 1:58–70

Grapentine T (2008) Top concerns for our industry. Market Res 20(2):4

Greene J, Speizer H, Wiitala W (2008) Telephone and web: Mixed-mode challenge. Health Serv Res 43(1):230–248

Groves RM, Peytcheva E (2008) The impact of nonresponse rates on nonresponse bias. Public Opin Q 72(2):167–189

Internet World Stats (2009) Internet Usage World Stats – Internet and Population. Retrieved from Internet World Stats website http://www.internetworldstats.com/

Kroth PJ, McPherson L, Leverence R, Pace W, Daniels E, Rhyne RL et al (2009) Combining web-based and mail surveys improves response rates: A PBRN study from PRIME Net. Ann Fam Med 7(3):245–248

Kwak N, Radler B (2002) A comparison between mail and web surveys: Response pattern, respondent profile and data quality. J Off Stats 18(2):257–273

Lee S (2006) Propensity score adjustment as a weighting scheme for volunteer panel web surveys. J Off Stat 22(2):329–349

Lee S, Valliant RL (2009) Estimation of volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociolo Methods Res 37(3):319–343

Lozar K, Bosnjak M, Berzelak J, Haas I, Vehovar V (2008) Web surveys versus other survey modes: A meta-analysis comparing response rates. Int J Market Res 50(1):79–104

McCabea SE, Diezb A, Boydc CJ, Nelsond TF, Weitzmand CJER (2006) Comparing web and mail responses in a mixed mode survey in college alcohol use research. Addict Behav 31(9): 1619–1635

Sax LJ, Gilmartin SK, Hagedorn LS, Lee JJ (2008) Using web surveys to reach community college students: An analysis of response rates and response bias. Comm College J Res Pract 32(9):712–729

Schonlau M, van Soest A, Kapteyn A, Couper MP (2009) Selection bias in web surveys and the use of propensity scores. Sociolog Methods Res 37(3):291–318

Stoop I (2005) *The hunt for the last respondent: Nonresponse in sample surveys*. The Hague, the Netherlands: Social and Cultural Planning Office of the Netherlands

Taylor PA, Nelson NM, Grandjean BD, Anatchkova B, Aadland D (2009) *Mode effects and other potential biases in panel-based Internet surveys: Final report.* Retrieved from University of Wyoming website http://yosemite.epa.gov/ee/epa/eerm.nsf/vwAN/EE-0519-01.pdf//EE-0519-01.pdf

The American Association for Public Opinion Research (2009) *Stan-ll dard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 6th edition*. Retrieved from AAPOR website http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions1&Template =/CM /ContentDisplay.cfm & ContentID = 1814

Vehovar V, Batagelj Z, Lozar K (1999) Web surveys: Can the weighting solve the problem? Paper presented at the the Section on Survey Research Methods, Alexandria

Vehovar V, Lozar K, Zaletel M, Batagelj Z (2002) Survey nonresponse. In: Groves RM, Eltinge JL, Little RJA (eds) Survey nonresponse New York. Wiley, NY, pp 229–242

Nonsampling Errors in Surveys

Judith M. Tanur
Distinguished Teaching Professor Emerita
Stony Brook University, Stony Brook, NY, USA

The term "nonsampling errors" is usually employed to refer to all the problems that can occur in sample surveys with the exception of sampling error (which is the inherent variability of estimates from sample to sample (See the entry Sampling distributions). Mosteller (1978), in a classic article that is outdated in detail but still very much worth reading, uses the term more broadly so that it refers to experiments as well as surveys. In his usage it includes such things as the Hawthorne Effect and bad breaks in random assignment; the usage here will be confined to sample surveys.

Broadly, nonsampling error can be broken down into five categories as shown in Fig. 1: specification errors, coverage errors, nonresponse, response errors, and processing errors. In this article these types of errors will be discussed separately, but the reader should note that in reality the issues overlap and interact – for example, a question that tends to produce confusion in a mail survey may work perfectly well in person. In addition, a brief article such as this cannot address all possible nonsampling errors that are discussed in the literature. Please see the references at the end of the article for fuller treatments, some older and some newer; see also the article on Total Survey Error.

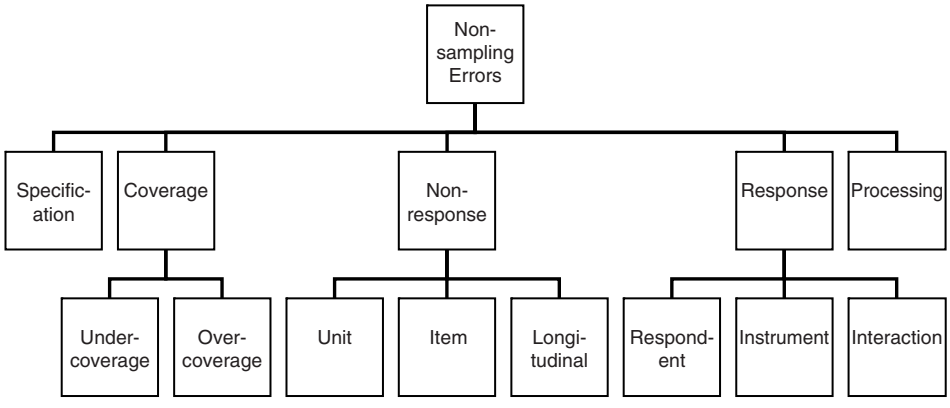
Specification Errors

Specification errors occur in the planning stage of a survey if researchers prepare an instrument that does not collect the data necessary to meet the objectives of the study. Leaving out a question the answer to which would be crucial to the analysis would be an example.

Coverage Errors

Undercoverage errors occur when members of the target population are missing from the frame from which the sample is drawn. Early in the twentieth century, for example, sample surveys in the United States had to be done via in-person interviewing or by mail because telephone penetration was sufficiently small as to make telephone numbers a frame that suffered from undercoverage in representing the US population. As telephone penetration grew to close to 100% towards the end of the twentieth century, telephone interviewing (especially random digit dialing using CATI, computer-assisted telephone interviewing) became the method of choice for many governmental and non-governmental surveys. Now as more and more households replace their landlines with cell phones, methods using random digit dialing sampling again suffer from undercoverage, and methods are being devised to reach the cell-phone only population. (This is especially difficult in the United States where there are legal prohibitions against using automated dialers to call cell phones.) A turn toward what has been dubbed "address-based sampling" is one current solution to these undercoverage problems – the method uses post-office address lists (which are usually quite complete) and then reaches households either via phone if they can be traced through reverse directories, via mail, or as a last resort because of the expense involved, in person.

Overcoverage errors occur when the frame includes individuals who are not members of the target population, as when those who will not vote in an election are interviewed about their voting intentions and their answers



Nonsampling Errors in Surveys. Fig. 1 Conceptualization of nonsampling errors

included in estimates of candidates' popularity with voters. Survey researchers have devised careful questioning and weighting procedures (most of them proprietary) to deal with this problem of overcoverage in pre-election surveys. The issues are less straightforward in other types of surveys.

Nonresponse Errors

If a unit from the population designated to be in the sample cannot be reached or refuses to participate in the survey, then what is called unit nonresponse occurs. To avoid unit nonresponse survey researchers make repeated call-backs and often offer incentives to participate. Nonresponse *error* occurs to the extent that nonresponders are different from responders and hence estimates of the quantities of interest in the survey are biased. Often a subsample of nonresponders is pursued with special efforts, both to reduce unit nonresponse and to estimate nonresponse error on the theory that hard-to-reach respondents are more similar to nonrespondents than are other respondents.

If a unit from the population responds to the survey but leaves out answers to some of the questions, those omissions are referred to as item nonresponse. When there is item nonresponse the editing process for a survey often "fills in" those blanks by a process known as ▶imputation. (Imputation is also sometimes used for unit nonresponse.)

When a survey is longitudinal, a so-called panel survey, in which respondents are questioned several times over a period of months or years, there is a special kind of nonresponse when a respondent who has given answers during previous rounds of the survey cannot be contacted or refuses to answer in the current round. Survey researchers who carry out longitudinal surveys put in place careful procedures to keep in touch with respondents to help avoid this kind of panel attrition.

Response Errors

When a respondent gives an answer to a survey question, but that answer is "wrong," a response error occurs.

Of course, respondents can simply lie, and one of the reasons for lying is to maintain one's self presentation. Hence there is a class of so-called sensitive questions that researchers often suspect respondents will be reluctant to answer truthfully – for example, questions about illegal acts or about sexual behavior. Methods for reducing respondents' tendency to respond untruthfully to sensitive questions include introducing the question with a statement that many people actually indulge in that behavior, through arranging for the respondent to answer privately (which always is the case in a mailed survey and which can be done in-person through a sealed ballot box or by having the respondent him/herself enter responses on the computer screen out of view of the interviewer). A randomized response protocol has been developed in which the respondent does a randomization (e.g., tosses a coin) and depending on the outcome (which is unknown to the interviewer) answers the sensitive question or an innocuous one (e.g., does your social security number end in an even digit?). Since both the probability distribution for the outcome of the randomization device and that for the answer to the innocuous question are known, an estimate of the proportion of respondents admitting to the sensitive behavior can be generated. Of course, complications arise with any attempt to relate those answers to other variables.

Sometimes respondents give an incorrect answer because of memory problems. Survey researchers have developed memory aids to help solve this problem – for example, a respondent can fill in a time line in order to remember events in his/her own history. In a longitudinal study, often the events reported in the first interview are not used as data but only to "bound" the respondent's memory in the subsequent interview. In the second and subsequent interviews, the respondent is reminded about the events s/he reported up to the time of the previous interview and asked what has happened since that time. A special kind of forgetting is called "telescoping," in which the respondent reports events that happened before the start of the reference period as having happened during the reference period. Huttenlocher et al. (1990) have shown that telescoping can be a natural result of how elapsed time is stored in memory.

Instrument Errors

The differing modes of survey presentation, whether an in-person interview (using either a paper ▶questionnaire or a computer), a mailed questionnaire, a survey on the web, or a telephone interview can make a difference in how the respondent understands and answers the questions. This is a broad field of study; a current reference is Biemer and Lyberg (2003). Don Dillman and his colleagues have made deep studies of the effects of the physical form of the questionnaire and of how to design both paper surveys and computer displays to avoid errors generated by the physical layout. See, for example, Dillman et al. (2009).

The problem of wording of questions and their effect on nonsampling error was for many years more of an art than a science. In recent years some progress has been made in systematizing this knowledge and providing theoretical underpinnings. The movement to study cognitive aspects of survey methodology (CASM) was given impetus by a

seminar described the Jabine et al. (1984). It is now standard practice to do cognitive interviewing as part of the questionnaire design and testing process and thus avoid some misunderstandings as sources of nonsampling error. Useful works contributing to this part of the understanding of nonsampling errors are Tanur (1992), Tourangeau et al. (2000) and Sudman et al. (1996).

Interaction

The earliest surveys, Converse (1987) reports, were carried out by a researcher on his/her own behalf – interviewing respondents to find out what s/he wanted to know. It was an individual effort and conversations were often quite informal, with the aim of getting the information needed from the respondent and without any particular effort to standardize the questions asked. But as surveys got larger and the number of interviewers grew correspondingly large, researchers began to insist on standardization. The ideal was to have every interviewer ask every respondent the same questions in the same order using the exact same words and the same intonation. In a landmark article Suchman and Jordan (1990) showed that such standardization could sometime not only irritate respondents to the extent that they would be reluctant to continue the interview, but could also lead to such gross misunderstandings that incorrect data were recorded, thus providing another source of nonsampling error. Michael Schober and Fred Conrad and their colleagues (e.g., Conrad and Schober 2000) have experimented with conversational interviewing to try to alleviate that problem. They find, in particular, that when the respondent's situation is complicated, conversational interviewing is more likely to result in accurate data than more standardized procedures. When the respondent's situation is simple there seems to be little gain in conversational interviewing.

The race of the interviewer and of the respondent and whether they match have long been suspected as sources of nonsampling error. By the early 1980s survey researchers believed that the race of the interviewer had an effect only when the topic was itself concerned racial matters. In recent pre-election polling in the United States there was fear that a so-called Bradley effect (or Wilder effect) might occur. Such a putative effect posits that white respondents are unwilling to say that they would vote against a Black candidate when interviewed, but were willing to vote against a Black candidate in the privacy of the voting booth. There seems to be little evidence that a Bradley effect occurred in the polling the preceded the election of Barak Obama in 2008.

Processing

As survey data are edited for consistency and for imputation, errors can be introduced. Vigilance and both computer and human checking are the only ways such errors can be caught and corrected.

About the Author

Judith M. Tanur is Distinguished Teaching Professor Emerita in the Sociology Department, Stony Brook University. She has authored or co-authored over 75 papers and written or edited nine books. She has served on the Board of Directors of the American Statistical Association and received its Founders' Award (1997). She is a Fellow of the American Statistical Association (1980), the American Association for the Advancement of Science (1983), the American Psychological Society (2007), and an Elected member of the International Statistical Association (1987). She was a co-winner of the Innovators Award from the American Association for Public Opinion Research in 2005 for the creation of the Interdisciplinary Workshop on Cognitive Aspects of Survey Methodology. Professor Tanur is known as a co-editor (with William Kruskal) of the *International Encyclopedia of Statistics*, Free Press, l978.

Cross References

►Margin of Error
►Nonresponse in Surveys
►Total Survey Error

References and Further Reading

Biemer P, Lyberg L (2003) Introduction to survey process quality. Wiley, New York

Conrad FG, Schober MF (2000) Clarifying question meaning in a household telephone survey. Public Opin Q 64:1–28

Converse JM (1987) Survey research in the United States: roots and emergence, 1890–1960. University of California Press, Berkeley

Dillman DA, Smith JD, Christian LM (2009) Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method, 3rd edn. Wiley, Hoboken, NJ

Groves RM, Dillman D, Eltinge J, Little R (2002) Survey Nonresponse. Wiley, New York

Huttenlocher J, Hedges LV, Bradburn NM (1990) Reports of elapsed time: Bounding and Rounding Processes in Estimation. J Exp Psychol Learn 16:196–213

Jabine T, Straf M, Tanur J, Tourangeau R (1984) Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Report of the Advanced Research Seminar on Cognitive Aspects of Survey Methodology. National Academy Press, Washington, DC

Mosteller F (1978) Errors I: nonsampling errors. In: Kruskal WH, Tanur JM (eds) International encyclopedia of statistics. Free Press, New York, pp 208–229

Suchman L, Jordan B (1990) Interactional troubles in face-to-face survey interviews. J Am Stat Assoc 85(409):232–253

Sudman S, Bradburn N, Schwarz N (1996) Thinking about answers: the application of cognitive processes to survey methodology. Jossey-Bass, San Francisco, CA

Tanur JM (1983) Methods for large-scale surveys and experiments. In: Leinhardt S (ed) Sociological methodology 1983–1984. Josey-Bass, San Francisco, pp 1–71

Tanur JM (ed) (1992) Questions about Questions: Inquiries into the Cognitive Bases of Surveys. Russell Sage, New York

Tourangeau R, Rips LJ, Rasinski K (2000) The psychology of survey response. Cambridge University Press, Cambridge, UK

Non-uniform Random Variate Generations

Pierre L'Ecuyer
Professor, DIRO
Université de Montréal, Montréal, QC, Canada

Introduction

As explained in the entry ►Uniform Random Number Generators, the simulation of random variables on a computer operates in two steps: In the first step, uniform random number generators produce imitations of i.i.d. $U(0,1)$ (uniform over (0,1)) random variables, and in the second step these numbers are transformed in an appropriate way to imitate random variables from other distributions than the uniform ones, and other types of random objects. Here we discuss the second step only, assuming that infinite sequences of (truly) i.i.d. $U(0,1)$ random variables are available from the first step. This assumption is not realized exactly in software implementations, but good-enough approximations are available (L'Ecuyer 2004).

For some distributions, simple exact transformations from the uniform to the target distribution are available, usually based on the inversion method. But for many types of distributions and processes, in particular those having shape parameters, and multivariate distributions, one relies on approximations that require a compromise between efficiency and approximation error. That is, the sampling is not always done exactly from the target distribution, but the discrepancy between the sampling and target distributions can often be made smaller with more work. This work is generally divided in two parts: A one-time setup cost to compute constants and tables that depend on the distribution parameters, and a marginal cost for each random variate generated from this distribution. The marginal speed and also the quality of the approximation can often be improved by a larger investment in the setup time, sometimes to precompute larger tables. This investment can be worthwhile when a large number of random variates has to be generated from the same distribution, with the same shape parameters. Robustness with respect to shape parameters is another issue: Some methods provide a good approximation only in a certain range of values of these parameters, so one must be careful not to use the generator outside that range.

Generally speaking, inversion should be the preferred method whenever it is feasible and not too inefficient, because of its compatibility with important variance-reduction techniques such as common random numbers, antithetic variates, randomized quasi-Monte Carlo, etc. (Law and Kelton 2000; L'Ecuyer 2009).

Inversion is not always convenient, in particular for complicated multivariate distributions, for which the density is sometimes known only up to a multiplicative constant. One important approximation method for that situation is Markov chain Monte Carlo (MCMC) (*see* ►Markov Chain Monte Carlo Methods), which constructs an artificial Markov chain (see ►Markov Chains) whose stationary distribution is the target distribution. A random variate that follows approximately the target distribution is obtained by returning the current state after running the Markov chain long enough; *see* the entry ►Monte Carlo Statistical Methods.

In the remainder, we briefly summarize some of the most basic techniques. Detailed coverages of non-uniform variate generation can be found in (Devroye (1986; 2006) and Hörmann et al. 2004).

Inversion

A general transformation that provides a univariate random variable X having the cumulative distribution function (cdf) F from a $U(0,1)$ random variable U is $X = F^{-1}(U)$, where $F^{-1} : [0,1] \to \mathbb{R}$ is the inverse distribution function, defined as

$$F^{-1}(u) \stackrel{\text{def}}{=} \inf\{x \in \mathbb{R} \mid F(x) \geq u\}.$$

This is the *inversion method*.

As an illustration, if X is a binomial random variable with parameters $(n,p) = (2,0.4)$, then we have $P[X = i] = p_i$ where $p_0 = (0.6)^2 = 0.36$, $p_1 = 2 \times 0.6 \times 0.4 = 0.48$, $p_2 = (0.4)^2 = 0.16$, and $p_i = 0$ elsewhere. Inversion then returns $X = 0$ if $U < 0.36$, $X = 1$ if $0.36 \leq U < 0.84$, and $X = 2$ if $U \geq 0.84$. Another simple way of generating X here is to generate two Bernoulli random variables X_1 and X_2 by inversion from two independent uniforms U_1 and U_2, i.e., $X_1 = \mathbb{I}[U_1 < p]$ and $X_2 = \mathbb{I}[U_2 < p]$, where $\mathbb{I}$ is the indicator function, and return $X = X_1 + X_2$. This method requires two uniforms and is not inversion for X.

For certain distributions, there is a closed-form formula for F^{-1}. For example, if X has a discrete uniform distribution over $\{0,\ldots,k-1\}$, we have $X = F^{-1}(U) = \lfloor kU \rfloor$. If X is geometric with parameter p, so $P[X = x] = p(1-p)^x$ for $x = 0,1,\ldots$, we have $X = F^{-1}(U) = \lceil \ln(1-U)/\ln(1-p) \rceil - 1 = \lfloor \ln(1-U)/\ln(1-p) \rfloor$ with probability 1. If X is exponential, with rate λ, then $X = F^{-1}(U) = -\ln(1-U)/\lambda$. To generate X with cdf F but truncated to an interval $(a,b]$, it suffices to generate U uniform over $(F(a),F(b)]$, and to return $F^{-1}(U)$.

For certain distributions there is no closed-form expression for F^{-1} but good numerical approximations are available. For distributions having a location and a scale parameter, we only need a good approximation of F^{-1} for the standardized form of the distribution, say with location at 0 and scale 1. We generate from the standardized distribution, then multiply by the scale parameter and add the location parameter. This applies in particular to the normal distribution, for which good numerical approximations of the standard inverse cdf Φ^{-1} are available. For example, the `probdist` package of SSJ (L'Ecuyer 2008) implements a slight modification of a rational Chebyshev approximation proposed in Blair et al. (1976), which is quite fast and provides essentially machine-precision accuracy when using 64-bit floating point numbers. For any $u \in (0,1)$ that can be represented by such a floating-point number (given as input), the approximation procedure returns $\Phi^{-1}(u)$ with relative error smaller than 10^{-15}.

In general, given an approximation $\tilde{F}^{-1}$ of F^{-1}, the absolute error on X for a given U is $|\tilde{F}^{-1}(U) - F^{-1}(U)|$. The corresponding error on U is $|F(\tilde{F}^{-1}(U)) - U|$. This second error can hardly be less than the error in the representation of U, and we are also limited by the machine precision on the representation of X. If one of these two limits is reached for each $U \in [0,1]$, for practical purposes we have an exact inversion method.

When shape parameters are involved (e.g., for the gamma and beta distributions), things are more complicated because a different approximation of F^{-1} must be constructed for each choice of shape parameters.

When we have an algorithm for computing F but not F^{-1}, and F is continuous, as a last resort we can always approximate $X = F^{-1}(U)$ by a numerical method that finds a root of the equation $U = F(X)$ for a given U. For instance, we can run the robust Brent-Dekker iterative root finding algorithm (Brent 1973, Chapter 4) until we have reached the required precision, as done by default in SSJ (L'Ecuyer 2008).

Faster inversion algorithms for fixed shape parameters can be constructed if we are ready to invest in setup time. These methods are called *automatic* when the code that approximates F^{-1} is produced automatically by a general one-time setup algorithm (Hörmann et al. 2004). The setup computes tables that contain the interpolation points and coefficients. With these tables in hand, random variate generation is very fast. For example, a general adaptive method that constructs an accurate Hermite interpolation method for F^{-1}, given a function that computes F, is developed in Hörmann and Leydold (2003). In Derflinger (2010), the authors propose an algorithm that constructs an approximation of F^{-1} to a given accuracy (specified by the user) for the case where only the density of X is available. This algorithm is an improvement over similar methods in Ahrens and Kohrt (1981). These methods assume that the distribution has bounded support, but they can be applied to most other distributions by truncating the tails far enough for the error to be negligible.

For discrete distributions, say over the values $x_1 < \cdots < x_k$, inversion finds $I = \min\{i \mid F(x_i) \geq U\}$ and return $X = x_I$. To do this, one may first tabulate the pairs $(x_i, F(x_i))$ for $i = 1,\ldots,k$, and then find I by sequential or binary search in the table (L'Ecuyer 2004). However, the fastest implementation when k is large is obtained by using an index (Chen and Asau 1974; Devroye 1986). The idea is to partition the interval $(0,1)$ into c subintervals of equal sizes, $[j/c,\ (j+1)/c)$ for $j = 0,\ldots,c-1$, and store the smallest and largest possible values of X for each subinterval, namely $L_j = F^{-1}(j/c)$ and $R_j = F^{-1}((j+1)/c)$. Once U is generated, we find the corresponding interval number $J = \lfloor cU \rfloor$, and search for I only in that interval, with linear or binary search. The fastest average time per call is usually obtained by taking a large c (so that k/c does not exceed a few units), and linear search in the subintervals (to minimize the overhead). The resulting algorithm is as fast (on average) as the alias method (Law and Kelton 2000; Walker 1974), which is often presented as the fastest algorithm but does not preserve inversion. If k is large or even infinite, for example for the Poisson distribution or the ▶binomial distribution with a large n, the pairs $(x_i, F(x_i))$ are precomputed and tabulated only in the areas where the probabilities are not too small, usually around the center of the distribution. Other values are computed dynamically only in the very rare cases where they are needed. Similar indexing techniques can also be used for piecewise-polynomial approximations of F^{-1} for continuous distributions.

Rejection Methods and Thinning

When inversion is too costly, the alternative is often a rejection method. It works as follows. Suppose we want to generate X from density f. It suffices to know f up to a multiplicative constant, i.e., to know κf, where κ might be unknown. If f is known, we take $\kappa = 1$. We pick a density

r such that $\kappa f(x) \le t(x) \stackrel{\text{def}}{=} ar(x)$ for all x for some constant a, and such that sampling random variates Y from r is easy. The function t is called a *hat function*. Integrating this inequality with respect to x on both sides, we find that $\kappa \le a$. To generate X with density f, we generate Y from the density r and $U \sim U(0,1)$ independent of Y, repeat this until $Ut(Y) \le \kappa f(Y)$, and return $X = Y$ (Devroye 1986; von Neumann 1951). The number of times we have to retry is a geometric random variable with mean $a/\kappa - 1 \ge 0$. We want a/κ to be as small as possible.

If κf is expensive to compute, computations can often be accelerated by using *squeeze functions* q_1 and q_2 that are less costly to evaluate and such that $q_1(x) \le \kappa f(x) \le q_2(x) \le t(x)$ for all x. After generating Y, we first check if $Ut(Y) \le q_1(Y)$. If so we accept Y immediately. Otherwise if $Ut(Y) \ge q_2(Y)$, we reject Y immediately. We verify the condition $Ut(Y) \le \kappa f(Y)$ explicitly only when none of the two previous inequalities is satisfied. One may also use multiple levels of embedded squeezing, with crude squeezing functions that are very quick to evaluate at the first level, then tighter but slightly more expensive ones at the second level, and so on.

In most practical situations, rejection is combined with a change of variable to transform the original density into a nicer one, for which a more efficient implementation of the rejection method can be constructed. The change of variable can be selected so that the transformed density is concave and a piecewise linear hat function is easy to construct. Typical examples of transformations can be $T(x) = \log x$ and $T(x) = -x^{-1/2}$, for instance (Devroye 1986; Hörmann et al. 2004). The rejection method also works for discrete distributions; we just replace the densities by the probability mass functions.

One special case of change of variable combined with rejection leads to the *ratio-of-uniforms* method. It is based on the observation that if X has density f over the real line, κ is a positive constant, and the pair (U, V) has the uniform distribution over the set

$$C = \left\{(u,v) \in \mathbb{R}^2 \text{ such that } 0 \le u \le \sqrt{\kappa f(v/u)}\right\},$$

then V/U has the same distribution as X (Devroye 1986; Kinderman and Monahan 1977). Thus, one can generate X by generating (U, V) uniformly over C, usually by a rejection method, and returning $X = V/U$.

A special form of rejection called *thinning* is frequently used to generate non-homogeneous ►point processes. For example, suppose we want to generate the jump times of a Poisson process (see ►Poisson Processes) whose time-varying rate is $\{\lambda(t), t \ge 0\}$, where $\lambda(t) \le \bar{\lambda}$ at all time t for some constant $\bar{\lambda}$. Then we can generate *pseudo-jumps* at constant rate $\bar{\lambda}$ by generating the times between successive jumps as i.i.d. exponentials with mean $1/\bar{\lambda}$. A pseudo-jump at time t is accepted (becomes a real jump) with probability $\lambda(t)/\bar{\lambda}$.

Multivariate Distributions

A d-dimensional *random vector* $\mathbf{X} = (X_1, \ldots, X_d)^t$ has distribution function F if $\mathbb{P}[X_1 \le x_1, \ldots, X_d \le x_d] = F(x_1, \ldots, x_d)$ for all $\mathbf{x} = (x_1, \ldots, x_d)^t \in \mathbb{R}^d$. The distribution function of a random vector does not have an inverse in general, so the inversion method does not apply directly to multivariate distributions. There are situations where one can generate X_1 directly by inversion from its marginal distribution, then generate X_2 by inversion from its marginal distribution conditional on X_1, then generate X_3 by inversion from its marginal distribution conditional on (X_1, X_2), and so on. But this is not always possible or convenient.

There are important classes of multivariate distributions for which simple and elegant methods are available. For example, suppose $\mathbf{X}$ has a *multinormal distribution* with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. When $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$ (the identity), we have a *standard multinormal* distribution. This one is easy to generate: the coordinates are independent standard normals, and they can be generated separately by inversion. For the general case, it suffices to decompose $\Sigma = \mathbf{A}\mathbf{A}^t$, generate $\mathbf{Z}$ standard multinormal, and return $\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$. The most popular way to decompose $\boldsymbol{\Sigma}$ is the Cholesky decomposition, for which $\mathbf{A}$ is lower triangular, but there are in fact many other possibilities, including for example the eigendecomposition as in ►principal component analysis. The choice of decomposition can have a large impact on the variance reduction in the context of randomized quasi-Monte Carlo integration, by concentrating much of the variance on just a few underlying uniform random numbers (Glasserman 2004; L'Ecuyer 2009).

Multivariate normals are useful for various purposes. For example, to generate a random point on a sphere in d dimensions centered at zero, generate a standard multinormal vector $\mathbf{Z}$, then normalize its length to the desired radius. This is equivalent to generating a *random direction*. A more general class of multivariate distributions named *radially symmetric* are defined by putting $\mathbf{X} = R\mathbf{Z}/\|\mathbf{Z}\|$ where $\mathbf{Z}/\|\mathbf{Z}\|$ is a random direction and R has an arbitrary distribution over $(0, \infty)$. For example, if R has the Student distribution, then $\mathbf{X}$ is multivariate Student. A further generalization yields the *elliptic multivariate* random variable: $\mathbf{X} = \boldsymbol{\mu} + R\mathbf{A}\mathbf{Z}/\|\mathbf{Z}\|$ where $\mathbf{Z}$ is a multinormal in k dimensions and $\mathbf{A}$ is a $d \times k$ matrix. It is easy to generate $\mathbf{X}$ if we know how to generate R.

A very rich class of multivariate distributions are defined via *copula methods* (Hörmann and Derflinger 2002; Nelsen 1999). Start with an arbitrary d-dimensional cdf G with continuous marginals G_j, generate $\mathbf{Y} = (Y_1, \ldots, Y_d)^t$ from G, and let $\mathbf{U} = (U_1, \ldots, U_d) = (G_1(Y_1), \ldots, G_d(Y_d))^t$. At this point, the U_j have the uniform distribution over $(0,1)$, but they are not independent in general. The cdf of $\mathbf{U}$ is the *copula associated with* G and it specifies the dependence structure of the vector $\mathbf{U}$. In fact, any cdf over $(0,1)^d$ with uniform marginals can act as a copula. To generate a vector $\mathbf{X} = (X_1, \ldots, X_d)^t$ with arbitrary marginal cdf's F_j and a dependence structure specified by this copula, just put $X_j = F_j^{-1}(U_j)$ for each j. A popular choice for G is the multinormal cdf with standard normal marginals; then $\mathbf{Y}$ and $\mathbf{U}$ are easy to generate, and one can select the correlation matrix of $\mathbf{Y}$ to approximate a target correlation (or rank correlation) matrix for $\mathbf{X}$. This is known as the NORTA (normal to anything) method. It can usually match the correlations pretty well. But to approximate the whole dependence structure in general, a much richer variety of ▶copulas is required (Hörmann and Derflinger 2002; Nelsen 1999).

The rejection method extends rather straightforwardly to the multivariate case. For a known target d-dimensional density f, pick a d-dimensional density r such that $f(\mathbf{x}) \le ar(\mathbf{x})$ for all $\mathbf{x}$ and some constant a, and such that sampling random vectors $\mathbf{Y}$ from r is easy. To generate $\mathbf{X}$ with density f, generate $\mathbf{Y}$ from r and $U \sim U(0,1)$ independent of $\mathbf{Y}$, until $Uar(\mathbf{Y}) \le f(\mathbf{Y})$, and return $\mathbf{X} = \mathbf{Y}$.

Stochastic Processes

Various types of ▶stochastic processes can be simulated in a way that becomes obvious from their definition. The *Lévy processes* form an important class; they are continuous-time stochastic processes $\{Y(t), t \ge 0\}$ with $Y(0) = 0$ and whose increments over disjoint time intervals are independent, and for which the increment over a time interval of length t has a distribution that depends only on t (the mean and standard deviation must be proportional to t) (Asmussen and Glynn 2007; Bertoin 1996). Special instances include the (univariate or multivariate) *Brownian motion* (*see* ▶Brownian Motion and Diffusions), the stationary *Poisson process*, the *gamma process*, and the *inverse Gaussian process*, for example, whose increments have the (multi)normal, Poisson, gamma, and inverse Gaussian distributions (*see* ▶Inverse Gaussian Distribution), respectively. A natural way to generate a Lévy process observed at times $0 = t_0 < t_1 < \cdots < t_c$ is to generate the independent increments $Y(t_j) - Y(t_{j-1})$ successively, for $j = 1, \ldots, c$. This is the *random walk* method.

For certain ▶Lévy processes (such as those mentioned above), for any $t_1 < s < t_2$, it is also easy to generate $Y(s)$ from its distribution *conditional* on $\{Y(t_1) = y_1, Y(t_2) = y_2\}$ for arbitrary y_1, y_2. Then the trajectory can be generated via the following *Lévy bridge sampling* strategy, where we assume for simplicity that c is a power of 2. We start by generating $Y(t_c)$ from the distribution of the increment over $[0, t_c]$, then we generate $Y(t_{c/2})$ from its distribution conditional on $(Y(t_0), Y(t_c))$, then we apply the same technique recursively to generate $Y(t_{c/4})$ conditional on $(Y(t_0), Y(t_{c/2}))$, $Y(t_{3c/4})$ conditional on $(Y(t_{c/2}), Y(t_c))$, $Y(t_{c/8})$ conditional on $(Y(t_0), Y(t_{c/4}))$, and so on. This method is convenient if one wishes to later refine the approximation of a trajectory. It is also effective for reducing the effective dimension in the context of quasi-Monte Carlo methods (L'Ecuyer 2009).

For the Poisson process, one usually wishes to have the individual jump times, and not only the numbers of jumps in predetermined time intervals. For a stationary Poisson process, the times between successive jumps are easy to generate, because they are independent exponential random variables. For a non-stationary Poisson process, one way is to apply a nonlinear time transformation to turn it into a standard stationary Poisson process of rate 1, generate the jumps times of the standard process, and apply the reverse time transformation to recover the jump times of the target non-stationary Poisson process. This idea applies to other continuous-time stochastic processes as well, as we now explain.

Given a process $X = \{X(t), t \ge 0\}$, and another process $T = \{T(t), t \ge 0\}$ with nondecreasing trajectories, called a *subordinator*, we can define a new process $Y = \{Y(t) \stackrel{\text{def}}{=} X(T(t)), t \ge 0\}$, which is the process X to which we have applied a random time change. This can be applied to any process X with index $t \in \mathbb{R}$. If both X and T are Lévy processes, then so is Y. If X is a stationary Poisson process with rate 1 and we want a nonstationary Poisson process Y with rate function $\{\lambda(t), t \ge 0\}$, then we must take $T(t) = \Lambda(t) \stackrel{\text{def}}{=} \int_0^t \lambda(s)\mathrm{d}s$. To simulate Y, we generate the jump times $Z_1 < Z_2 < \cdots$ of the stationary process X by generating $Z_j - Z_{j-1}$ as independent exponentials with mean 1, and define the jump times of Y as $T_j = \Lambda^{-1}(Z_j)$ for $j \ge 1$.

If X is a one-dimensional Brownian motion, the random time change is equivalent to replacing the constant volatility parameter σ of the Brownian motion by a stochastic (time-varying) volatility process $\{\sigma(t), t \ge 0\}$. Two notable examples of Lévy processes that can act as subordinators are the gamma process and the inverse Gaussian process, respectively. Their use as subordinators

for the Brownian motion yields the variance gamma and the normal inverse Gaussian processes. Brownian motions with such a random time change do provide a much better fit to various types of financial data (such as the log prices of stocks and commodities, etc.) than standard Brownian motions, and they recently became popular for this reason.

Acknowledgement

This work has been supported by the Natural Sciences and Engineering Research Council of Canada Grant No. ODGP0110050 and a Canada Research Chair to the author.

About the Author

For biography *see* the entry ►Uniform Random Number Generators.

Cross References

►Brownian Motion and Diffusions
►Computational Statistics
►Copulas
►Inverse Gaussian Distribution
►Lévy Processes
►Markov Chain Monte Carlo
►Monte Carlo Methods in Statistics
►Multivariate Statistical Distributions
►Point Processes
►Poisson Processes
►Stochastic Processes
►Uniform Random Number Generators

References and Further Reading

Ahrens JH, Kohrt KD (1981) Computer methods for efficient sampling from largely arbitrary statistical distributions. Computing 26:19–31

Asmussen S, Glynn PW (2007) Stochastic simulation. Springer, New York

Bertoin J (1996) Lévy Processes. Cambridge University Press, Cambridge

Blair JM, Edwards CA, Johnson JH (1976) Rational Chebyshev approximations for the inverse of the error function. Math Comput 30:827–830

Brent RP (1973) Algorithms for minimization without derivatives. Prentice-Hall, Englewood Cliffs

Chen HC, Asau Y (1974) On generating random variates from an empirical distribution. AIEE Trans 6:163–166

Derflinger G, Hörmann W, Leydold J (2010) Random variate generation by numerical inversion when only the density is known. ACM Trans Model Comput Simul 20(4)

Devroye L (1986) Non-Uniform random variate generation. Springer, New York

Devroye L (2006) Nonuniform random variate generation. In: Henderson SG, Nelson BL(eds) Simulation, handbooks in operations research and management science, Chap. 4. Elsevier, Amsterdam, pp 83–121

Glasserman P (2004) Monte Carlo methods in financial engineering. Springer, New York

Hörmann W, Derflinger G (2002) Fast generation of order statistics. ACM Trans Model Comput Simul 12(2):83–93

Hörmann W, Leydold J (2003) Continuous random variate generation by fast numerical inversion. ACM Trans Model Comput Simul 13(4):347–362

Hörmann W, Leydold J, Derflinger G (2004) Automatic nonuniform random variate generation. Springer, Berlin

Kinderman AJ, Monahan JF (1977) Computer generation of random variables using the ratio of uniform deviates. ACM Trans Math Softw 3:257–260

L'Ecuyer P (2004) Random number generation. In: Gentle JE, Haerdle W, Mori Y (eds) Handbook of computational statistics, Chap. II.2, Springer, Berlin, pp 35–70

L'Ecuyer P (2008) SSJ: a java library for stochastic simulation. Software user's guide. http://www.iro.umontreal.ca/~lecuyer

L'Ecuyer P (2009) Quasi-Monte Carlo methods with applications inhh finance. Finance Stochastics 13(3):307–349

Law AM, Kelton WD (2000) Simulation modeling and analysis, 3rd edn. McGraw-Hill, New York

Nelsen RB (1999) An introduction to Copulas. Lecture Notes in Statistics, vol 139. Springer, New York

von Neumann J (1951) Various techniques used in connection with random digits. In: Householder As et al (ed) The Monte Carlo method, vol 12. National Bureau of Standards Applied Mathematics Series, Washington, pp 36–38

Walker AJ (1974) New fast method for generating discrete random numbers with arbitrary frequency distributions. Electron Lett 10:127–128

N

Normal Distribution, Univariate

Pushpa Narayan Rathie
Professor
University of Brasilia, Brasilia, Brazil

Brief Historical Background

The normal distribution is used extensively in probability theory, statistics, and the natural and social sciences. It is also called the Gaussian distribution because Carl Friedrich Gauss (1809) used it to analyze astronomical data. The normal distribution was (a) first introduced by Abraham de Moivre (1733), as an approximation to a ►binomial distribution (b) used by Laplace (1774), as an approximation to hypergeometric distribution to analyze errors of experiments and (c) employed, in the past, by Legendre, Peirce, Galton, Lexis, Quetelet, etc. The normal distribution can be used as an approximation to other distributions because the standardized sum of a large number of independent and identically distributed random variables is approximately normally distributed. Thus, the normal distribution can be used when a large number

of non-normal distribution correspond more closely to observed values. The 1816 work of Gauss (Gauss 1816) is the earliest result of this kind when he derived the normal distribution as the sum of a large number of independent astronomical data errors. The central limit theorem (see ►Central Limit Theorems) for independent and identically distributed random variables was established by Lyapunov (1900). This result was extended to non identically distributed random variables by Lindeberg (1922) and for non-independent random variables later on (see Loève 1963; Gnedenko and Kolmogorov 1954).

Important Properties and Results

The normal distribution has a continuous probability density function which is bell-shaped and has its maximum at the mean (Figs. 1 (Normal densities) and 2 (Cumulative distribution functions)).

The expression for density function along with various basic important results are listed below:

1. Notation: $X \sim N(\mu, \sigma^2)$
2. Probability density function:

$$f(x) = ((2\pi)^{1/2}\sigma)^{-1} \exp[-(x-\mu)^2/(2\sigma^2)], -\infty < x, \mu < \infty, \sigma > 0$$

3. Mean (location parameter), Median, and Mode: $E(X) = \mu$ = median = mode.
4. Variance (shape parameter): $V(X) = \sigma^2$
5. Coefficient of skewness: $\tau_3 = 0$
6. Coefficient of kurtosis: $\tau_4 = 3$
7. Standard normal variate: $Y = (X-\mu)/\sigma \sim N(0,1)$
8. Standard normal density function: $\phi(y) = (2\pi)^{-1/2} \exp(-y^2/2), -\infty < y < \infty$
9. Folded normal density function: $g(y) = 2((2\pi)^{1/2}\sigma)^{-1} \exp[-y^2/(2\sigma^2)], y > 0, \sigma > 0$
10. Cumulative distribution function corresponding to (8):

$$\varphi(x) = \left[1 + erf\left(x/\sqrt{2}\right)\right]/2,$$

$$\text{where } erf(x) = (2/\pi^{1/2}) \int_0^x e^{-t^2} dt = (2/\pi^{1/2}) \sum_{n=0}^{\infty} [(-1)^n x^{2n+1}/\{n!(2n+1)\}].$$

11. ►Moment generating function:

$$M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$$

12. Characteristic function:

$$\phi_X(t) = \exp(i\mu t - \sigma^2 t^2/2), i = (-1)^{1/2}$$

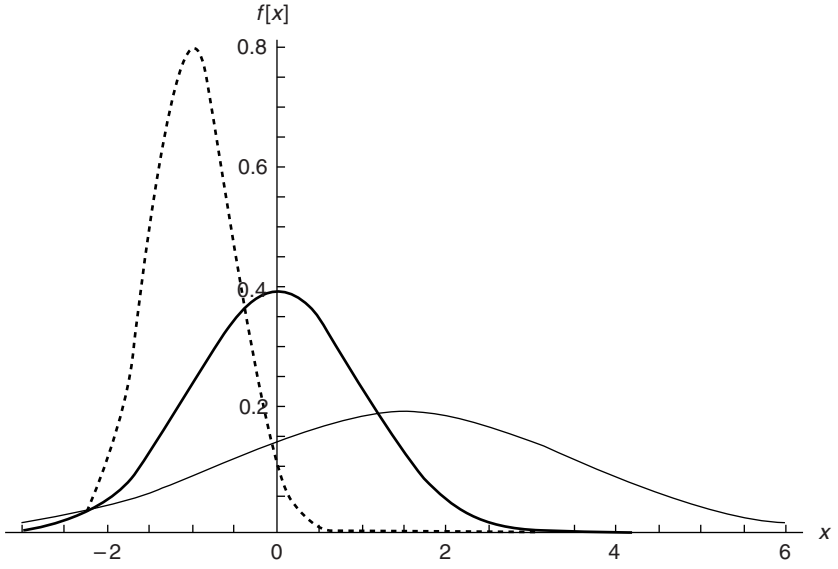
13. nth central moment:

$$\mu_n = E(X-\mu)^n = \begin{cases} 0, & n\text{-odd} \\ n!\sigma^n/[2^{n/2}(n/2)!], & n\text{-even} \end{cases}$$

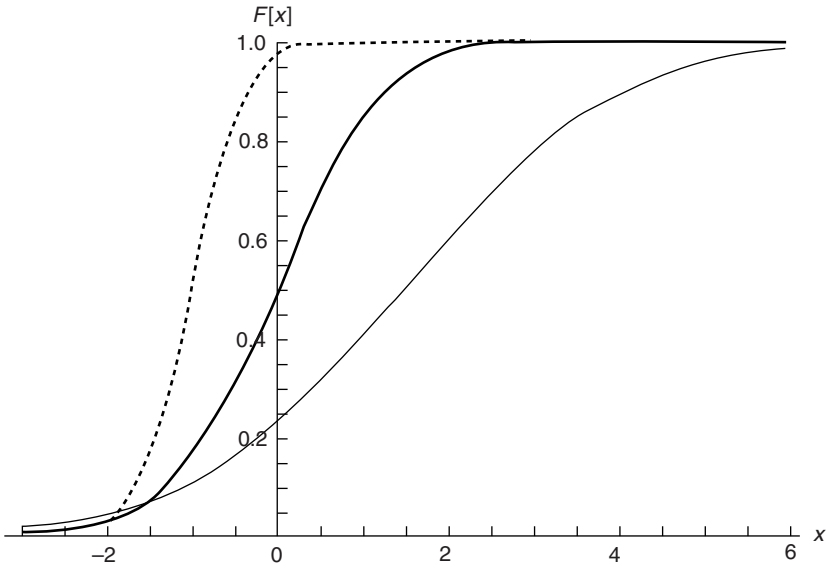
14. L-moments: $\lambda_1 = \mu, \lambda_2 = \sigma/\pi^{1/2}, \lambda_3 = 0$
15. Shannon entropy: $H(X) = (1/2)\ln(2\pi e\sigma^2)$
16. Rényi entropy of order α:

$$H_\alpha(X) = (1/2)\ln(2\pi\sigma^2) - (\ln\alpha)/[2(1-\alpha)]$$

17. Song measure: $V(\ln f(X)) = -2 \lim_{\alpha \to 1} \frac{dH_\alpha(X)}{d\alpha} = 1/2$



Normal Distribution, Univariate. Fig. 1 $N(-1,1/4)$: . . . ; $N(0,1)$: _____; $N(1.5,4)$: _____



Normal Distribution, Univariate. Fig. 2 $N(-1,1/4):\ldots;N(0,1):$ _____; $N(1.5,4):$ _____

18. Fisher information: (a) of the distribution of X is $I(X)=\sigma^{-2}$, (b) with respect to μ is $I(\mu)=\sigma^{-2}$ and (c) with respect to σ^2 is $I(\sigma^2)=(2\sigma^4)^{-1}$

Some properties of the normal distribution are given below:

19. If $X\sim N(\mu,\sigma^2)$, then (a) $aX+b\sim N(a\mu+b,a^2\sigma^2)$, where a and b are real numbers and (b) $Y=e^X\sim \log normal(\mu,\sigma^2)$.
20. If $X_1,\ldots,X_n$ with $X_i\sim N\left(\mu_i,\sigma_i^2\right),i=1,\ldots,n$ are independent, then $\sum_{i=1}^{n}a_iX_i\sim N\left(\sum_{i=1}^{n}a_i\mu_i,\sum_{i=1}^{n}a_i^2\sigma_i^2\right)$ where $a_1,\ldots,a_n$ are real numbers.
21. If $X_1,\ldots,X_n$, $X_i\sim N(0,1),i=1,\ldots,n$ are independent, then $\sum_{i=1}^{n}X_i^2\sim\chi_n^2$.
22. For a sample of size n from $N(\mu,\sigma^2)$, the sample mean $\overline{X}=\sum_{i=1}^{n}X_i/n$ and sample variance $S^2=\sum_{i=1}^{n}(X_i-\overline{X})^2/(n-1)$ are independent.
23. If $X_i\sim N(0,1),i=1,2$, then $Y=X_1/X_2$ has the Cauchy distribution with density function $g(y)=[\pi(1+y^2)]^{-1},y\in\Re$.
24. If $X\sim N(0,\sigma^2)$, then $|X|$ is folded normal.
25. Central Limit Theorem (Lyapunov): If $X_1,\ldots,X_n$ are independent and identically distributed random variables with finite mean and variance, then $[\overline{X}-E(\overline{X})]/[V(\overline{X})^{1/2}]\sim N(0,1)$, as $n\to\infty$.

Importance of Normal Distribution in Statistics

26. The sample mean $\overline{X}$: (a) is the maximum likelihood estimator for the population mean μ and (b) has the distribution $N(\mu,\sigma^2/n)$.
27. The sample variance S^2 is the unbiased estimator for the population variance σ^2.
28. $(n-1)S^2/\sigma^2\sim\chi_{n-1}^2$.
29. $S^2\sim\Gamma((n-1)/2,2\sigma^2/(n-1))$.
30. Moment Estimators: The method of moments leads to estimating μ by $\overline{X}$ and σ^2 by $\sum_{i=1}^{n}(X_i-\overline{X})^2/n$.
31. The sample mean of i.i.d random variables $X_i\sim N(\mu,\sigma^2),i=1,\ldots,n$ is consistent for μ and the sample variance for σ^2.
32. The statistic $(\overline{X},S^2)$ is a sufficient statistic for (μ,σ^2).
33. ▶Cramér-Rao inequality implies that $\overline{X}$ is the minimum variance unbiased estimator of μ. In other words, $\overline{X}$ is efficient.
34. Confidence Intervals: Let $X_1,\ldots,X_n$ be i.i.d random variables with $X_i\sim N(\mu,\sigma^2),i=1,\ldots,n$. Then the symmetric $(1-\alpha)$-level confidence interval:
    (a) for μ, given σ^2, is $(\overline{X}-z_{\alpha/2}\sigma/n^{1/2},\overline{X}+z_{\alpha/2}\sigma/n^{1/2})$, where $z_{\alpha/2}$ is given by $P(Y\geq z_{\alpha/2})=\alpha/2$, for $Y\sim N(0,1)$;
    (b) for σ^2, with μ unknown, is $((n-1)S^2/\chi_{n-1,\alpha/2}^2,(n-1)S^2/\chi_{n-1,1-\alpha/2}^2)$, and
    (c) for σ^2, with μ known, is

$$\left(\sum_{i=1}^{n}(X_i-\mu)^2/\chi_{n,\alpha/2}^2,\sum_{i=1}^{n}(X_i-\mu)^2/\chi_{n,1-\alpha/2}^2\right).$$

The confidence intervals are used in hypotheses testing problems involving μ and σ^2.

An Approximation to the Normal Distribution

Several approximations to the normal distribution are available in the literature. In 2006, Rathie and Swamee (see Rathie et al. 2008) defined a family of invertible distributions by taking the generalized logistic distribution function as

35. $F(x) = \left[1 + \exp\{-x(a + b|x|^p)\}\right]^{-1}, x \in \Re, a, b, p > 0.$

This distribution is a very good approximation for $a = 1.59413, b = 0.07443$, and $p = 1.939$ with a maximum error of (a) 4×10^{-4} at $x = 0$, to the standard normal density function, and (b) 7.757×10^{-5} at $x = \pm 2.81$ to the cumulative standard normal distribution function. It may be pointed out that for certain sets of values of the parameters a, b, and p, this distribution approximates very well the Student t-distribution. The density function corresponding to (35) is given by

36. $f(x) = [a + b(1+p)|x|^p] \exp[-x(a + b|x|^p)] / [1 + \exp\{-x(a + b|x|^p)\}]^2,$
$x \in \Re, a, b, p > 0.$

Some Applications and Computational Aspects

The normal distribution is used in various practical applications occurring in many areas such as Economics (used earlier to analyze exchange rates, stock markets, etc.; nowadays heavy tailed distributions, such as Lévy distribution, are used), Medical Sciences (blood pressure of adults (males or females)), Physics (measurement errors, heat equation), Election predictions, etc.

Most statistical and mathematical packages may be used for numerical and symbolical calculations of several results concerning the normal distribution as well as for verifying if a given data set is approximately normally distributed. There are several methods to generate values that have normal distribution. Box and Muller (see Rubinstein and Kroese 2008) gave an easy method of generating variables from $N(0,1)$ based on the following result. Let U_1 and U_2 be independent random variables from uniform distribution $U(0,1)$. Then the random variables

37. $X = (-2\ln U_1)^{1/2} \cos(2\pi U_2)$

and

38. $Y = (-2\ln U_1)^{1/2} \sin(2\pi U_2)$

are independent $N(0,1)$.

Alternatively, variables from approximate $N(0,1)$ given in Eq. (35) can be generated from the following expressions derived by Rathie and Swamee (2006) (see Rathie et al. 2008):

39. $$x = \begin{cases} -\sum_{k=0}^{\infty} [(-b/a)^k \Gamma(kp+k+1)/\{k\Gamma(kp+2)\}] \\ \quad \times [a^{-1} \ln\{(1-F)/F\}]^{kp+1}, \quad F \le 0.5 \\ \sum_{k=0}^{\infty} [(-b/a)^k \Gamma(kp+k+1)/\{k!\Gamma(kp+2)\}] \\ \quad \times [a^{-1} \ln\{F/(1-F)\}]^{kp+1}, \quad F \ge 0.5 \end{cases}$$

About the Author

Dr. Rathie is a Professor, Department of Statistics, University of Brasilia, Brasilia, Brazil. He was awarded merit scholarship of the Govt. of India and two gold and two silver medals for academic distinction from the University of Rajasthan and the University of Jodhpur. He has authored or co-authored more than 125 research papers and 2 books: *Some Basic Concepts in Information Theory and Statistics* (Wiley, 1975, co-author: A.M. Mathai) and *Probability and Statistics* (Macmillan, 1977, co-author: A.M. Mathai). He was Managing Editor of *International Journal of Mathematical and Statistical Sciences* (1992–2000).

Cross References

- ▶Approximations to Distributions
- ▶Asymptotic Normality
- ▶Bivariate Distributions
- ▶Central Limit Theorems
- ▶Logistic Normal Distribution
- ▶Multivariate Normal Distributions
- ▶Multivariate Statistical Distributions
- ▶Normality Tests
- ▶Normality Tests: Power Comparison
- ▶Skew-Normal Distribution
- ▶Statistical Distributions: An Overview
- ▶Statistics, History of

References and Further Reading

De Moivre A (1733) Approximatio ad Summam Ferminorum Binomii $(a + b)^n$ in Seriem expansi. Supplementum II to Miscellanae Analytica, pp 1–7

Gauss CF (1809) Theoria Motus Corporum Coelestum. Perthes and Besser, Hamburg

Gauss CF (1816) Bestimmung der Genauigkeit der Beobachtungen. Z Astronom 1:185–197

Gnedenko BV, Kolmogorov AN (1954) Limit distributions for sums of independent random variables. Addison-Wesley, Reading
Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol 1. Wiley, New York
Laplace PS (1774) Determiner le milieu que l'on doit prendre entre trois obsevations données d'un même phénomené. In: Mémoires de Mathématique et Physique presentées à l'Académie Royale dês Sciences par divers Savans, vol 6, pp 621–625
Lindeberg JW (1922) Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. Math Z 15:211–225
Loève M (1963) Probability theory 3rd edn. D. Van Nostrand, New York
Luke YL (1969) The special functions and their approximations, vol 1. Academic, New York
Lyapunov A (1900) Sur une proposition de la théorie des probabilités. Izv Akad Nauk SSSR Ser V 13:359–386
Rathie PN, Swamee PK, Matos GG, Coutinho M, Carrijo TB (2008) H-functions and statistical distributions. Ganita 59(2): 23–37
Rubinstein RY, Kroese DP (2008) Simulation and Monte Carlo method. Wiley, New York

Normal Scores

Lelys Bravo de Guenni
Professor
Universidad Simón Bolívar, Caracas, Venezuela

Normal scores use the ranks of a data set to calculate standard normal quantiles of the same size than the original data set. The aim of this calculation is mainly to compare each sample value with the expected value of the order statistic of the same rank from a standard normal distribution sample of the same sample size than the original data set. The expected values of the ▶order statistics of a sample from a standard normal distribution are the sorted values in increasing order. Plots of the original data or sample quantiles versus the quantiles or scores derived from the standard normal distribution are best known as the *Normal Quantile–Quantile plots* or *Normal Q–Q plots* where Q stands for quantile. These plots are commonly used as a simple test for normality in a graphical way. If the data set is a sample from a normal probability distribution, the Q–Q plot should show a linear relationship (Barnnet 1975).

The best way to explain how to calculate the normal scores is through an example. Suppose we have the sample $X = (x_1, \ldots, x_n)$, which arises from a certain distribution function F with location and scale parameters μ and σ. The ordered sample values are denoted as $x_{(1)}, \ldots, x_{(n)}$. As a specific case assume $X = (2.3, 4.2, 0.1, 3.3, 2.1, 0.3, 2.7)$ with sample size $n = 7$. Associated to this sample we can calculate the standard normal quantile $z_{k/(n+1)}$ for each x_k where k is the rank of the observation; $k/(n+1)$ is the plotting position and Φ is the standard normal probability distribution function. In other words, $z_{k/(n+1)} = \Phi^{-1}(k/(n+1))$. In our example the rank of the observation 2.3 is $i = 4$. The quantile $z_{4/(7+1)} = z_{0.5}$ is 0. Detailed calculations for the sample data set are presented in Table 1.

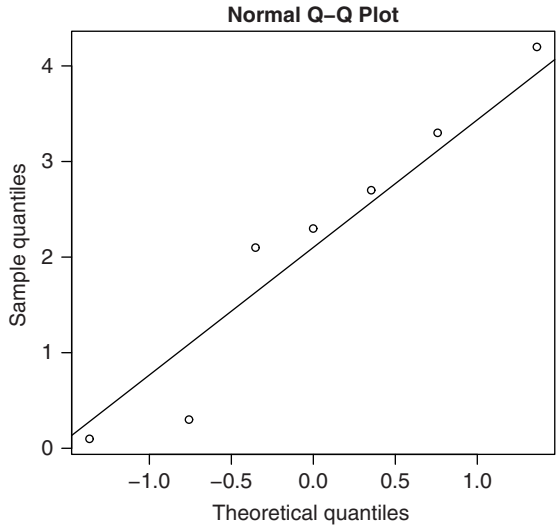
Plotting positions are plausible empirical estimates of $F[(x_{(k)} - \mu)/\sigma]$. To calculate the plotting position there are several options as summarized by Hyndman and Fan (1996). If $p(k)$ is the plotting position of rank k the following expressions for $p(k)$ may be considered:

1. $p(k) = k/n$
2. $p(k) = (k - 0.5)/n$
3. $p(k) = k/(n+1)$
4. $p(k) = (k-1)/(n-1)$
5. $p(k) = (k - 1/3)/(n + 1/3)$
6. $p(k) = (k - 3/8)/(n + 1/4)$

The differences of using different plotting position formulas in the resulting Q–Q plots will be very small. For small sample size ($n < 10$) R Development Core Team (2007) uses option 6 and has implemented the function *qqnorm* to produce Normal Q–Q plots. For our sample data set *qqnorm* produces Fig. 1. Other authors as Hyndman and Fan (1996) recommend the use of option 5.

Normal Scores. Table 1 Normal scores for the sample data *X*

Ordered sample data	Plotting position	Normal scores
0.1	1/8	−1.150
0.3	2/8	−0.674
2.1	3/8	−0.319
2.3	4/8	0
2.7	5/8	0.319
3.3	6/8	0.674
4.2	7/8	1.150



Normal Scores. Fig. 1 Quantile–Quantile plot

About the Author

Dr. Lelys Guenni is a Professor at the Department of Scientific Computing and Statistics, Universidad Simón Bol'ıvar, Caracas, Venezuela in the Graduate Programs of Statistics. She was director of the Statistical Center and Mathematical Software at Simón Bol'ıvar University for the period 2000–2004. She was a member of the Science Steering Committee of the project Biospherical Aspects of the Hydrological Cycle associated to the International Geosphere–Biosphere Programme, during the period 1996–2002. She has coauthored more than 25 papers with statistical applications to climate, hydrology, and environmental problems and was a lead author of the Ecosystem Millenium Assessment. Professor Guenni is a Board Member of the International Environmetrics Society and Associated Editor of the *Environmetrics Journal*.

Cross References

►Normal Distribution, Univariate
►Order Statistics
►Ranks
►Statistical Fallacies: Misconceptions, and Myths

References and Further Reading

Barnnet V (1975) Probability plotting methods and order statistics. Appl Stat 24(1):95–108
Hyndman RJ, Fan Y (1996) Sample quantiles in statistical packages. Am Stat 50:361–365
R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0

Normality Tests

Henry C. Thode
Associate Professor
Stony Brook University, Stony Brook, NY, USA

The Importance of Testing for Normality

Many statistical procedures such as estimation and hypothesis testing have the underlying assumption that the sampled data come from a normal distribution. This requires either an effective test of whether the assumption of normality holds or a valid argument showing that non-normality does not invalidate the procedure. Tests of normality are used to formally assess the assumption of the underlying distribution.

Much statistical research has been concerned with evaluating the magnitude of the effect of violations of the normality assumption on the true significance level of a test or the efficiency of a parameter estimate. Geary (1947) showed that for comparing two variances, having a symmetric non-normal underlying distribution can seriously affect the true significance level of the test. For a value of 1.5 for the kurtosis of the alternative distribution, the actual significance level of the test is 0.000089, as compared to the nominal level of 0.05 if the distribution sampled were normal. For a distribution with a kurtosis value of 6, the probability of rejection was 0.215. On the other hand, for the *t*-test the distortion of a Type I error is small if the underlying distribution is symmetric; however, if the underlying distribution is asymmetric or skewed, marked changes to the probability of rejection can occur. For the two sample *t*-test Geary concluded that if the underlying distribution is the same for both populations, regardless of the type of non-normality, the changes in the probability that the null hypothesis is rejected are small. Large differences can occur if the distributions are different. Others have corroborated these findings (e.g., Box (1953), Pearson and Please (1975), Subrahmaniam et al. (1975)): that comparative tests on means are not very sensitive when the departure from normality is the same in the different groups. Type I error in tests of variance and single sample

t-tests can be greatly affected depending on the type and degree of non-normality.

Tukey (1960) addressed the problem of robustness in estimation against slight departures from normality. He showed the effects of non-normality on estimates of location and scale parameters of a distribution which was an unbalanced mixture of two normal distributions with common mean and different variances. Under such circumstances he stated that "... neither mean nor variance is likely to be a wisely chosen basis for making estimates from a large sample." D'Agostino and Lee (1977) compared the efficiency of several estimates of location, including the sample mean, when the underlying distribution was either a Student's t or exponential power distribution, both of which are symmetric families of distributions. The efficiencies of the estimates were compared based on the kurtosis value of the underlying distributions. For the t distribution the relative efficiency of the sample mean only decreases to about 90% for a kurtosis value of 6 (corresponding to a t distribution with 6 degrees of freedom) compared to when the sample is from a normal distribution. For the exponential power distribution family, however, the relative efficiency of the sample mean drops quickly and decreases to about 50% when the kurtosis value is 6 (the Laplace distribution).

Tests for normality are useful for applications other than checking assumptions for estimates and tests. For example, some tests for normality have been found to be effective at detecting ►outliers in a sample. They also have promise as a tool in cluster analysis (see ►Cluster Analysis: An Introduction) where the alternative is a normal mixture model, with applications to genetics as an example.

Testing Distributional Hypotheses

There are more tests designed to assess normality than for any other specific distribution. Many normality tests take advantage of special properties of the normal distribution, for example, general absolute moment tests are based on unique relations among the moments of the normal distribution. The usual measure of the worth of a test for normality is its power, i.e., the probability of detecting a non-normal distribution.

Suppose you have a random sample of n independent and identically distributed observations of a random variable X, labeled $x_1, x_2, \ldots, x_n$, from an unspecified probability density $f(x)$. The general goodness of fit problem consists of testing the null hypothesis

$$H_0 : f(x) = f_0(x)$$

against an alternative hypothesis. The probability density in the null hypothesis $f_0(x)$ has a specified distributional form. When the parameters are completely specified, the null hypothesis is called a simple hypothesis. If one or more of the parameters in H_0 are not specified, H_0 is called a composite hypothesis. We will only consider tests of the composite hypothesis of normality

$$H_0 : f(x) = N(\mu, \sigma^2)$$

where both the mean μ and standard deviation σ are unknown. This is more commonly the case of interest in practice. The tests which we will describe are also location and scale invariant, i.e., they have the property that a change in the location or scale of the observations do not affect the test statistic T, i.e.,

$$T(x_1, x_2, \ldots, x_n) = T(kx_1 - u, kx_2 - u, \ldots, kx_n - u)$$

for constants k and u. This is a desirable property of a test since the parameters do not affect the shape of the normal distribution. For most statistical procedures, distribution assumptions which are made usually only concern shape.

Types of Normality Tests

Which of the many tests for normality is the best to use in a specific situation depends on how much is known or assumed about the alternative hypothesis. Normality tests have been derived based on alternative hypotheses ranging from specific distributions to non-normality of a completely unspecified nature.

Likelihood ratio tests and most powerful location and scale invariant (MPLSI) tests were derived for detecting specific alternative distributions to normality. These tests are based on the joint probabilities of the null and alternative distributions, conditional on the values of the observations. The likelihood ratio test (LRT) for the uniform (David et al. 1954) and for the double exponential alternatives (Uthoff 1973) are among the most powerful for these specific alternatives and distributions that are similar in shape; MPLSI tests show good power for specific alternatives of uniform, exponential and double exponential alternatives (Uthoff 1970, 1973) and Cauchy alternatives (Franck 1981). These would be the tests of choice when the alternative hypothesis is specified as one of these distributions.

More often the specific alternative is not known and a more general alternative hypothesis must be considered. The general shape of the alternative may be known, or only a certain type of departure from normality may be of concern. Conventionally, non-normal alternatives have been divided into three shape classes based on the comparison of their third and fourth standardized moments (denoted $\sqrt{\beta_1}$ and β_2, respectively) to those of the normal distribution. A distribution whose value of $\sqrt{\beta_1}$ (skewness) is

different from 0 has a skewed shape. The value of β_2 (kurtosis) for a normal distribution is 3, although a value of 3 does not necessarily indicate a normal distribution (e.g., Johnson et al. 1980). Symmetric alternatives are often separated into those shapes with β_2 less than 3 (sometimes called light-tailed distributions) and those shapes with β_2 greater than 3 (heavy-tailed distributions).

Among the best tests for detecting skewed or symmetric non-normality are the moment tests $\sqrt{b_1}$ (sample skewness) and b_2 (sample kurtosis), respectively. Defining the kth sample moment as

$$m^k = \sum (x_i - \bar{x})^k / n$$

then $\sqrt{b_1} = m_3/(m_2)^{3/2}$ and $b_2 = m_4/(m_2)^2$. These tests perform even better as directional or single tail tests, i.e., when the type of skewness or kurtosis is known (skewed left vs. skewed right; light- vs. heavy-tailed). For simplicity of calculation, the uniform likelihood ratio test u (range/standard deviation) and an asymptotic equivalent to the double exponential MPLSI test, Geary's (1936) absolute moment test a, are useful for detecting light- and heavy-tailed alternatives, respectively.

Omnibus tests are designed to cover all possible alternatives. They are not usually as powerful as specific or shape tests when the characteristics of the true alternative can be correctly identified. One of the better tests for detecting unspecified departures from normality is the Shapiro-Wilk W (Shapiro and Wilk 1965), which is the ratio of an estimate of σ^2 obtained from the regression of the sample order statistics on their expected values to s^2. The performance of $\sqrt{b_1}$ and b_2 at detecting shape alternatives suggest that combinations of moment tests would have acceptable power as omnibus tests; the Jarque-Bera test (1980, 1987) is computationally the simplest of these tests, calculated as $JB = n(\sqrt{b_1}/6 + b_2/24)$.

Empirical distribution function (EDF) goodness of fit tests are less powerful than tests for normality, with the most notable exception being the Anderson-Darling A^2 test (Anderson and Darling 1954) which performs very well from a power standpoint. Most of these procedures operate by using the cumulative distribution function to reduce the general problem to the specific one of testing the hypothesis of uniformity. The ►Kolmogorov–Smirnov test in particular has poor power compared to most normality tests in most situations. Similarly, the very common χ^2 and goodness of fit tests based on spacing should be avoided as tests for normality.

The following table shows some of the more powerful tests under a variety of alternatives

Test (symbol)	Source	Type of test	Good power at detecting this type of alternative
Anderson–Darling (A^2)	Anderson and Darling 1954	EDF	Any
Shapiro–Wilk (W)	Shapiro and Wilk 1965	Regression	Any
Jarque–Bera (JB)	Jarque and Bera 1980, 1987	Joint moment	Any
Skewness ($\sqrt{b_1}$)	Thode 2002	Moment	Skewed
Kurtosis (b_2)	Thode 2002	Moment	Symmetric
Geary (a)	Geary 1936	Absolute moment	Double exponential, symmetric $\beta_2 > 3$
Range (u)	David et al. 1954	LRT	Uniform, symmetric $\beta_2 < 3$

Test Recommendations

As indicated previously, the number of normality tests is large, too large for even the majority of them to be mentioned here. Overall the best tests appear to be the moment tests, Shapiro–Wilk W, Anderson–Darling A^2 (see ►Anderson-Darling Tests of Goodness-of-Fit), and the ►Jarque–Bera test. Specifics on these and many other normality tests and their characteristics can be found in Thode (2002) and on general goodness of fit issues, including normality tests, in D'Agostino and Stephens (1986).

About the Author

Dr. Henry C. Thode, Jr. is an Associate Professor In the Department of Emergency Medicine and Associate Director of Research of the Emergency Medicine Research Center at Stony Brook University, NY. He has authored and co-authored more than 110 papers and book chapters, and has authored the text *Testing for Normality* (Marcel-Dekker, 2002).

Cross References

►Anderson-Darling Tests of Goodness-of-Fit
►Jarque-Bera Test
►Kurtosis: An Overview
►Normal Distribution, Univariate
►Normality Tests: Power Comparison
►Omnibus Test for Departures from Normality
►Skewness
►Tests of Fit Based on The Empirical Distribution Function

References and Further Reading

Anderson TW, Darling DA (1954) A test of goodness of fit. J Am Stat Assoc 49:765–769
Box GEP (1953) Non-normality and tests on variances. Biometrika 40:318–335
D'Agostino RB, Lee AFS (1977) Robustness of location estimators under changes of population kurtosis. J Am Stat Assoc 72: 393–396
D'Agostino RB, Stephens MA (eds) (1986) Goodness-of-fit techniques. Marcel Dekker, New York
David HA, Hartley HO, Pearson ES (1954) The distribution of the ratio, in a single normal sample, of the range to the standard deviation. Biometrika 41:482–493
Franck WE (1981) The most powerful invariant test of normal versus Cauchy with applications to stable alternatives. J Am Stat Assoc 76:1002–1005
Geary RC (1936) Moments of the ratio of the mean deviation to the standard deviation for normal samples. Biometrika 28: 295–305
Geary RC (1947) Testing for normality. Biometrika, 34:209–242
Jarque C, Bera A (1980) Efficient tests for normality, homoskedasticity and serial independence of regression residuals. Econ Lett 6:255–259
Jarque C, Bera A (1987) A test for normality of observations and regression residuals. Int Stat Rev 55:163–172
Johnson ME, Tietjen GL, Beckman RJ (1980) A new family of probability distributions with application to Monte Carlo studies. J Am Stat Assoc 75:276–279
Pearson ES, Please NW (1975) Relation between the shape of population distribution and the robustness of four simple test statistics. Biometrika 62:223–241
Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52:591–611
Subrahmaniam K, Subrahmaniam K, Messeri JY (1975) On the robustness of some tests of significance in sampling from a compound normal distribution. J Am Stat Assoc 70:435–438
Thode HC Jr (2002) Testing for normality. Marcel-Dekker, New York
Tukey JW (1960) A survey of sampling from contaminated distributions. In: Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB (eds) Contributions to probability and statistics. Stanford University Press, CA, pp 448–485
Uthoff VA (1970) An optimum test property of two well-known statistics. J Am Stat Assoc 65:1597–1600
Uthoff VA (1973) The most powerful scale and location invariant test of the normal versus the double exponential. Ann Stat 1:170–174

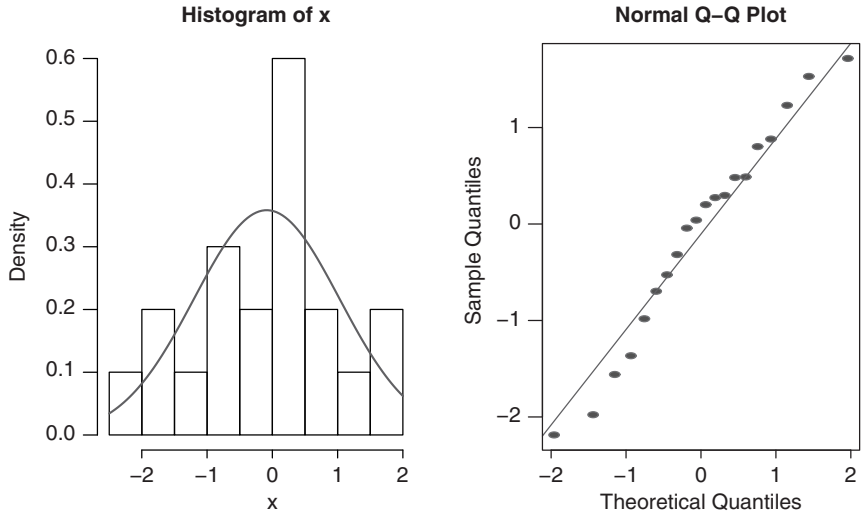
Normality Tests: Power Comparison

Edith Seier
Professor
East Tennessee State University, Johnson City, TN, USA

The assumption of normality is required by several methods in parametric statistical inference, some of which are robust toward mild or moderate non-normality. The histogram in Fig. 1a was prepared with 20 values randomly generated with the standard normal distribution; however, one could argue that the normality of the variable is not evident from the histogram. If these were real observations instead of generated data, a test would be applied with the null hypothesis being that the variable has a normal distribution; frequently no specific distribution is mentioned as alternative hypothesis. The normal quantile plot (Fig. 1b) compares the ordered values $x_{(i)}$ in the sample, also called ►order statistics, with the n corresponding quantiles of the normal distribution. If the sample comes from a normal distribution, the dots tend to suggest a linear pattern. Other versions of the quantile plot do exist.

Numerous tests for normality have been defined. Tests are generally compared in terms of their power, i.e. the probability of finding out when the sample does not come from a normal distribution; every few years a systematic comparison of tests is published. A recent and comprehensive comparison that involves thirty three different tests for normality can be found in Romão et al (2010). Previous comparisons include those of Gan and Koehler (1990) and Seier (2002), among others. All those articles provide tables that compare the power of tests against several non-normal distributions for different sample sizes. That detailed information is useful when we want to test for normality, specially if we have in mind a certain type of alternative distribution. Historical information and a discussion on power comparison is included in Chap. 7 of Thode (2002). There is not one single test that is the most powerful against all the possible non-normal distributions. There are tests that are more powerful against skewed distributions, or tests that are more powerful against distributions that are symmetric but have heavier tails than the normal. However, if the practitioner does not have in mind a specific type of non-normal distribution, a test that is more powerful on average against a wide range of distributions is usually preferred.

To test the hypothesis of normality, based on a data set, one can use one of various statistical software or write a computer program if the test of our preference is not



Normality Tests: Power Comparison. Fig. 1 Histogram and normal quantile plot for 20 values generated with a normal distribution

implemented by the available software. Different tests will produce different ▶p-values for the same data set. For the data in Fig. 1, the p-values for four different tests are as follows: Shapiro–Wilk 0.7277, Pearson 0.4337, Anderson–Darling 0.7988, Lilliefors (Kolmogorov–Smirnov) 0.691. In this example, the four tests agree in not rejecting the null hypothesis of normality. However, for a different example some tests might reject the hypothesis of normality and some others might not, it is from such cases that the differences in power originate.

A special program needs to be written in order to systematically compare tests for normality in terms of power. A non-normal distribution is used to generate a large number of samples of a given size and each one of the tests is applied to each sample, keeping track of the number of times that the null hypothesis is rejected. The empirical power of the normality test, when the true distribution is that specific one, is the number of times that the null hypothesis is rejected divided by the number of samples generated. To make the comparison fair, tests are first compared in terms of their control of α. The proportion of samples for which the hypothesis of normality is rejected, when the samples are generated by a normal distribution, should be as close as possible to the nominal probability (α) of type I error.

The normal distribution has well known density and distribution functions, it is symmetric, has Pearson's kurtosis equal to 3, and if $X \sim N(\mu, \sigma^2)$ then $X = \mu + \sigma Z$ where $Z \sim N(0,1)$. To define a new test of normality one usually focuses on one or more of those characteristics. Tests might use Pearson's skewness and kurtosis statistics or some other ones. The empirical distribution function can be compared with the normal CDF using different criteria. Some tests compare the order statistics with the expected values of the order statistics under normality, the summary of that comparison can be done in different manners. Some tests rely on concepts such as ▶entropy or ▶likelihood. After the test statistic is defined, it is necessary to derive the distribution of the statistic under normality in order to calculate either critical values or p-values. Some tests work with exact distributions that require special tables, some others strive for simplicity using one of the standard statistical distributions as an approximation (sometimes sacrificing a little power in the process). The diverse possibilities in the definition of the tests explains in part why there are so many tests for normality and why they might differ in power depending on the true distribution of the variable. Thode (2002) describes thirty tests of normality and in recent years new ones have been defined.

Most tests of normality have high power when the sample is large, it is the small and moderate sample sizes the ones that provide the setting for interesting comparisons among tests. The more different from the normal the true distribution looks like, the smaller the sample size needed to achieve high power. For example, the Student's t-distributions look very similar to the normal except that they have heavier tails, as the number of degrees of freedom n increases the $t_{(n)}$ and the normal look more alike. A certain normality test requires a sample size 150 to have a power 0.793 when the true distribution is $t_{(5)}$, the same test requires a sample size of only 35 to have power 0.812 when the true distribution is $t_{(2)}$ (Bonett and Seier 2002).

When the true distribution is severely skewed, e.g. the log-normal(0,1), there are several tests that have a power larger than 0.95 when the sample size is 25.

Considering the way tests are defined helps to understand why some of them have more power than others. One of the first approaches used to test for normality was to apply the Chi-square goodness of fit test, comparing the observed frequency of each interval with the expected frequency (something similar to the visual comparison of the histogram to the normal curve in Fig. 1a) Currently more powerful tests are available.

Kurtosis is a characteristic of distributions related to the heaviness of the tails and the concentration of mass toward the center of the distribution. There are some tests defined specifically to test for normality against symmetric distributions that have heavier tails or large kurtosis; these distributions are of interest in the fields of economics and finance. Two such tests are the ones defined by Bonnet and Seier (2002) and Gel et al. (2007), both represent kurtosis with a ratio that has the standard deviation in the numerator and another measure of spread in the denominator. They have good power when the true distribution is symmetric with high kurtosis such as Laplace, $t_{(2)}$, some *SU* distributions, Cauchy and some scale contaminated normals where the contaminating distribution has smaller variance. The kurtosis test in D'Agostino et al. (1990) has good performance against symmetric distributions with kurtosis moderately higher than 3 and against some scale contaminated distributions where the contaminating distribution has larger variance. It is recommended to accompany kurtosis based tests by a test of symmetry.

There is a class of tests that first summarize the empirical distribution using some skewness and kurtosis statistics and then compare their values with the skewness and kurtosis of the normal distribution. Tests described by D'Agostino et al. (1990) and the ▶Jarque–Bera test, popular in econometrics, are based on Pearson's skewness and kurtosis statistics. A modification to the Jarque–Bera test (Gel and Gastwirth 2007) uses a robust measure of variability and improves the power of the JB test when the true distribution is symmetric with moderately higher kurtosis than the normal and the sample size is small.

The Anderson–Darling (see ▶Anderson-Darling Tests of Goodness-of-Fit) and Kolmogorov–Smirnov tests are both based on the comparison of the empirical distribution function and the normal CDF. The Lilliefors test is the version of the ▶Kolmogorov–Smirnov test that uses the sample mean and variance. The Lilliefors or KS test focuses on the maximum difference between the empirical and theoretical distribution. The Anderson–Darling test summarizes all the differences between the empirical and theoretical distributions, tending to be more powerful than the KS test. The Shapiro–Wilk test is more powerful than the Anderson–Darling test against many non-normal distributions, but the AD test is more powerful than the SW when the true distribution is the Tukey (10), a distribution that has a very sharp peak and tails that end abruptly.

There is a group of tests of normality called regression type tests. Those tests, in one way or other, compare one by one sample quantiles with quantiles from the normal distribution. The most well known of these tests is the Shapiro–Wilk test, defined in terms of the order statistics and the expected values of the order statistics. Originally defined in 1965 for $n \leq 50$, it required special constants and critical values. Several extensions and modifications of that test have been published since then, such as the one in Royston (1992). The Chen–Shapiro test (1995) compares the spacings in between the quantiles in both the empirical and the normal distribution. The Chen–Shapiro test is also quite powerful against a wide range of non-normal distributions, it is even slightly more powerful than the Shapiro–Wilk test against some distributions when the sample size is small. When working with rounded data, both the CS test and the SW test benefit from the adjustment for ties proposed by Royston (1989).

Romão et al. (2010) conclude that when testing normality against symmetric alternatives, the best choices are the already mentioned tests by Gel et al. (2007), Bonett and Seier (2002), Chen and Shapiro (1995), and the test defined by Coin (2008). The first two are kurtosis based tests and the last two are regression type tests that work with the order statistics. Coin's test consists in fitting a polynomial model to a standardized version of the normal quantile plot and testing for the coefficient ($Ho : \beta_3 = 0$) of the third power of the expected values of the order statistics under normality. To test for normality against skewed distributions they recommend the Chen–Shapiro test, Shapiro–Wilk test and two tests based on the likelihood ratio defined by Zhang and Wu (2005). To detect normal distributions with ▶outliers Romão et al. (2010) recommend tests based on a trimmed version (defined in 2003 by Elamir and Seheult) of Hosking's L-moments, which are considered to be less sensitive to outliers than the classical moments. To test for normality in general without having a particular type of alternative distribution in mind, Romão et al. (2010) recommend: Shapiro–Wilk, Chen–Shapiro and Del Barrio et al. (2005), the three of them are regression type tests.

Most statistical software include some tests of normality. However, some powerful tests have not been implemented by statistical software yet. The base package of

R calculates the Shapiro–Wilk test and five more tests with the *normtest* package: Anderson–Darling, Cramer–von Mises, Kolmogorov–Smirnov (Lilliefors), Pearson chi-square test, and Shapiro–Francia. MINITAB has three options: KS, AD and Ryan–Joiner (similar to SW). SPSS calculates KS(L) and SW. SAS calculates SW, CM, AD and KS(L).

Cross References

►Anderson-Darling Tests of Goodness-of-Fit
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Jarque-Bera Test
►Kolmogorov-Smirnov Test
►Normal Distribution, Univariate
►Normality Tests
►Omnibus Test for Departures from Normality
►Power Analysis
►R Language

References and Further Reading

Bonett DG, Seier E (2002) A test of normality with high uniform power. Comput Stat Data An 40:435–445
Chen L, Shapiro S (1995) An alternative test for normality based on normalized spacings. J Stat Comput Sim 53:269–287
Coin DA (2008) Goodness-of-fit test for normality based on polynomial regression. Comput Stat Data An 52: 2185–2198
D'Agostino RB, Belanger A, D'Agostino RB Jr (1990) A suggestion for using powerful and informative tests of normality. Am Stat 44:316–322
Del Barrio E, Giné E, Utzet F (2005) Asymptotics for *L*2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. Bernoulli 11: 131–189
Gan FF, Koehler KJ (1990) Goodness of fit tests based on P-P probability plots. Technometrics 32:289–303
Gel YR, Gastwirth JL (2008) A robust modification of the JarqueBera test of normality. Econ Lett 99:30–32
Gel YR, Miao W, Gastwirth JL (2007) Robust directed test of normality against heavy-tailed alternatives. Comput Stat Data An 51:2734–2746
Romão X, Delgado R, Costa A (2010) An empirical power comparison of univariate goodness-of-fit tests for normality. J Stat Comp Sim 80:545–591
Royston P (1989) Correcting the Shapiro–Wilk W for ties. J Stat Comput Sim 31:237–249
Royston P (1992) Approximating the Shapiro–Wilk W-test for non-normality. Stat Comput 2:117–119
Seier E (2002) Comparison of tests for univariate normality. Interstat, January 2002
Thode HC (2002) Testing for normality. Marcel Dekker, New York
Zhang J, Wu Y (2005) Likelihood-ratio tests for normality. Comput Stat Data An 49:709–721

Null-Hypothesis Significance Testing: Misconceptons

Raymond S. Nickerson
Research Professor
Tufts University, Medford, MA, USA

Null-hypothesis significance testing (NHST) has for many years been the most widely used statistical tool for evaluating the outcomes of psychological experiments. It is routinely taught to college students in elementary statistics courses and courses in experimental methodology and design. Despite these facts, there are many misconceptions about null-hypothesis significance testing—about what conclusions the results of such testing do or do not justify. Here several of these misconceptions are briefly summarized. More substantive treatments of these and related misconceptions may be found in several publications, including Rozeboom (1960), Clark (1963), Bakan (1966), Morrison and Henkel (1970), Carver (1978), Lakatos (1978), Berger and Sellke (1987), Falk and Greenbaum (1995), Gigerenzer (1998), and Wilkinson and APA Task Force on Statistical Inference (1999).

Many criticisms have been directed at the use of NHST. Some of these criticisms challenge the logic on which it is based; some contend that its use is not productive because it does not advance an understanding of the phenomena of interest; others focus on what are seen as arbitrary aspects of the rules for its use and interpretation of the statistics it provides; still others contend that its widespread use has undesirable effects on the way experiments are designed and on the reporting – or the failure to report – experimental results.

Although *null hypothesis* has more than one connotation as it has been used by different people in a variety of contexts, the meaning that is taken for purposes of this article is one that is likely to be found in statistics texts: The hypothesis that two groups – usually an experimental group and a control group – do not differ with respect to a specified quantitative measure of interest, and that any observed difference between the means of that measure is due strictly to chance. Assuming the data obtained from the two groups satisfy certain assumptions underlying the use of the null-hypothesis test, the test yields a statistic and a p value, the latter of which indicates the probability that the value of the statistic, or a larger one, would be obtained if the observed difference between the means were due strictly to chance. Conventionally, the null hypothesis is rejected when the value of p that is obtained is less than some specified criterion, generally referred to

as *alpha*, and typically set at .05 or .01. The theory admits of only two possible outcomes of null-hypothesis testing: either the hypothesis is rejected on the basis of obtaining a *p* value not exceeding *alpha*, or it is not rejected if the *p* value does exceed *alpha*. Never, strictly speaking, is the null hypothesis said to have been shown to be true. The obvious bias in the test reflects the view that rejecting the null hypothesis when it is true (generally referred to as a *Type-I* error) is a much less acceptable outcome of testing than is failing to reject it when it is false (*Type-II* error).

Apparently null-hypothesis statistical testing is easily misunderstood, and the misunderstandings can take a variety of forms. Space does not permit an exhaustive, or even extensive, listing and explanation of all the misunderstandings that can occur. This note focuses on certain misconceptions that are fairly common. The evidence of their commonality is found in numerous published reports of experiments in which NHST has been used in the data analyses. The format for what follows is a statement of each of several misconceptions, accompanied by a brief explanatory comment. The misconceptions noted have been discussed in more detail in Nickerson (2000); many have also been discussed in other publications cited in that article and in the reference list for this one.

- *The value of p is the probability that the null hypothesis is true and that of 1-p is the probability that the alternative to the null hypothesis is true*. Given this belief, a *p* value of .001 would be taken to mean that there is one chance in one-thousand that the obtained difference between the means is due to chance, and that it is almost certain that the alternative to the null hypothesis – the hypothesis that the observed difference is not due to chance – is true. This misconception is an example of confusion between the probability of *A* conditional on *B*, $p(A|B)$, and the probability of *B* conditional on *A*, $p(B|A)$. Letting *A* represent "the observed value of the statistic is *X*," and *B* "the null hypothesis is true," obtaining a *p* of .001 means that $p(A|B) = 0.001$; it does not mean that $p(B|A) = 0.001$.
- *Rejection of the null hypothesis establishes the truth of a theory that predicts it to be false*. The reasoning, expressed as a conditional syllogism is: *If the theory is true, the null hypothesis will prove to be false; the null hypothesis proved to be false, therefore the theory is true*. This is a case of *affirming the consequent*, a common fallacy in conditional reasoning.
- *A small p is evidence that the results are replicable* (*sometimes accompanied by the belief that 1-p is the probability that the results will replicate*). A case can be made that a small *p* indicating a statistically significant difference justifies a greater expectation than does a large *p* that a follow-up experiment under the same conditions will also produce a statistically significant difference, but a small *p* does not guarantee that the results will replicate, and 1-*p* is not the probability that they will do so.
- *A small p means a large treatment effect*. Although, other things equal, a large effect is likely to yield a small *p*, and a small *p* is suggestive of a large effect, *p* is also sensitive to sample size (with a large sample, a small effect can yield a small *p*), so a small *p* is not a reliable indication of a large effect.
- *Statistical significance means theoretical or practical significance*. Statistical significance indicates only that two samples selected at random from the same population are unlikely to produce the observed result; it reveals nothing about whether the finding has any theoretical or practical significance.
- *Alpha is the probability that if one has rejected the null hypothesis, one has made a Type-I error*. This is another case of confusing $p(A|B)$ and $p(B|A)$. *Alpha* – or more specifically $p \leq alpha$ – is the probability of rejecting the null hypothesis, given that the hypothesis is true, which is different from the probability that the null hypothesis is true, given that one has rejected it.
- *The value of alpha selected for a given experiment is the probability that a Type-I error will be made in interpreting the results of that experiment*. *Alpha* is the probability of making a Type-I error when the null hypothesis is true; a Type-I error cannot be made when the null hypothesis is false. So *alpha* could be the probability of a Type-I error only on the assumption that the null hypothesis is always true, but presumably for most experiments there are prior reasons to believe that the null hypothesis is false.
- *The value of alpha is the probability of Type-I error across a large set of experiments in which alpha is set at the same value*. The counter for this misconception is similar to that for the last one. It would be true only on the assumption that across the set of experiments of interest the null hypothesis is always true.

Just as there are faulty beliefs about Type-I errors and *alpha*, there are faulty beliefs about Type-II errors and *beta*. A Type-II error is made when a null hypothesis that is false is not rejected; *beta* is the probability of failing to reject the null hypothesis, given that it is false. Sometimes *beta* is taken to be the probability that the null hypothesis is false, given that it has not been rejected, and sometimes it is taken to be the absolute probability of making a Type-II

error. Both of these beliefs are wrong, and for reasons comparable to those given for the corresponding beliefs about Type-I errors and *alpha*.

- *Failing to reject the null hypothesis is equivalent to demonstrating it to be true.* In view of the convention of selecting a very small value for *alpha* (0.05, 0.01), thus, in effect, setting a high bar for rejecting the null hypothesis, failure to reach this bar to claim the hypothesis to be false is very weak evidence that it is true. Nevertheless, experimenters often take the failure to find a statistically significant difference between two groups as sufficient evidence to proceed as though no difference exists.
- *Failure to reject the null hypothesis is evidence of a failed experiment.* Probably most experiments for which null hypothesis testing is used to analyze data are planned and run with the expectation that the results will reveal a hypothesized difference, or differences, of interest .If a difference that is sufficiently distinct to permit rejection of the null hypothesis is not obtained, it may be because the experiment was not adequately planned or executed, in which case one might argue that it should be considered a failed experiment. On the other hand, it may be too that there is no difference to be found. The failure to reject the null hypothesis, by itself, does not suffice to distinguish between these possibilities.

The focus here has been on misconceptions of NHST. Critics of the use of NHST in the analysis and interpretation of data from psychological experiments have raised other issues as well, discussion of which is beyond the scope of this article. It must be noted, too, that NHST has many defenders (Abelson 1995; Baril and Cannon 1995; Chow 1996; Cortina and Dunlap 1997; Dixon 1998; Frick 1996; Harris 1997; Mulaik et al. 1997; Wilson et al. 1967; Winch and Campbell 1969), but defense of the NHST is also beyond the scope of this article. Defenders of NHST generally acknowledge that it is often misunderstood and misapplied, but contend that it is an effective tool when used appropriately.

About the Author

Raymond S. Nickerson is a research professor at Tufts University, from which he received a PhD in experimental psychology, and is retired from Bolt Beranek and Newman (BBN), where he was a senior vice president. He is a Fellow of the American Association for the Advancement of Science, the American Psychological Association, the Association for Psychological Science, the Human Factors and Ergonomics Society, and the Society of Experimental Psychologists. He is a past Chair of the U.S. National Research Council's Committee on Human Factors and a recipient of the Franklin V. Taylor Award from the American Psychological Association. Dr. Nickerson was the Founding Editor of *The Journal of Experimental Psychology: Applied* and of *Reviews of Human Factors and Ergonomics*, an annual publication of the Human Factors and Ergonomics Society. He is author or co-author of about 200 published works, including eight books, the latest of which is *Mathematical Reasoning: Patterns, Problems, Conjectures, and Proofs* (Psychology Press, 2010).

Cross References

►Effect Size
►Frequentist Hypothesis Testing: A Defense
►Psychology, Statistics in
►P-Values
►Role of Statistics
►Significance Testing: An Overview
►Significance Tests, History and Logic of
►Significance Tests: A Critique
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Inference: An Overview
►Statistics: Controversies in Practice

References and Further Reading

Abelson RP (1995) Statistics as principled argument. Erlbaum, Hillsdale
Bakan D (1966) The test of significance in psychological research. Psychol Bull 66:1–29
Baril GL, Cannon JT (1995) What is the probability that null hypothesis testing is meaningless? Am Psychol 50:1098–1099
Berger JO, Sellke T (1987) Testing a point null hypothesis: The irreconcilability of P values and evidence. J Am Stat Assoc 82:112–122
Carver RP (1978) The case against statistical significance testing. Harvard Educ Rev 48:378–399
Chow SL (1996) Statistical significance: rationale, validity, and utility. Sage, Beverly Hills, CA
Clark CA (1963) Hypothesis testing in relation to statistical methodology. Rev Educ Res 33:455–473
Cortina JM, Dunlap WP (1997) On the logic and purpose of significance testing. Psychol Meth 2:161–172
Dixon P (1998) Why scientists value *p* values. Psychon Bull Rev 5:390–396
Falk R, Greenbaum CW (1995) Significance tests die hard: The amazing persistence of a probabilistic misconception. Theory Psychol 5:75–98
Frick RW (1996) The appropriate use of null hypothesis testing. Psychol Meth 1:379–390
Gigerenzer G (1998) Surrogates for theories. Theory Psychol 8:195–204
Harris RJ (1997) Significance tests have their place. Psychol Sci 8:8–11

Lakatos I (1978) Falsification and the methodology of scientific research programmes. In: Worral J, Currie G (eds) The methodology of scientific research programs: Imre Lakatos' philosophical papers, vol 1. Cambridge University Press, Cambridge, UK
Morrison DE, Henkel RE (eds) (1970) The significance test controversy: A reader. Aldine, Chicago
Mulaik SA, Raju NS, Harshman RA (1997) There is a time and place for significance testing. In: Harlow LL, Mulaik SA, Steiger JH (eds) What if there were no significance tests? Erlbaum, Mahwah, NJ, pp 65–116
Nickerson RS (2000) Null hypothesis statistical testing: A review of an old and continuing controversy. Psychol Meth 5:241–301
Rozeboom WW (1960) The fallacy of the null hypothesis significance test. Psychol Bull 57:416–428
Wilkinson L, APA Task Force on Statistical Inference (1999) Statistical methods in psychology journals: Guidelines and explanations. Am Psychol 54:594–604
Wilson W, Miller HL, Lower JS (1967) Much ado about the null hypothesis. Psychol Bull 68:188–196
Winch RF, Campbell DT (1969) Proof? No. Evidence? Yes. The significance of tests of significance. Am Sociol 4:140–143

Numerical Integration

James E. Gentle
Professor of Computational Statistics
George Mason University, Fairfax, VA, USA

One of the most common mathematical operations in scientific computing is quadrature, the evaluation of a definite integral. It is used to determine volume, mass, or total charge, for example. In the evaluation of probabilities, of expectations, and of marginal or conditional densities, integration is the basic operation.

Most of the integrals and differential equations of interest in real-world applications do not have closed-form solutions; hence, their solutions must be approximated or estimated numerically.

The general problem is to approximate or estimate

$$I = \int_D f(x)\,\mathrm{d}x. \tag{1}$$

There are basically two approaches. One is based on sums of integrals of approximations of the integrand over subregions of the domain:

$$\int_D f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n} \int_{D_i} \tilde{f}_i(x)\,\mathrm{d}x, \tag{2}$$

where $\cup_{i=0}^{n} D_i = D$ and $\tilde{f}_i(x) \approx f(x)$ within D_i.

In the other approach, the integrand is decomposed into a probability density function (PDF) and another factor:

$$\int_D f(x)\,\mathrm{d}x = \int_D h(x) p_X(x)\,\mathrm{d}x, \tag{3}$$

where p_X is the probability density function of a random variable X with support on D. In this formulation, we have

$$I = \mathrm{E}(h(X)), \tag{4}$$

where $\mathrm{E}(\cdot)$ represents the expectation operator with respect to the probability distribution of X. Notice that this is exact.

The second approach, called Monte Carlo integration, is fundamentally different from the first, because it leads to a statistical *estimate* rather than a *numerical approximation*.

Each approach has many variations and there are many details that we cannot address in the brief space here. The references at the end of this article contain fuller descriptions.

Numerical Approximations

One type of numerical approximation is based on direct approximation of the Riemann sum, which we take as the basis for the definition of the integral. Another type of approximation is based on an approximation of the function using one of the methods discussed above. We begin with approximations that are based on Riemann sums. We also generally limit the discussion to univariate integrals.

Although some of the more interesting problems are multivariate and the region of integration is not rectangular, we begin with the simple integral,

$$I = \int_a^b f(x)\,\mathrm{d}x. \tag{5}$$

Newton–Cotes Quadrature

The Riemann integral is defined as the limit of the *Riemann sums*:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - x_{i-1}) f(\widetilde{x}_i), \tag{6}$$

where $a = x_0 < x_1 < \cdots < x_n = b$ and $\widetilde{x}_i \in [x_{i-1}, x_i]$.

Instead of a simple step function, the function $f(x)$ may be approximated by a piecewise linear function $p_1(x)$ that agrees with f at each of the points. The integral (5) can be approximated by a sum of integrals, each of which is particularly easy to evaluate:

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &\approx \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} p_1(x)\,\mathrm{d}x,\\ &= h\big(f(a) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{n-1})\\ &\quad + f(b)\big)/2.\end{aligned} \tag{7}$$

The expression (7) is called the *trapezoid rule*.

Many other quadrature rules can be built using this same idea of an approximating function that agrees with f at each of some set of points. Quadrature formulas that result from this kind of approach are called *Newton–Cotes formulas*.

Rather than the linear functions of the trapezoid rule, a more accurate approximation would probably result from use of polynomials of degree k that agree with f at each of $k+1$ successive points. Use of polynomials in this way leads to what are called *Simpson's rules*.

Error in Newton–Cotes Quadrature

There are generally multiple sources of error. The error in approximations must be considered separately from the error in rounding, although at some level of discretization, the rounding error may prevent any decrease in approximation error, even though the approximation is really the source of the error.

In Newton–Cotes quadrature, we get an expression for the error of the general form $O(g(h))$.

To approximate the error, we consider a polynomial of degree n over that region, because we could have a single such polynomial that corresponds to f at each of the break points. Using a Taylor series, we find that the error in use of the trapezoid rule can then be expressed as

$$-\frac{1}{12}(b-a)h^2 f''(x^*)$$

for some $x^* \in [a,b]$. This is not very useful in practice. It is important, however to note that the error is $O(h^2)$.

Using similar approaches we can determine that the error for Simpson's rules is $O(h^4)$.

Extrapolation in Quadrature Rules

We can use Richardson extrapolation (see, for example, Gentle 2009) to improve the approximation in Newton–Cotes formulas. In the trapezoid rule, for example, we consider various numbers of intervals. Let T_{0k} represent the value of the expression in Eq. 7 when $n = 2^k$; that is, when there are n intervals, and we examine the formula for $1, 2, 4, \ldots$ intervals, that is, $T_{00}, T_{01}, T_{02}, \ldots$. Now, we use Richardson extrapolation to form

$$T_{1,k} = (4T_{0,k+1} - T_{0,k})/3. \tag{8}$$

Generalizing this, we define

$$T_{m,k} = (4^m T_{m-1,k} - T_{m-1,k-1})/(4^m - 1). \tag{9}$$

Notice in $T_{m,k}$, m represents the extent of extrapolation, and k determines the number of intervals. (This kind of scheme is used often in numerical analysis. It can be represented as a triangular table in which the ith row consists of the i terms $T_{0,i-1}, \ldots, T_{i-1,0}$.)

This application of Richardson extrapolation in quadrature with the trapezoid rule is called *Romberg quadrature*.

Adaptive Quadrature Rules

It is not obvious how to choose the interval width in Newton–Cotes formulas. Obviously if the interval width is too large, finer structure in the integrand will be missed. On the other hand, if the interval width is too small, in addition to increased cost of evaluation of the integrand, rounding error can become significant. There are various ways of trying to achieve a balance between accuracy and number of function evaluations. In most cases these involve approximation of the integral over different subintervals with different widths used in each of the subintervals. Initially, this may identify subintervals of the domain of integration that require smaller widths in the Newton–Cotes formulas (that is, regions in which the integrand is rougher). Evaluations at different widths over the different subintervals may lead to a good choice of both subintervals and widths within the different subintervals. This kind of approach is called *adaptive quadrature*.

Gaussian Quadrature

We now briefly discuss another approach to the evaluation of the integral (5) called *Gaussian quadrature*. Gaussian quadrature uses the idea of expansion of the integrand. Like the Newton–Cotes approaches, Gaussian quadrature arises from the Riemann sum (6), except here we interpret the interval widths as weights:

$$\sum_{i=0}^{n} w(x_i) f(x_i). \tag{10}$$

In Gaussian quadrature, we put more emphasis on choosing the points, and by so doing, we need a smaller number of points. Whether or not this is a good idea of course depends on how we choose the points and how we define $w(x_i)$.

Determination of Weights in Gaussian Quadrature

If f is a polynomial of degree $2n-1$, it is possible to represent the integral $\int_a^b f(x)\mathrm{d}x$ exactly in the form (10). Use of the formula (10) yields $2n$ equations in $2n$ unknowns to determine the x_i's and w_i's. In Gaussian quadrature, the x_i's and w_i's are chosen so that the approximation is correct when f is a polynomial, and it can provide a approximation in many cases with a relatively small n. Often only five or six points provide a good approximation.

To make this a useful method for any given (reasonable) integrand over a finite range, the obvious approach is to represent the function as a series in a standard sequence of orthogonal polynomials.

Error in Gaussian Quadrature

The error in Gaussian quadrature is

$$\int_a^b f(x)\,\mathrm{d}x - \sum_{i=0}^{n} w_i g(x_i),$$

which we can write as

$$\int_a^b g(x)w(x)\,\mathrm{d}x - \sum_{i=0}^{n} w_i g(x_i) = \frac{g^{(2n)}(x^*)}{(2n)!c_n^2}, \qquad (11)$$

for some point x^* in (a, b).

A problem with Gaussian quadrature is that it is not easy to use the results for n to compute results for $\tilde{n}$, and hence the kinds of extrapolation and adaptation we discussed above for Newton–Cotes quadrature are not very useful for Gaussian quadrature.

Monte Carlo Methods for Quadrature

In the Monte Carlo method of quadrature we first formulate the integral to be evaluated as an expectation of a function of a random variable, as in Eq. 4. To estimate this expected value, that is, the integral, we simulate realizations of the random variable, and take the average of the function evaluated at those realizations,

$$\widehat{I} = \frac{1}{m}\sum_{i=1}^{m} h(x_i), \qquad (12)$$

where $x_1, x_2, \ldots, x_m$ is a random sample (or pseudorandom sample) of the random variable X.

If we formulate the estimator $\widehat{I}$ as a sum of functions of independent random variables, each with density p_X, instead of a sum of realizations of random variables, the estimator itself is a random variable. An obviously desirable property of this random variable is that its expectation be equal to the quantity being estimated. Assuming the expectations exist, this is easily seen to be the case:

$$\mathrm{E}\left(\widehat{I}\right) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}(h(X_i)) = \frac{1}{m}\sum_{i=1}^{m} I = I.$$

We therefore say the estimator is unbiased.

An advantage of Monte Carlo quadrature is that the nature of the domain of integration is not as critical as in the other quadrature methods we have discussed above. Monte Carlo quadrature can be performed equally easily for improper integrals as for integrals over finite domains.

Another advantage of Monte Carlo quadrature is that the computations can be performed in parallel without any special coding.

Variance of Monte Carlo Estimators

Monte Carlo methods are sampling methods; therefore the estimates that result from Monte Carlo procedures have associated *sampling errors*.

In the case of scalar functions, the variance of the estimator $\widehat{I}$ is a rather complicated function involving the original integral (assuming the integrals exist):

$$\begin{aligned}\mathrm{V}\left(\widehat{I}\right) &= \frac{1}{m}\mathrm{E}\left(\left(h(X) - \mathrm{E}(h(X))\right)^2\right)\\ &= \tfrac{1}{m}\int_D \left(h(x) - \int_D h(y)p_X(y)\,\mathrm{d}y\right)^2 p_X(x)\,\mathrm{d}x.\end{aligned} \qquad (13)$$

We see that the magnitude of the variance depends on the variation in

$$h(x) - \int_D h(y)p_X(y)\,\mathrm{d}y,$$

which depends in turn on the variation in $h(x)$. If $h(x)$ is constant, the variance of $\widehat{I}$ is 0. Of course, in this case, we do not need to do the Monte Carlo estimation; we have the solution $I = h(\cdot)\int_D \mathrm{d}y$.

While the variance in (13) is complicated, we have a very simple estimate of the variance; it is just the sample variance of the observations $h(x_i)$.

Reducing the Variance

As we see from Eq. 13 the variance of the Monte Carlo estimator is linear in m^{-1}; hence, the variance is reduced by increasing the Monte Carlo sample size. More effective methods of variance reduction include use of antithetic variates, importance sampling, and stratified sampling (see Gentle 2003).

Error in Monte Carlo Quadrature

As we have emphasized, Monte Carlo quadrature differs from quadrature methods such as Newton–Cotes methods and Gaussian quadrature in a fundamental way; Monte Carlo methods involve random (or pseudorandom) sampling. The expressions in the Mont Carlo quadrature formulas do not involve any approximations, so questions of bounds of the error of approximation do not arise. Instead of error bounds or order of the error as some function of the integrand, we use the variance of the random estimator to indicate the extent of the uncertainty in the solution.

The square root of the variance, that is, the standard deviation of the estimator, is a good measure of the

range within which different estimators of the integral may fall.

Because of the dependence of the confidence interval on the standard deviation the standard deviation is sometimes called a "probabilistic error bound." The word "bound" is misused here, of course, but in any event, the standard deviation does provide some measure of a sampling "error."

The important thing to note from Eq. 13 is the order of error in the Monte Carlo sample size; it is $O\left(m^{-\frac{1}{2}}\right)$. This results in the usual diminished returns of ordinary statistical estimators; to halve the error, the sample size must be quadrupled.

Higher Dimensions

The most significant difficulties in numerical quadrature occur in multiple integration.

The Monte Carlo quadrature methods extend directly to multivariate integrals, although, obviously, it takes larger samples to fill the space. It is, in fact, only for multivariate integrals that Monte Carlo quadrature should ordinarily be used. The preference for Monte Carlo in multivariate quadrature results from the independence of the pseudoprobabilistic error bounds and the dimensionality mentioned above.

An important fact to be observed in Eq. 13 the order of the error in terms of the number of function evaluations is independent of the dimensionality of the integral so the order of the error remains $O\left(m^{-\frac{1}{2}}\right)$. On the other hand, the usual error bounds for numerical quadrature are $O\left((g(n))^{-\frac{1}{d}}\right)$, where d is the dimensionality, and $g(n)$ is the order for one-dimensional quadrature. This is one of the most important properties of Monte Carlo quadrature.

The papers in the book edited by Flournoy and Tsutakawa (1991) provide good surveys of specific methods for multiple integrals, especially ones with important applications in statistics. Evans and Schwartz (2000) provide a good summary of methods for numerical quadrature, including both the standard deterministic methods of numerical analysis and Monte Carlo methods.

About the Author

James E. Gentle is University Professor of Computational Statistics in the Department of Computational and Data Sciences at George Mason University. He is a Fellow of the American Statistical Association and of the American Association for the Advancement of Science, and he is an elected member of the International Statistical Institute. He is author of *Computational Statistics* (Springer, 2009) and of other books on statistical computing.

Cross References

►Computational Statistics
►Monte Carlo Methods in Statistics

References and Further Reading

Evans M, Schwartz T (2000) Approximating integrals via Monte Carlo and deterministic methods. Oxford University Press, Oxford, UK

Flournoy N, Tsutakawa RK (1991) Statistical multiple integration. American Mathematical Society, Providence, Rhode Island

Gentle JE (2003) Random number generation and Monte Carlo methods. Springer, New York

Gentle JE (2009) Computational statistics. Springer, New York

Numerical Methods for Stochastic Differential Equations

Peter E. Kloeden
Professor
Goethe-Universität, Frankfurt, Germany

A stochastic differential equation (SDE)

$$dX_t = f(t, X_t)\, dt + g(t, X_t)\, dW_t$$

is, in fact, not a differential equation at all, but only a symbolic representation for the stochastic integral equation

$$X_t = X_{t_0} + \int_{t_0}^{t} f(s, X_s)\, ds + \int_{t_0}^{t} g(s, X_s)\, dW_s,$$

where the first integral is a deterministic Riemann integral for each sample path. The second integral is an Itô stochastic integral, which is defined as the mean-square limit of sums of products of the integrand g evaluated at the start of each discretization subinterval times the increment of the Wiener process W_t (which is often called a Brownian motion, see ►Brownian Motion and Diffusions). It is not possible to define this stochastic integral pathwise as a Riemann–Stieltjes integral, because the sample paths of a Wiener process, although continuous, are nowhere differentiable and not even of bounded variation on any bounded time interval.

The simplest numerical method for the above SDE is the *Euler-Maruyama scheme* given by

$$Y_{n+1} = Y_n + f(t_n, Y_n)\, \Delta_n + g(t_n, Y_n)\, \Delta W_n,$$

where $\Delta_n = t_{n+1} - t_n$ and $\Delta W_n = W_{t_{n+1}} - W_{t_n}$. This is intuitively consistent with the definition of the ►Itô integral. Here Y_n is random variable, which is supposed to be an approximation on X_{t_n}. The stochastic increments

ΔW_n, which are $\mathcal{N}\left(0, \sqrt{\Delta_n}\right)$ distributed, can be generated using, for example, the Box-Muller method. In practice, however, only individual realizations can be computed.

Depending on whether the realizations of the solutions or only their probability distributions are required to be close, one distinguishes between strong and weak convergence of numerical schemes, respectively, on a given interval $[t_0, T]$. Let $\Delta = \max_n \Delta_n$ be the maximum step size. Then a numerical scheme is said to converge with *strong order* γ if, for sufficiently small Δ,

$$\mathbb{E}\left(\left|X_T - Y_{N_T}^{(\Delta)}\right|\right) \le K_T\, \Delta^{\gamma}$$

and with *weak order* β if

$$\left|\mathbb{E}\left(p(X_T)\right) - \mathbb{E}\left(p\left(Y_{N_T}^{(\Delta)}\right)\right)\right| \le K_{p,T}\, \Delta^{\beta}$$

for each polynomial p. These are global discretization errors, and the largest possible values of γ and β give the corresponding strong and weak orders, respectively, of the scheme for a whole class of stochastic differential equations, e.g., with sufficiently often continuously differentiable coefficient functions. For example, the Euler-Maruyama scheme has strong order $\gamma = \frac{1}{2}$ and weak order $\beta = 1$, while the *Milstein scheme*

$$\begin{aligned} Y_{n+1} = {} & Y_n + f(t_n, Y_n)\,\Delta_n + g(t_n, Y_n)\,\Delta W_n \\ & + \frac{1}{2}\, g(t_n, Y_n)\frac{\partial g}{\partial x}(t_n, Y_n)\left\{(\Delta W_n)^2 - \Delta_n\right\} \end{aligned}$$

has strong order $\gamma = 1$ and weak order $\beta = 1$. Note that these convergence orders may be better for specific SDE within the given class, e.g., the Euler-Maruyama scheme has strong order $\gamma = 1$ for SDE with additive noise, i.e., for which g does not depend on x, since it then coincides with the Milstein scheme.

The Milstein scheme is derived by expanding the integrand of the stochastic integral with the Itô formula, the stochastic chain rule. The additional term involves the double stochastic integral $\int_{t_n}^{t_{n+1}} \int_{t_n}^{s} dW_u\, dW_s$, which provides more information about the non-smooth Wiener process inside the discretization subinterval and is equal to $\frac{1}{2}\left\{(\Delta W_n)^2 - \Delta_n\right\}$. Numerical schemes of even higher order can be obtained in a similar way. In general, different schemes are used for strong and weak convergence. The strong stochastic Taylor schemes have strong order $\gamma = \frac{1}{2}$, 1, $\frac{3}{2}$, 2, ..., whereas weak stochastic Taylor schemes have weak order $\beta = 1{,}2, 3, \ldots$. See Kloeden and Platen (1992) for more details. In particular, one should not use heuristic adaptations of numerical schemes for ordinary differential equations such as Runge–Kutta schemes, since these may not converge to the right solution.

The proofs of convergence rates in the literature assume that the coefficient functions in the above stochastic Taylor schemes are uniformly bounded, i.e., the partial derivatives of appropriately high order of the SDE coefficient functions f and g exist and are uniformly bounded. This assumption, however, is not satisfied in many basic and important applications, for example with polynomial coefficients such as

$$dX_t = -(1 + X_t)\left(1 - X_t^2\right)\, dt + \left(1 - X_t^2\right)\, dW_t,$$

or with square-root coefficients such as in the Cox-Ingersoll-Ross volatility model

$$dV_t = \kappa\left(\vartheta - V_t\right)\, dt + \mu\sqrt{V_t}\, dW_t,$$

which requires $V_t \ge 0$. The second is more difficult because there is a small probability that numerical iterations may become negative and various ad hoc methods have been suggested to prevent this. The paper (Jentzen et al. 2009) provides a systematic method to handle both of these problems by using pathwise convergence, i.e.,

$$\sup_{n=0,\ldots,N_T} \left|X_{t_n}(\omega) - Y_n^{(\Delta)}(\omega)\right| \longrightarrow 0 \text{ as } \Delta \to 0, \qquad \omega \in \Omega.$$

It is quite natural to consider pathwise convergence since numerical calculations are actually carried out path by path. Moreover, the solutions of some SDE do not have bounded moments.

Vector valued SDE with vector valued Wiener processes can be handled similarly. The main new difficulty is how to simulate the multiple stochastic integrals since these cannot be written as simple formulas of the basic increments as in the double integral above when they involve different Wiener processes.

About the Author

For biography *see* the entry ►Stochastic Differential Equations.

Cross References

►Brownian Motion and Diffusions
►Itô Integral
►Stochastic Difference Equations and Applications
►Stochastic Differential Equations

References and Further Reading

Jentzen A, Kloeden PE, Neuenkirch A (2009) Convergence of numerical approximations of stochastic differential equations on domains: higher order convergence rates without global Lipschitz coefficients. Numerische Mathematik, 112(1):41–64

Kloeden PE, Platen E (1992) The numerical solution of stochastic differential equations. Springer, Berlin (3rd revised printing 1999)

O

Omnibus Test for Departures from Normality

Kimiko O. Bowman[1], L. R. Shenton[2]
[1]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]Professor Emeritus of Statistics
University of Georgia, Athens, GA, USA

An omnibus *test* for departures from normality is an idea developed by E.S. Pearson (letter to Bowman); he thought a test including skewness b_1 and kurtosis b_2, both of which scale and location free, would give more information than the test using lower moments only. For the normal distribution, the population skewness is $\sqrt{\beta_1} = 0$ and the population kurtosis is $\beta_2 = 3$.

D'Agostino and Pearson (1973) introduced a goodness-of-fit test for departures from normality using sample skewness $\sqrt{b_1}$ and sample kurtosis b_2, thus

$$K_S^2 = X_S^2\left(\sqrt{b_1}\right) + X_S^2(b_2)$$

where $\sqrt{b_1} = m_3/m_2^{3/2}$, $b_2 = m_4/m_2^2$, and $m_s = \sum_1^n (x_j - \bar{x})^s/n, j = 1, 2, \cdots, n$, n sample size. D'Agostino and Pearson considered the Johnson's (1965) S_U and S_B transformed distribution for $\sqrt{b_1}$ and b_2. Johnson's system of distributions has the advantage that it transforms the distribution to the normal distribution, so the K_S^2 is considered as χ^2 with degree of freedom $v = 2$. Bowman and Shenton (1975b) continued the study further and introduced the contours for this test. They found S_U gives a good fit to $\sqrt{b_1}$ for $n \geq 8$ or so, and reasonable fit to b_2 for $n \geq 25$. For smaller sample size, the S_B system was used for the b_2.

$$X_S\left(\sqrt{b_1}\right) = \delta_1 \sinh^{-1}\left(\sqrt{b_1}/\lambda_1\right),$$
$$X_S(b_2) = \gamma_2 + \delta_2 \sinh^{-1}[(b_2 - \zeta)/\lambda_2],$$

and small sample $n < 25$ for b_2

$$X_S(b_2) = \gamma_2 + \delta_2 \ln\left(\frac{b_2 - \zeta}{\lambda_2 - b_2}\right).$$

γ, δ, λ and ζ are parameters for the Johnson system of distributions and derived by sample moments m_1', m_2, m_3 and m_4. The 90%, 95%, and 99% contours are made for $n = 20(5)65, 75, 85, 100, 120, 150, 200, 250, 800, 500, 1000$. The contours are shown in Fig. 1.

Extensive simulation study found the areas of the contours will contain the percentiles contents satisfactory; however they are not the smallest areas that represent the true shape of distribution. For the normal distribution, $\sqrt{b_1}$ and b_2 are not independent only asymptotically independent.

Pearson et al. (1977) studied the power of the test of normality for a variety of tests and found the omnibus test to be one of the most powerful tests for the departures from normality.

Bowman and Shenton (1986) further studied the subject and improved the K_S^2 test to include the correlation of $\left(\sqrt{b_1}, b_2\right)$. The new improved test is

$$K_{SR}^2 = \frac{X_S^2\left(\sqrt{b_1}\right) - 2RX_S\left(\sqrt{b_1}\right)X_S(b_2) + X_S^2(b_2)}{1 - R^2}$$

where $R = \rho\left(X_S\left(\sqrt{b_1}\right), X_S(b_2)\right)$, and is treated as χ^2 $(v = 2)$.

The new bivariate model contours follow the true distribution shapes and are shown in Fig. 2.

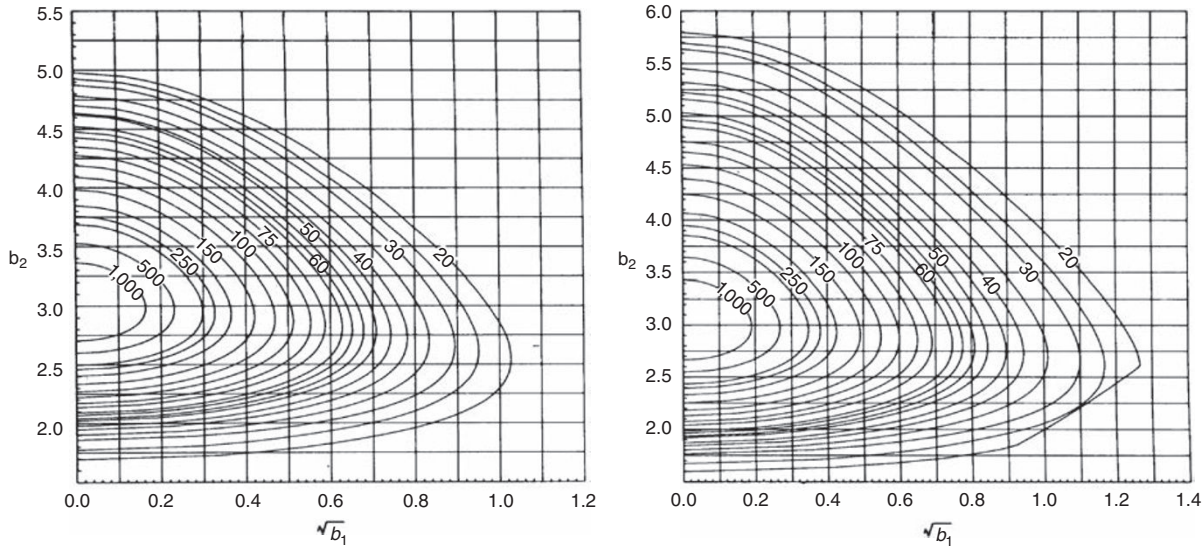
Figure 3 shows the contrast of two sets of contours old and new.

If the 99% level is chosen, then reject samples 2, 11, 12, 14, 15, and 17 immediately without further complicated calculations. This brings out the striking simplicity of the omnibus test approach.

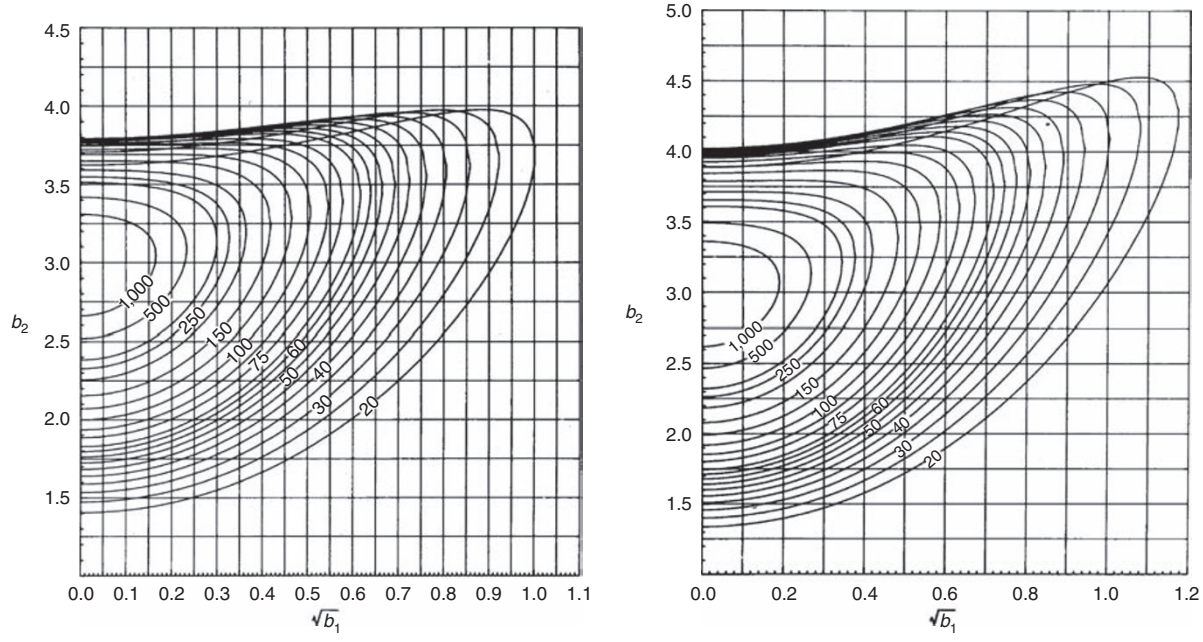
In concluding remarks, a contour can be constructed for any distribution which has moments. Bowman and Shenton (1975a) tabulated the values of moments of the skewness and kurtosis statistics in non-normal sampling, power series in power of (n^{-1}). Figure 4 shows several Pearson type I distributions including normal distribution for 90% with sample size $n = 200$. It is surprising to find 90% area of normal distribution is not necessarily smallest. Also the shape of contours are different for each distribution. The test could be used for non-normal distribution as well as computing the power of test for a particular distribution.

Patrinos and Bowman (1980) used the test to determine the thermal effect of rainfall amount around the nuclear plant.

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,



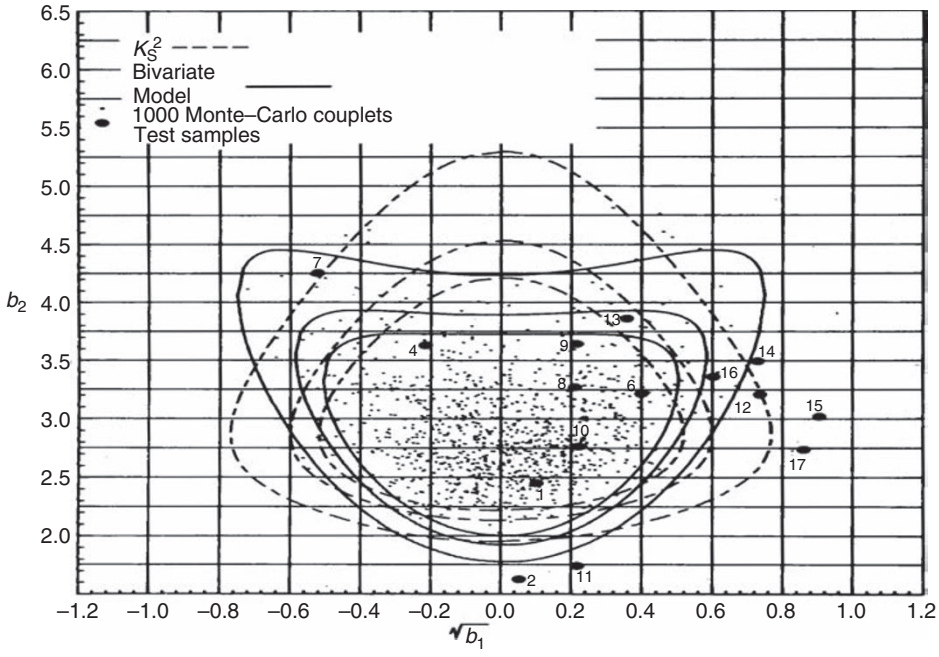
Omnibus Test for Departures from Normality. Fig. 1 Contours for K_S^2 test; $n = 20, 35, \cdots, 1,000$; 90% and 95%



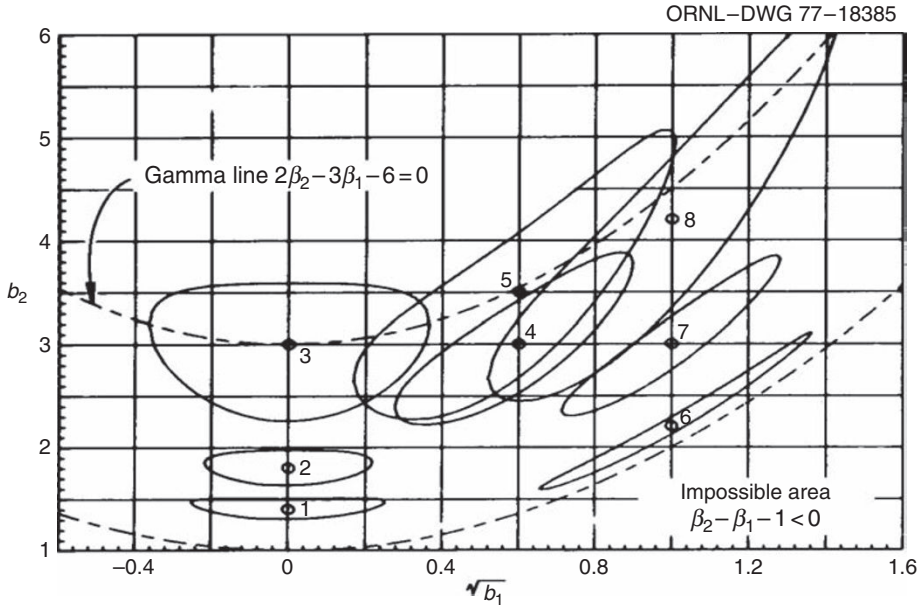
Omnibus Test for Departures from Normality. Fig. 2 Normal sampling, bivariate contours, 90 and 95% level, $n = 20, 35, \cdots, 1,000$

About the Authors

Dr. K.O. Bowman is a retired Senior Research Scientist, Oak Ridge National Laboratory, USA. She has received BS in Mathematics with Honor from Radford College, MS and Ph.D. in Statistics from Virginia Polytechnic Institute and State University and Doctor of Engineering degree in Mathematical Engineering from University of Tokyo. She is an Elected member of the International Statistical Institute, an Elected Fellow of the American Association for the Advancement of Science, an Elected fellow of the American Statistical Association, an Elected Fellow of the Institute of Mathematical Statistics. She has published many papers including three books: *Maximum Likelihood Estimation in Small Samples* (Charles Griffin 1977), *Properties*



Omnibus Test for Departures from Normality. Fig. 3 Normal sampling, contours of 90, 95, and 99% content, n=100, with random sample of 17 sets of 100



Omnibus Test for Departures from Normality. Fig. 4 Examples of different shapes of bivariate contours of 90% content for several Type I populations (n = 200); numbers 1–8 (small circles) indicate $\left(\sqrt{\beta_1}, \beta_2\right)$ of population samples

of Estimators for the Gamma Distribution (Marcel Dekker Inc, 1988), *Continued Fractions in Statistical Applications* (Marcel Dekker 1989), all three books are coauthored with Dr. Shenton. In particular, the paper coauthored with Dr. Marvin Kastenbaum, "Tables for Determining the Statistical Significance of Mutation Frequencies," *Mut. Res.*, 9,

(1970), was designated as **Citation Classic** in 1989. It was cited as the 4th most cited paper in the history of Mutation Research. She has served as an Associated Editor to many journals and as a committee member or chair for many professional societies, and received many awards.

Dr. Leonard R. Shenton was born on February 4, 1909. He is an Emeritus Professor of Statistics, University of Georgia, USA. He was a Reader in Mathematics (1960) in the faculty of Technology, University of Manchester (now University of Manchester, Institute of Science and Technology). He was an Associate Editor of the *Journal of American Statistical Association* (1964–1966). He is an Elected member of the International Statistical Institute, an Elected Fellow of the American Association of the Advancement of Science, and the Elected Fellow of the American Statistical Association. He has published many papers including three books: *Maximum Likelihood Estimation in Small Samples* (Charles Griffin 1977), *Properties of Estimators for the Gamma Distribution* (Marcel Dekker 1988), *Continued Fractions in Statistical Applications* (Marcel Dekker 1989), all three books are coauthored with Dr. Bowman. His interest is in symbolic computing.

"He is 101 years old and working on mathematics every day" (Kimiko Bowman).

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Anderson-Darling Tests of Goodness-of-Fit
►Jarque-Bera Test
►Kurtosis: An Overview
►Normal Distribution, Univariate
►Normality Tests
►Normality Tests: Power Comparison
►Skewness

References and Further Reading

Bowman KO, Shenton LR (1975a) Tables of moments of the skewness and kurtosis statistics in non-normal sampling. Union Carbides Nuclear Division report UCCND-CSD-8

Bowman KO, Shenton LR (1975b) Omnibus test contours for departures from normality based on $\sqrt{b1}$ and b2. Biometrika 62: 243–250

Bowman KO, Shenton LR (1986) Moment ($\sqrt{b1}$; b2) techniques, Goodness-of-fit techniques, Chapter 7. Marcel Dekker, New York

D'Agostino RB, Pearson ES (1973) Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{b1}$. Biometrika 60(3):613–622

Johnson NL (1965) Tables to facilitate fitting SU frequency curves. Biometrika 52:547–548

Patrinos AAN, Bowman KO (1980) Weather modification from cooling tower: a test based on the distributional properties of rainfall. J Appl Meteorol 19(3):290–297

Pearson ES, D'Agostine RB, Bowman KO (1977) Test of departure from normality: comparison of powers. Biometrika 64: 231–246

Online Statistics Education

David Lane
Rice University, Houston, TX, USA

The Internet is a great resource for statistics education providing numerous online resources including textbooks, interactive simulations/demonstrations, practical applications of statistics, assessment tools, and data analysis facilities. With the continuing advances in web technologies and means for accessing the Internet, the number of resources and their use will very likely increase greatly in the coming years.

Textbooks

Textbooks on the Internet offer many advantages over their written counterparts. One advantage is the capability of having links to glossary items embedded in the text. Over 20 years ago Lachman (1989) showed that the availability of glossary items enhances the comprehension of online texts. Links to glossary items are incorporated in several online statistics books (e.g., Online Statistics Education: A Multimedia Course of Study, SurfStat, SticiGui, and Learner.org). Similarly, many online statistics books incorporate hyperlinks to related materials (e.g., Hyperstat, SurfStat, Online Statistics Education: A Multimedia Course of Study, StatSoft, and New View of Statistics). Hyperlinks are particularly valuable in technical topics such as statistics because they make it easy for students to find prerequisite material that they may need to review before proceeding. Hyperlinks can be especially useful for students who consult a book for information about a topic covered in one of the later chapters of the book.

A second advantage of online statistics textbooks is that they can provide interactive exercises. For example, Learner.org offers interactive hints as well as solutions to homework problems. SticiGui's exercises are graded interactively with correct answers and explanations shown after the student answers. Online Statistics Education: A Multimedia Course of Study also provides automatically-graded questions with explanations. In addition, some questions from this site involve randomly-generated data. For example, a student may be asked to compute an independent-groups *t* test. Data are generated randomly and presented to the student for analysis. After the student enters the

answer, the website calculates the correct answer and provides feedback. The student then has the option to do a similar problem with newly generated data.

Gregory Francis of Purdue University developed an innovative way to take advantage of the online aspect of an online textbook. Using a modified version of Online Statistics Education: A Multimedia Course of Study, Dr. Francis added the capability to monitor the time students spent on individual pages. Each page had an assigned time period and students were given credit for the page only if they had the page open for the required amount of time. Although there was no way to know if students were really reading the page during that time, the idea was that most would be reading the page since they had it open anyway. The class using this method performed remarkably better than previous classes who had used a traditional text book: the modal grade jumped from a C in the previous year's class that used a traditional text book to an A with the online textbook. There is no way to know, of course, how much of this difference was due to student differences, differences in the quality of the textbooks per se, or other advantages of online texts. In any case, this finding, if confirmed in a more controlled study, would indicate that online statistics texts have the potential to greatly improve student learning.

Most online statistics textbooks contain non-mathematical introductions to statistical analysis. However, more advanced methods such as multivariate statistics are covered by some (PsychStat, Electronic Statistics Textbook, VisualStatistics, StatNotes: Topics in Multivariate Analysis).

Interactive Simulations/Demonstrations

Many of the concepts in statistics are inherently abstract. Interactive simulations and demonstrations can help make these concepts more concrete and facilitate understanding. Accordingly, the well-respected Guidelines for Assessment and Instruction in Statistics Education (GAISE) report recommended that technology tools should be used to help students visualize concepts and develop an understanding of abstract ideas by simulations.

Technologies such as Java and Flash have made it possible for interactive simulations to be hosted on the web and run within a web browser. A large number of these educational resources based on these technologies have become available over the last 10–15 years. Among the sites offering large numbers of interactive simulations and demonstrations are "Statistics Online Computational Resource (SOCR)," the "Web Interface for Statistics Education (WISE)," the Rice Virtual Laboratory in Statistics (RVLS)," and "Computer-Assisted Statistics Textbooks (CAST). Collections of educational resources including interactive simulations and demonstrations can also be found at the websites of the "Consortium for the Advancement of Undergraduate Statistics Education (CauseWeb)" and "Multimedia Educational Resource for Learning and Online Teaching (MERLOT)."

Real World Applications

One of the key recommendations of the GAISE report was that real data be used to illustrate statistical principles. This recommendation is in accordance with Weinberg and Abramowitz (2000) who concluded that statistical principles "come alive" through the use of real data.

There are many online resources for statistics education that include real data. Among them are the "Data and Story Library (DASL)," "Australasian Data and Story Library (OzDASL)," "Journal of Statistics Education Data Archive," "Stat Labs: Mathematical Statistics through Applications," "UCLA Statistics Case Studies," and "Online Statistics Education: A Multimedia Course of Study." These datasets can play an integral part of an online statistics course.

Chance News is an outstanding resource reviewing issues in the news that use probability or statistical concepts. Back issues of this monthly publication are available online starting with May 2005.

Assessment

Assessment is a critical part of instruction. The "Assessment Resource Tools for Improving Statistical Thinking (ARTIST)" project contains numerous resources for assessment including an item database, references to articles on assessment, and research instruments. The items in the database are categorized in terms of whether they measure statistical thinking, statistical reasoning, and/or basic statistical literacy. Other assessment items can be found in the "Database of Sample Statistics Quiz Questions" and on the "Chance Evaluation" page.

Data Analysis

The Internet contains a large number of resources for online statistical analysis. These resources are valuable for online statistics education because they enable students to do statistical analyzes without downloading additional software or learning complex procedures. Most of the resources are free although a few have small costs associated with their use.

StatCrunch can perform a wide range of statistical analyzes and create several types of statistical graphs. Statistical analyzes include ANOVA, simple, multiple and ►logistic regression, and ►control charts. StatCrunch also has calculators for several statistical distributions.

The site "Web Pages that Perform Statistical Calculations" contains an extensive list of online statistics analysis calculators. Topics include (1) Selecting the right kind of analysis, (2) Calculators, plotters, function integrators, and interactive programming environments, (3) Probability distribution functions: tables, graphs, random number generators, (4) Descriptive statistics, histograms, charts, (5) Confidence intervals, single-population tests, (6) Sample comparisons: t-tests, ANOVAs, non-parametric comparisons, (7) Contingency tables, cross-tabs, ►Chi-Square tests, (8) Regression, correlation, least squares curve-fitting, non-parametric correlation, (9) Analysis of ►survival data, (10) Bayesian Methods, and (11) Power, sample size and experimental design.

About the Author

Dr. David Lane is an Associate Professor of Statistics, Psychology, and Management, Rice University. He has authored and co-authored more than 90 articles. He is an Associate Editor of *Numeracy* and has reviewed articles for over 25 scientific journals. Dr. Lane received the Merlot Classics Award in Statistics in 2007 for his sampling distribution simulation (http://onlinestatbook.com/stat_sim/sampling_dist/index.html).

Cross References

►Careers in Statistics
►Data Analysis
►Learning Statistics in a Foreign Language
►Promoting, Fostering and Development of Statistics in Developing Countries
►Rise of Statistics in the Twenty First Century
►Role of Statistics in Advancing Quantitative Education
►Statistical Literacy, Reasoning, and Thinking
►Statistics Education

References and Further Reading

Assessment Resource Tools for Improving Statistical Thinking. https://app.gen.umn.edu/artist/index.html/

Australasian Data and Story Library (OzDASL) Journal of Statistics Education Data Archive. http://www.statsci.org/data/index.html

Chance Evaluation Page. http://www.dartmouth.edu/%7Echance/course/eval.html/

Chance News. http://chance.dartmouth.edu/chancewiki/

Computer-Assisted Statistics Textbooks. http://cast.massey.ac.nz/collection_public.html/

Consortium for the Advancement of Undergraduate Statistics Education. http://www.causeweb.org/

Data and Story Library. http://lib.stat.cmu.edu/DASL/

Database of Sample Statistics Quiz Questions. http://www2.gsu.edu/~dscbms/ibs/ibsd.html/

Electronic Statistics Textbook. http://statsoft.com/textbook/stathome.html

GAISE: Guidelines for Assessment and Instruction in Statistics Education College Report (2005) American Statistical Association. Retrieved December 2, 2008 from http://www.amstat.org/Education/gaise/GAISECollege.htm

HyperStat. http://davidmlane.com/hyperstat/index.html/

Journal of Statistics Education Data Archive. http://www.amstat.org/publications/JSE/jse_data_archive.html/

Lachman R (1989) Comprehension aids for on-line reading of expository text. Hum Factors 31(1):1–15

Learner.org. http://www.learner.org/courses/learningmath/data/index.html/

Multimedia Educational Resource for Learning and Online Teaching. http://www.merlot.org/merlot/

New View of Statistics. http://www.sportsci.org/resource/stats/index.html/

Online Statistics Education: A Multimedia Course of Study. http://onlinestatbook.com/

PsychStat. http://www.psychstat.missouristate.edu/multibook/mlt00.htm/

Rice Virtual Laboratory in Statistics. http://onlinestatbook.com/rvls.html

Stat Labs: Mathematical Statistics Through Applications. http://www.stat.berkeley.edu/users/statlabs/index.html/

StatCrunch. http://www.statcrunch.com/

StatNotes: Topics in Multivariate Analysis. http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm

SticiGui. http://www.stat.Berkeley.EDU/users/stark/SticiGui/index.htm/

Statistics Online Computational Resource. http://socr.ucla.edu/SOCR.html/

UCLA Statistics Case Studies. http://www.stat.ucla.edu/cases/

Visual Statistics. http://www.visualstatistics.net/

Web Interface for Statistics Education. http://wise.cgu.edu/tutor.asp/

Web Pages that Perform Statistical Calculations. http://statpages.org/

Weinberg SL, Abramowitz SK (2000) Making general principles come alive in the classroom using an active case studies approach. J Stat Educ 8(2). http://www.amstat.org/publications/jse/secure/v8n2/weinberg.cfm/

Optimal Designs for Estimating Slopes

Shahariar Huda
Professor
Kuwait University, Safat, Kuwait

Introduction

In classical designs, comparison of the treatment effects is of primary interest. Response surface designs are concerned with experiments in which treatments are combinations of various levels of factors that are quantitative. Consequently, the response is assumed to be a smooth function of the factors and the experimenter is usually interested in estimating the absolute response at various

points in the factor space. However, even in response surface designs, sometimes the experimenter may have greater interest in estimating the differences between response at various points rather than the response at individual locations (Herzberg 1967; Box and Draper 1980; Huda 1985). If differences in response at points close together in the factor space are involved, estimation of local slopes of the response surface becomes important. Optimal designs for estimating slopes are concerned with developing various meaningful optimality criteria and deriving designs that are best according to these criteria when the experimenter's primary objective is to estimate slopes. Such designs can be useful to investigators wishing to optimize (minimize or maximize) the response in a given region of the factor space.

Preliminaries

Typical response surface design set-up involves an univariate response y and k quantitative factors $x_1, \ldots, x_k$. It is assumed that $y = \phi(\mathbf{x}, \boldsymbol{\theta})$, a smooth function, where $\mathbf{x} = (x_1, \ldots, x_k)'$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ is a p-component column vector of unknown parameters. A design ξ is a probability measure on the experimental region χ. Let y_i be the observation on the response at point $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ $(i = 1, \ldots, N)$ selected according to the design. It is assumed that $y_i = \phi(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i$ where the ε_i's are uncorrelated, $E(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2$ $(i = 1, \ldots, N)$.

The parameters θ_j's are usually estimated by the method of ▶least squares. Let $\hat{\boldsymbol{\theta}}$ be the estimate of $\boldsymbol{\theta}$, then $\hat{y}(\mathbf{x}) = \phi(\mathbf{x}, \hat{\boldsymbol{\theta}})$ is the corresponding estimate of the response at the point $\mathbf{x}$. The column vector of estimated slopes along the factor axes at point $\mathbf{x}$ is given by $d\hat{y}(\mathbf{x})/d\mathbf{x} = (\partial\hat{y}(\mathbf{x})/\partial x_1, \ldots, \partial\hat{y}(\mathbf{x})/\partial x_k)'$. Let $\mathbf{V}(\xi, \mathbf{x}) = (N/\sigma^2)Cov(d\hat{y}/d\mathbf{x})$, the normalized covariance matrix of estimated slopes. $\mathbf{V}(\xi, \mathbf{x})$ depends on the point at which slopes are estimated as well as the design used. The vector $dy(\mathbf{x})/d\mathbf{x}$ not only displays the rates of change in y along the axial directions but also provides information about the rates of change in other directions. The directional derivative at point $\mathbf{x}$ in the direction specified by the vector of direction cosines $\mathbf{c} = (c_1, \ldots, c_k)'$ is $\mathbf{c}'\, dy(\mathbf{x})/d\mathbf{x}$. Also, the direction in which the derivative is largest is given by $\{(dy(\mathbf{x})/d\mathbf{x})'(dy(\mathbf{x})/d\mathbf{x})\}^{-1/2} dy(\mathbf{x})/d\mathbf{x}$.

Linear Model Set-Up

Most of the available work on response surface designs is concerned with situations where the model is linear in the parameters, that is, $\phi(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{f}'(\mathbf{x})\boldsymbol{\theta}$ with $\mathbf{f}'(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_p(\mathbf{x}))$ containing p linearly independent functions of $\mathbf{x}$. In this case for the least squares estimate $\hat{\boldsymbol{\theta}}$,

$$(N/\sigma^2)Cov(\hat{\boldsymbol{\theta}}) = \mathbf{M}^{-1}(\xi),$$

assuming the information matrix $\mathbf{M}(\xi) = \int_{\chi} \mathbf{f}(\mathbf{x})\mathbf{f}'(\mathbf{x})\, \xi(d\mathbf{x})$ to be nonsingular. Then

$$(N/\sigma^2)Cov(d\hat{y}(\mathbf{x})/d\mathbf{x}) = \mathbf{H}(\mathbf{x})\mathbf{M}^{-1}(\xi)\mathbf{H}'(\mathbf{x}) = \mathbf{V}(\xi, \mathbf{x}),$$

where $\mathbf{H}(\mathbf{x})$ is a $k \times p$ matrix whose ith row is $\partial\mathbf{f}'(\mathbf{x})/\partial x_i = (\partial f_1(\mathbf{x})/\partial x_i, \ldots, \partial f_p(\mathbf{x})/\partial x_i)$.

The commonly used linear models are the polynomial models for which $\mathbf{f}(\mathbf{x})$ consists of terms of a polynomial of order (degree) d in $\mathbf{x}$. If all the terms of a polynomial of degree d are included in the model then $\mathbf{f}(\mathbf{x})$ (and $\boldsymbol{\theta}$) contains ${}_{k+d}C_d$ components. A design ξ is called a dth order design if it permits estimation of all the parameters of a dth order model. A design ξ of order d is called symmetric if all the "odd moments" up to order $2d$ are zero, i.e., if

$$\int_{\chi} x_1^{d_1} \ldots x_k^{d_k}\, \xi(d\mathbf{x}) = 0$$

whenever one or more of the d_i's are odd integers and $\sum_{i=1}^{k} d_i \le 2d$. A design is balanced (permutation invariant) if the moments are invariant with respect to permutations of the factors $x_1, \ldots, x_k$. The class of "symmetric balanced" designs is simpler to analyze and also very rich in the sense that it contains optimal designs under many commonly used criteria.

Optimality Criteria

Research into design of experiments to estimate the slopes of a response surface was initiated by Atkinson (1970) who proposed minimization of trace of the "integrated mean squared error matrix" of the estimated slopes as a design criterion and investigated first-order designs when the true model may be a second-order model. Since then the problem of optimal design for estimating slopes has been investigated by many other researchers. Ott and Mendenhall (1972) considered univariate second-order model over an interval and used criterion of minimizing the variance of the estimated slope maximized over the design region. Murty and Studden (1972) also considered univariate polynomial model of order d over an interval and obtained optimal designs to minimize variance of the estimated slope at a fixed point in the interval as well as averaged over the interval. Myers and Lahoda (1975) considered integrated mean squared error criterion for first- and second-order designs in the presence of second- and third-order terms in the true model, respectively when the integration is done with respect to an uniform measure. Vaughan (1993) obtained a new optimal second-order design for

estimating slopes near a stationery region, taking account of both variance and bias in the estimation.

If the model assumed is correct and primary goal of the experimenter is to estimate slopes, it is natural to consider design criteria based on $\mathbf{V}(\xi,\eta)$ rather than $Cov(\hat{\theta})$ or $\mathrm{var}(\hat{y}(\mathbf{x}))$ where

$$\mathbf{V}(\xi,\eta) = \int_R (N/\sigma^2) Cov(d\hat{y}(\mathbf{x})/d\mathbf{x})\eta(d\mathbf{x}) = \int_R \mathbf{H}(\mathbf{x})\mathbf{M}^{-1}(\xi)\mathbf{H}'(\mathbf{x})\eta(d\mathbf{x}),$$

with R being the region of interest and η a measure reflecting the pattern of the experimenters interest.

Then in analogy with the traditional set-up A-, D- and E-average optimal designs for estimating slopes are defined as those minimizing

$$\beta = tr\mathbf{V}(\xi,\eta)/k = \sum_{i=1}^{k}\beta_i/k, \quad |\mathbf{V}(\xi,\eta)|^{1/k} = \prod_{i=1}^{k}\beta_i^{1/k}, \quad Max\{\beta_1,\ldots,\beta_k\}$$

respectively, where β_i are e-values of $\mathbf{V}(\xi,\eta)$.

One possibility is to take η as the measure putting all its mass at a single point $\mathbf{x}^*$ where $tr\mathbf{V}(\xi,\mathbf{x}), |\mathbf{V}(\xi,\mathbf{x})|$ and $Max\{\beta_1(\xi,\mathbf{x}),\ldots,\beta_k(\xi,\mathbf{x})\}$ is maximized, respectively, the $\beta_i(\xi,\mathbf{x})$ being e-values of $\mathbf{V}(\xi,\mathbf{x})$. These objective functions are not necessarily maximized at the same point and the measure η may be different in each case. This is the minimax approach and A-, D- and E-minimax optimality criteria of designs for estimating slopes are defined as minimization (with respect to ξ) of

$$\beta(\xi,\mathbf{x}) = \sum_{i=1}^{k}\beta_i(\xi,\mathbf{x})/k, \; |\mathbf{V}(\xi,\mathbf{x})|^{1/k} = \prod_{i=1}^{k}\{\beta_i(\xi,\mathbf{x})\}^{1/k}, \; Max\{\beta_1(\xi,\mathbf{x}),\ldots,\beta_k(\xi,\mathbf{x})\}$$

maximized with respect to $\mathbf{x} \in R$, respectively. Under "MV-minimax optimality" criterion the objective is to minimize the largest diagonal element of $\mathbf{V}(\xi,\mathbf{x})$ maximized with respect to $\mathbf{x} \in R$.

Available Results

The A-criterion is the easiest to handle and has been extensively used in the past. The A-minimax second- and third-order designs for regression over spheres were derived in Mukerjee and Huda (1985), some results for the set-up with $R \neq \chi$ being given in Huda (1990). For the cubic regions the A-minimax second-order designs were presented in Huda and Shafiq (1992) while the third-order designs were derived in Huda and Al-Shiha (1998).

The D-criterion is much more difficult to tackle. For second-order models over spherical regions the D-minimax designs were derived in Huda and Al-Shiha (1999) while the E-minimax designs were presented in Al-Shiha and Huda (2001). More recently, Huda and Al-Shingiti (2004) obtained A-, D- and E-minimax designs over spherical regions when $R \neq \chi$. Huda and Al-Shiha (2000) provided the D- and E-minimax second-order designs over cubic regions.

A-average optimal second- and third-order designs for spherical regions were studied in Huda (1986). The A-average second-order designs for cubic regions were studied in Huda (1998).

Concluding Remarks

Many interesting problems remain unsolved in optimal design of experiments for estimating the slopes. For example, the D- and E-minimax criteria have not yet been applied for third- and higher-order designs. Very little work in general has been done for third-order designs and practically none for higher order designs in more than one variable. The derivation of optimal designs for models of order three and higher is likely to be very difficult but certainly deserves the attention of researchers in the field. Researchers also need to consider situations that take account of the possibility of bias in the assumed models.

About the Author

Dr. Shahariar Huda is Professor of Statistics, Department of Statistics and O.R., Faculty of Science, Kuwait University. He is Editor-in-Chief for *International Journal of Statistical Sciences* (IJSS) (Bangladesh, 2007–2012). He is Associate editor for the following journals: *Journal of Applied Statistical Sciences* (USA), *Aligarh Journal of Statistics* (India), *Journal of Statistical Research* (Bangladesh), and *Journal of Statistical Sciences* (Bangladesh). He has received an award for the best paper on Experimental Design (during 1986–1987) in *Journal of Indian Society of Agricultural Statistics*. He is Elected Ordinary Member of International Statistical Institute (1993), Member of Royal Statistical Society, American Statistical Association and Institute of Mathematical Statistics. Professor Huda has written 26 papers in international journals on slope estimation. He is co-author of the book *Markov Models with Covariate Dependence for Repeated Measures* (with M.A. Islam and R.I. Chowdhury, Nova Science, USA, 2009).

Cross References

►Least Squares
►Optimal Regression Design
►Optimum Experimental Design
►Response Surface Methodology

References and Further Reading

Al-Shiha AA, Huda S (2001) On E-optimal designs for estimating slopes. J Appl Stat Sci 10:357–364

Atkinson AC (1970) The design of experiments to estimate the slope of a response surface. Biometrika 57:319–328

Box GEP, Draper NR (1980) The variance functions of the difference between two estimated responses. J R Stat Soc B 42: 79–82

Herzberg AM (1967) The behaviour of the variance functions of the difference between two estimated responses. J R Stat Soc B 29:174–179

Huda S (1985) Variance of the difference between two estimated responses. J Stat Plann Inf 11:89–93

Huda S (1986) On minimizing the average variance of the estimated slope of a response surface. Pak J Stat 2:31–37

Huda S (1990) Further results on minimax designs to estimate the slope of a response surface. Biom J 32:189–194

Huda S (1998) On optimal designs to estimate the slope of a second-order response surface over cubic regions. Parisankhyan Samikkha 5:11–19

Huda S, Al-Shiha AA (1998) Minimax designs for estimating the slope of a thirdorder response surface in a hypercubic region. Commun Stat Sim 27:345–356

Huda S, Al-Shiha AA (1999) On D-optimal designs for estimating slopes. Sankhya B 61:488–495

Huda S, Al-Shiha AA (2000) On D- and E-minimax optimal designs for estimating the axial slopes of a second-order response surface over hypercubic regions. Commun Stat Theor Meth 29:1827–1849

Huda S, Al-Shingiti AM (2004) On second-order A-, D- and E-minimax designs for estimating slopes in extrapolation and restricted interpolation regions. Commun Stat Sim Comp 33:773–785

Huda S, Shafiq M (1992) Minimax designs for estimating the slope of a second-order response surface in a cubic region. J Appl Stat 19:501–507

Mukerjee R, Huda S (1985) Minimax second- and third-order designs to estimate the slope of a response surface. Biometrika 72:173–178

Murty VN, Studden J (1972) Optimal designs for estimating the slopes of a polynomial regression. J Am Stat Assoc 67: 869–873

Myers RH, Lahoda SJ (1975) A generalization of the response surface mean square error criterion with a specific application to the scope. Technometrics 17:481–486

Ott L, Mendenhall W (1972) Designs for estimating the slope of a second order linear model. Technometrics 14: 341–353

Vaughan TS (1993) Experimental design for response surface gradient estimation. Commun Stat Theor Meth 22:1535–1555

Optimal Regression Design

Saumen Mandal
Associate Professor
University of Manitoba, Winnipeg, MB, Canada

Introduction

There are a variety of problems in statistics, which demand the calculation of one or more probability distributions or measures. Optimal regression design is a particular example. Other examples include parameter estimation, adaptive design and stratified sampling.

Consider the problem of selecting an experimental design to furnish information on models of the type $y \sim \pi(y|\underline{x}, \underline{\theta}, \sigma)$, where y is the response variable; $\underline{x} = (x_1, x_2, \ldots, x_m)^T$ are design variables, $\underline{x} \in \mathcal{X} \subseteq \mathbb{R}^m$, $\mathcal{X}$ is the design space; $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$ are unknown parameters; σ is a nuisance parameter, fixed but unknown; and $\pi(.)$ is a probability model. In most applications, $\mathcal{X}$ is taken to be compact. For each $\underline{x} \in \mathcal{X}$, an experiment can be performed whose outcome is a random variable $y(\underline{x})$, where $var(y(\underline{x})) = \sigma^2$. In linear models, it is further assumed that $y(\underline{x})$ has an expected value of the explicit form $E(y|\underline{v}) = \underline{v}^T\underline{\theta}$, where $\underline{v} \in \mathcal{V}$, $\mathcal{V} = \{\underline{v} \in \mathbb{R}^k : \underline{v} = \underline{\eta}(\underline{x}), \underline{x} \in \mathcal{X}\}$ with $\underline{\eta}(\underline{x}) = (\eta_1(\underline{x}), \eta_2(\underline{x}), \ldots, \eta_k(\underline{x}))^T$. That is, $\mathcal{V}$ is the image under a set of regression functions η of $\mathcal{X}$, called the induced design space.

Clearly choosing a vector $\underline{x}$ in the design space $\mathcal{X}$ is equivalent to choosing a k-vector $\underline{v}$ in the closed bounded k-dimensional space $\mathcal{V}$. Typically this design space is continuous but we can assume that $\mathcal{V}$ is discrete. Suppose that $\mathcal{V}$ consists of J distinct vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_J$. In order to obtain an observation on y, a value for $\underline{v}$ must first be chosen from the J elements of $\mathcal{V}$ to be the point at which to take this observation.

Exact and Approximate Designs

A natural question to consider is at what values of $\underline{v}$ should observations, say n, on y be taken in order to obtain a "best" inference or as reliable an inference as possible for all or some of the parameters $\underline{\theta}$. Such a "best" selection of $\underline{v}$ (or $\underline{x}$ values) or allocation of the n observations to the elements of $\mathcal{V}$ (or $\mathcal{X}$) is termed an optimal regression design. Given n observations, we must decide how many of these, say n_j, to take at $\underline{v}_j$, $\sum n_j = n$. Given that the n_j's must be integer this is an integer programming problem and in the optimal design context is described as an "exact design" problem. Typically interger programming

problems are difficult or at least laborious to solve, mainly because the theory of calculus cannot be used to define the existence of or to identify optimal solutions.

However, we could find a simpler or more flexible problem to solve and yet is not much visibly different from the original problem. If $\underline{\hat{\theta}}$ is the least squares estimator of $\underline{\theta}$, then $cov(\underline{\hat{\theta}}) \propto M^{-1}(p)$, where $M(p)$ is the per observation information matrix:

$$M(p) = \sum_{j=1}^{J} p_j \underline{v}_j \underline{v}_j^T = VPV^T = \sum_{j=1}^{J} p_j \underline{\eta}(\underline{x}_j)\underline{\eta}^T(\underline{x}_j)$$

where V is the $k \times J$ matrix $[\underline{v}_1\, \underline{v}_2 \ldots \underline{v}_J]$, $P = \text{diag}\,(p_1, p_2, \ldots, p_J)$, $p_j = n_j/n$. So p_j is the proportion of observations taken at $\underline{v}_j$, so that $p_j \geq 0$, $\sum p_j = 1$; and $p = (p_1, p_2, \ldots, p_J)$ represents the resultant distribution on $\mathcal{V}$. Thus our problem becomes that of choosing p to make $M(p)$ large subject to $p_j = n_j/n$. Relaxing the latter to $p_j \geq 0$ and $\sum p_j = 1$ yields an "approximate design" problem. Naturally an approximate solution that would be preferred to the original exact design problem would be np, rounded to a nearest exact design.

We wish to choose the proportion p_j of observations, taken at $\underline{x}_j$ (or $\underline{v}_j$) to ensure good estimation of $\underline{\theta}$ by optimizing some criterion. The most important criterion in design applications is that of D-optimality, in which the criterion $\phi(p) = \psi\{M(p)\} = log\{det(M(p))\}$, is maximized. A D-optimal design minimizes the volume of the conventional ellipsoidal confidence region for the parameters of the linear model. Other choices of maximizing criteria are $\psi\{M(p)\} = -\underline{c}^T M^{-1}(p)\underline{c}$ for a given vector $\underline{c}$ (c-optimality; appropriate if there is interest only in $\underline{c}^T\underline{\theta}$) or $\psi\{M(p)\} = -Trace(AM^{-1}(p)A^T)$ and $\psi\{M(p)\} = -logdet(AM^{-1}(p)A^T)$ for a given $s \times k$ matrix A, $s < k$ (linear optimality and D_A-optimality respectively and appropriate if there is interest in inference only for $A\underline{\theta}$). A-optimality is the special case of linear optimality with $A = I$, the identity matrix. There is a vast statistical literature on optimal design and optimality criteria. Useful texts in optimal design are Fedorov (1972), Atkinson and Donev (1992), Pukelsheim (1993), and Silvey (1980).

Optimality Conditions and Algorithms

Our general problem is to maximize a criterion $\phi(p)$ subject to $p_j \geq 0, j = 1, 2, \ldots, J$ and $\sum p_j = 1$. To solve the problem, we first define optimality conditions in terms of point to point directional derivatives. Making use of differential calculus, we exploit the directional derivative of Whittle (1973). The directional derivative $F_\phi\{p, q\}$ of a criterion function $\phi(.)$ at p in the direction of q is defined as

$$F_\phi\{p, q\} = \lim_{\varepsilon \downarrow 0} \frac{\phi\{(1-\varepsilon)p + \varepsilon q\} - \phi(p)}{\varepsilon}.$$

The derivative $F_\phi\{p, q\}$ exists even if $\phi(.)$ is not differentiable. If $\phi(.)$ is differentiable, $F_\phi(p, q) = (q - p)^T \partial\phi/\partial p$. Let

$$F_j = F_\phi(p, e_j) = \frac{\partial \phi}{\partial p_j} - \sum_{i=1}^{J} p_i \frac{\partial \phi}{\partial p_i} = d_j - \bar{d},$$

$$d_j = \frac{\partial \phi}{\partial p_j},\ \bar{d} = \sum_{i=1}^{J} p_i d_i$$

where e_j is the jth unit vector in $\mathbb{R}^J$. We call F_j the vertex directional derivative of $\phi(.)$ at p. If $\phi(.)$ is differentiable at an optimizing distribution p^*, then the first-order conditions for $\phi(p^*)$ to be a local maximum of $\phi(.)$ in the feasible region of the problem are

$$F_j^* = F_\phi\{p^*, e_j\} \begin{cases} = 0 & \text{for } p_j^* > 0 \\ \leq 0 & \text{for } p_j^* = 0. \end{cases}$$

If $\phi(.)$ is concave on its feasible region, then the above first-order stationarity conditions are both necessary and sufficient for optimality, a result known as the general equivalence theorem in optimal design (Kiefer 1974).

In order to determine the optimal weights, we often require an algorithm because it is typically not possible to evaluate an optimal solution explicitly. Several algorithms exist in the literature. A class of algorithms which neatly satisfy the basic constraints of the optimal weights take the form $p_j^{(r+1)} \propto p_j^{(r)} f\left(d_j^{(r)}\right)$, where $d_j^{(r)} = \partial\phi/\partial p_j$ at rth iterate $p = p^{(r)}$ and the function $f(.)$ satisfies certain conditions and may depend on a free positive parameter δ. Torsney (1977) first proposed this type of iteration by taking $f(d) = d^\delta$ with $\delta > 0$. Subsequent empirical studies include Silvey et al. (1978) and Torsney (1988). Torsney (1983) explores monotonicity of particular values of δ for particular $\phi(p)$. Mandal and Torsney (2006) modified the algorithm for more than one optimizing distributions based on a clustering approach. Mandal et al. (2005) used the algorithm for constructing designs subject to additional constraints.

There are many other algorithms in the literature. Vertex direction algorithms which perturb one p_j and change the others proportionately were first proposed by Fedorov (1972) and Wynn (1972). When all p_j are positive at the optimum or when it has been established which are positive, constrained steepest ascent or Newton type iterations may be appropriate. See Wu (1978) and Atwood (1980) on these respectively. Molchanov and Zuyev (2001)

consider steepest descent algorithms based on the gradient function.

Conclusion

Finally, as a concluding remark, one important advantage of optimal regression designs is that we can find the best selection of the inputs (factors) for which the optimal value of each response occurs. Another advantage of optimal design is that it reduces the costs of experimentation by allowing statistical models to be estimated with fewer experimental runs.

Acknowledgments

The author thanks the reviewer for helpful suggestions that led to some improvement over an earlier version of the manuscript. The research of the author is supported by a Discovery Grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Cross References

►Optimal Designs for Estimating Slopes
►Optimum Experimental Design
►Robust Statistics

References and Further Reading

Atkinson AC, Donev AN (1992) Optimum experimental designs. Oxford Statistical Science Series-8, Clarendon, Oxford
Atwood CL (1980) Convergent design sequences for sufficiently regular optimality criteria, II: singular case. Ann Stat 8:894–912
Fedorov VV (1972) Theory of optimal experiments. Academic, New York
Kiefer J (1974) General equivalence theory for optimum designs (approximate theory). Ann Stat 2:849–879
Mandal S, Torsney B (2006) Construction of optimal designs using a clustering approach. J Stat Plann Inf 136:1120–1134
Mandal S, Torsney B, Carriere KC (2005) Constructing optimal designs with constraints. J Stat Plann Inf 128:609–621
Molchanov I, Zuyev S (2001) Variational calculus in the space of measures and optimal design. Opt Design 2000, Kluwer Academic, Dordrecht, pp 79–90
Pukelsheim F (1993) Optimal design of experiments. Wiley Series in Probability and Mathematical Statistics, New York
Silvey SD (1980) Optimal design. Chapman and Hall, London
Silvey SD, Titterington DM, Torsney B (1978) An algorithm for optimal designs on a finite design space. Commun Stat A 14:1379–1389
Torsney B (1977) Contribution to discussion of "maximum likelihood estimation via the EM algorithm" by Dempster et al. J R Stat Soc B 39:26–27
Torsney B (1983) A moment inequality and monotonicity of an algorithm. In: Kortanek KO, Fiacco AV (eds) Proceedings of the International Symposium on Semi-Infinite Programming and Application at the University of Texas at Austin. Lecture Notes in Economics and Mathematical systems 215:249–260
Torsney B (1988) Computing optimizing distributions with applications in design, estimation and image processing. In: Dodge Y, Fedorov VV, Wynn HP (eds) Optimal design and analysis of experiments. Elsevier Science BV, North Holland, pp 361–370
Whittle P (1973) Some general points in the theory of optimal experimental design. J R Stat Soc B 35:123–130
Wu CFJ (1978) Some iterative procedures for generating nonsingular optimal designs. Commun Stat A 14:1399–1412
Wynn HP (1972) Results in the theory and construction of *D*-optimum experimental designs (with Discussion). J R Stat Soc B 34:133–147, 170–186

Optimal Shrinkage Estimation

S. Ejaz Ahmed[1], T. Quadir[2], S. Nkurunziza[2]
[1]Professor and Head Department of Mathematics and Statistics
University of Windsor, Windsor, ON, Canada
[2]University of Windsor, Windsor, ON, Canada

To fix the idea, we consider the estimation of means in the following one-way ANOVA model,

$$Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, , \ldots, n_i. \qquad (1)$$

For brevity sake, we consider that ϵ_{ij} are identically and independently normally distributed with mean 0 and common finite variance σ^2. The statistical objective is to estimate simultaneously the mean parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. Let $\bar{Y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$, $i = 1, 2, \ldots, k$. If σ is known, the vector $\left(\bar{Y}_1, \ldots, \bar{Y}_k\right)'$ is a complete sufficient statistic for $\boldsymbol{\theta}$. Further, it is the best unbiased, maximum likelihood, and minimax estimator of $\boldsymbol{\theta}$. However, we wish to improve the performance of the maximum likelihood estimator (MLE), $\bar{Y}_i$ by incorporating the information (which may not be certain) regarding the parameter vector of interest, $\boldsymbol{\theta}$. In other words, it is possible that $\boldsymbol{\theta} = \boldsymbol{\theta}^o$, where $\boldsymbol{\theta}^o$ is a known prior guess of $\boldsymbol{\theta}$. On the other hand, in many applications one considers that $\boldsymbol{\theta} = \theta 1_k$, where θ is the unknown common parameter of interest and 1_k is a k-column vector with all entries equal to 1. Indeed, it is not unusual to encounter the above information regarding the parameter of interest in many practical situations. Hence, in either situation we assume that a prior guess of $\boldsymbol{\theta}$, say $\tilde{\boldsymbol{\theta}}$, is available or can be evaluated from the data. For example, $\tilde{\boldsymbol{\theta}}$ is the restricted maximum likelihood estimator (RMLE) under the assumption that $\boldsymbol{\theta} = \theta 1_k$. In this case, $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}, \ldots, \tilde{\theta})' = \tilde{\theta} 1_k$, where $\tilde{\theta} = \bar{Y} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}/n$, where $n = n_1 + \ldots + n_k$. It is well-documented in the scientific literature that restricted estimator yields smaller risk (under quadratic loss) when a priori information is

nearly correct, however at the expense of poorer performance in the rest of the parameter space induced by the restriction. The incorrect or imprecise restrictions on $\boldsymbol{\theta}$ may lead to biased (or even inconsistent) and inefficient estimators of $\boldsymbol{\theta}$. Now, the question is how to combine $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ to get a better estimation strategy for $\boldsymbol{\theta}$. A natural way to balance the potential bias of the estimator under the restriction against the benchmark estimator is to take a weighted average of $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$. Such shrinkage estimator (SE) or integrated estimator is defined as

$$\hat{\boldsymbol{\theta}}^{S} = \pi\hat{\boldsymbol{\theta}} + (1-\pi)\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}} + \pi(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}), \tag{2}$$

for a judiciously chosen weight π $(0 \le \pi \le 1)$. Many of the estimators proposed in the reviewed literature, both design-based and model-based, have the integrated form (2). A major drawback of $\hat{\boldsymbol{\theta}}^{S}$ is that it is not uniformly better than either component estimators in terms of risk. Another approach is to employ shrinkage estimation based on Stein rule, which in turn yields the optimal weight for $\hat{\boldsymbol{\theta}}^{S}$. To construct a shrinkage estimator, we consider the preliminary test approach as advocated by Sclove (1968) or empirical Bayes consideration. This approach had been implemented by Ahmed (2001) and others. In this technique, the prior information regarding $\boldsymbol{\theta}$ can be displayed in the form of the null hypothesis. Let us consider the following null hypothesis:

$$H_o: \theta_1 = \cdots = \theta_k = \theta \text{ (unknown)}. \tag{3}$$

For the preliminary test on the null hypothesis in (3), we consider the following F-test statistic

$$D = \frac{(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})'\boldsymbol{\Lambda}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})}{(k-1)S^2}, \quad \boldsymbol{\Lambda} = \text{diagonal}(n_1, \ldots, n_k),$$
$$S^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}). \tag{4}$$

A preliminary test estimator (PTE) is defined as

$$\hat{\boldsymbol{\theta}}^{P} = \tilde{\boldsymbol{\theta}} + (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})I(D \ge d_\alpha), \tag{5}$$

where $I(A)$ is an indicator function of the set A and d_α is the upper $100\alpha\%$ $(0 < \alpha < 1)$ point of the test statistic under the null hypothesis. The PTE is essentially obtained by repacking π by a random quantity $I(D \ge d_\alpha)$; however, this is not an optimal value of π. The PTE does not uniformly improve upon $\hat{\boldsymbol{\theta}}$ which merits further enhancement. Hence we replace this indicator function by a smooth function of D to obtain James–Stein (J–S) type estimator (see ▶James-Stein Estimator).

Optimal Shrinkage Estimation Strategy

The J-S type shrinkage estimator of $\boldsymbol{\theta}$ is defined by

$$\hat{\boldsymbol{\theta}}^{JS} = \tilde{\boldsymbol{\theta}} + \{1 - cD^{-1}\}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}), \quad k > 2, \tag{6}$$

where c is the shrinkage constant chosen in an interval in such a way that $\hat{\boldsymbol{\theta}}^{JS}$ dominates $\hat{\boldsymbol{\theta}}$. We notice that although this estimator resembles the J-S rule, its construction is based on the preliminary test approach. Further, as it shrinks the $\hat{\boldsymbol{\theta}}$ toward $\tilde{\boldsymbol{\theta}}$, this estimator is generally called a shrinkage estimator. Clearly, if the value of D is small then a relatively large weight is placed on $\tilde{\boldsymbol{\theta}}$. Otherwise, more weight is placed on $\hat{\boldsymbol{\theta}}$. Consequently, $\hat{\boldsymbol{\theta}}^{JS}$ is a special case of $\hat{\boldsymbol{\theta}}^{S}$ with $\pi = (1 - cD^{-1})$. It is important to note that $\hat{\boldsymbol{\theta}}^{JS}$ may over-shrink $\hat{\boldsymbol{\theta}}$ toward the $\tilde{\boldsymbol{\theta}}$, thus causing a possible inversion of the sign of the benchmark estimator $\hat{\boldsymbol{\theta}}$. A positive-rule shrinkage estimator (PSE) $\hat{\boldsymbol{\theta}}^{JS+}$ is obtained from (6) by changing the factor $1-c/D$ to 0 whenever $D \le c$, that is,

$$\hat{\boldsymbol{\theta}}^{JS+} = \tilde{\boldsymbol{\theta}} + \left(1 - cD^{-1}\right)^{+}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}), \tag{7}$$

where $z^{+} = max(0, z)$. The PSE is particularly important to control the over-shrinking inherent in $\hat{\boldsymbol{\theta}}^{JS+}$. For this reason, Ahmed (2001) recommended using the shrinkage estimator as a tool to develop the PSE instead of as an estimator in its own right. Rewriting the above relation in the following weighted form

$$\hat{\boldsymbol{\theta}}^{JS+} = \tilde{\boldsymbol{\theta}} + \left(1 - cD^{-1}\right)[1 - I(D > c)](\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}), \tag{8}$$

we easily see that $\hat{\boldsymbol{\theta}}^{JS+} = \hat{\boldsymbol{\theta}}^{S}$ with $\pi = (1 - cD^{-1})I(cD^{-1} > 1)$ keeping in mind all the shrinkage estimators are biased estimators of $\boldsymbol{\theta}$. Since bias is a part of the risk, we shall only compare the risk function of the estimators. Assuming that $\hat{\boldsymbol{\theta}}^{*}$ is an estimator of $\boldsymbol{\theta}$ and Q is a positive semi-definite matrix, let us define the quadratic loss function

$$\mathcal{L}(\hat{\boldsymbol{\theta}}^{*}; \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}^{*} - \boldsymbol{\theta})'Q(\hat{\boldsymbol{\theta}}^{*} - \boldsymbol{\theta}) = trace[Q(\hat{\boldsymbol{\theta}}^{*} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}^{*} - \boldsymbol{\theta})']. \tag{9}$$

The risk of $\hat{\boldsymbol{\theta}}^{*}$ is $\mathcal{R}(\hat{\boldsymbol{\theta}}^{*}; \boldsymbol{\theta}) = \mathcal{E}[\mathcal{L}(\hat{\boldsymbol{\theta}}^{*}; \boldsymbol{\theta})] = trace(Q\boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma}$, the mean squared error matrix, is defined as $\boldsymbol{\Gamma} = \mathcal{E}([\hat{\boldsymbol{\theta}}^{*} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}^{*} - \boldsymbol{\theta})']$. Noting that $\mathcal{R}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = \sigma^2\boldsymbol{\Lambda}^{-1}$, so to appraise the relative risk performance in meaningful way it makes sense that we consider $Q = \sigma^{-2}\boldsymbol{\Lambda}$ what follows in the remaining discussions. For this particular choice of Q, we have $\mathcal{R}(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}) = k$. The risk of the shrinkage estimators are given below. These results originated from the work of

James and Stein (1961). For detailed derivation we refer to Ahmed and Ullah (1999).

$$R(\hat{\boldsymbol{\theta}}^{JS};\boldsymbol{\theta}) = k - c_o m(k-1)\mathcal{E}\left[2\chi_{k+1}^{-2}(\Delta) - (k-3)\chi_{k+1}^{-4}(\Delta)\right] + c_o m(k+1)\Delta\mathcal{E}\left(\chi_{k+3}^{-4}(\Delta)\right),$$

where $c_o = (k-3)(m+2)^{-1}$ is optimal shrinkage constant with $m = n-k$. Further, $\Delta = \sigma^{-2}(A\boldsymbol{\theta})'\boldsymbol{\Lambda}(A\boldsymbol{\theta})$, with $A = I_k - n^{-1}1_k 1_k'\boldsymbol{\Lambda}$, I_k is the $k\times k$ identity matrix. The above relation reveals that $R(\hat{\boldsymbol{\theta}}^{JS};\boldsymbol{\theta}) \le R(\hat{\boldsymbol{\theta}};\boldsymbol{\theta})$ for all the values of Δ and strict inequality holds for some Δ. The maximum reduction in the risk for $\hat{\boldsymbol{\theta}}^{JS}$ is achieved at $\Delta = 0$, i.e., $\boldsymbol{\theta} = \theta 1_k$.

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}^{JS+};\boldsymbol{\theta}) = R(\hat{\boldsymbol{\theta}}^{JS};\boldsymbol{\theta}) &- (k-1)G_{k+1,m}(k_1;\Delta) \\ &- m(k-1)c_o\mathcal{E}\left[\left\{(k-3)\chi_{k+1}^{-4}(\Delta)\right.\right. \\ &\left.\left.-2\chi_{k+1}^{-2}(\Delta)\right\} I\left(\chi_{k+1}^2(\Delta)/\chi_m^2 \le c_o\right)\right] \\ &- mc_o\Delta\left[\left\{(k-3)\chi_{k+1}^{-2}(\Delta)\right.\right. \\ &\left.\left.-2\chi_{k+3}^{-4}(\Delta)\right\} I\left(\chi_{k+3}^2(\Delta)/\chi_m^2 \le c_o\right)\right. \\ &\left.+2\chi_{k+1}^{-2}(\Delta) I\left(\chi_{k+1}^2(\Delta)/\chi_m^2 \le c_o\right)\right] \\ &+ \Delta\{G_{k+1,m}(k_1;\Delta) - G_{k+3,m}(k_2;\Delta)\}, \end{aligned}$$

where $G_{p_1,\,p_2}(.\,,\,\Delta)$ is the cumulative distribution of a non-central F-distribution with p_1 and p_2 degrees of freedom and non-centrality parameter Δ with $k_1 = c_o m(k+1)^{-1}$ and $k_2 = c_o m(k+3)^{-1}$. The above risk expression leads to the relation that $R(\hat{\boldsymbol{\theta}}^{JS+};\boldsymbol{\theta}) \le R(\hat{\boldsymbol{\theta}}^{JS});\boldsymbol{\theta}) \le R(\hat{\boldsymbol{\theta}});\boldsymbol{\theta})$ for all the values of Δ, and strict inequality holds for some Δ. The maximum reduction in the risk is achieved at $\Delta = 0$. Therefore, the $\hat{\boldsymbol{\theta}}^{JS+}$ outperforms both $\hat{\boldsymbol{\theta}}^{JS}$ and $\hat{\boldsymbol{\theta}}$ in the entire parameter space induced by Δ and the upper of risk function of $\hat{\boldsymbol{\theta}}^{JS+}$ is obtained when $\Delta \to \infty$. Thus, for $\pi = \pi_o = (1 - c_o D^{-1})\ I(c_o D^{-1} \le 1)$ in $\hat{\boldsymbol{\theta}}^{S}$ given in (2), we obtain an optimal shrinkage strategy which provides a basis for optimally combining the estimation problems. This approach yields a well-defined data-based shrinkage estimator that combines estimation problem by shrinking a benchmark estimator to plausible alternative quantity. The optimal shrinkage estimator is similar to the most celebrated J–S estimator in which they shrink the benchmark estimator toward the null vector for estimating the mean vector of a multivariate normal distribution (see ►Multivariate Normal Distributions). There is no mystery about the origin; these estimators can shrink toward any point. Here we use the restricted estimator instead in the formulations and evaluations for optimal shrinkage strategy in one-way ANOVA. These formulations have been extended in regression problems pertaining to parametric, non-parametric and semi-parametric setup by a host of researchers including Ahmed (2001) and Ahmed et al. (2006, 2007). Now we present the formulation in a regression set-up.

Regression Model

Let us consider the linear regression model $\boldsymbol{y_n} = \boldsymbol{X_n}\boldsymbol{\beta} + \boldsymbol{\epsilon_n}$, where $\boldsymbol{y_n}$ is a $n\times 1$ random vector, $\boldsymbol{X_n}$ is a known $n\times k$ matrix $(n > k)$ of regression constant and, as the sample size n becomes infinitely large, $\lim_{n\to\infty}\left(\boldsymbol{X_n'}X_n\right)/n = \boldsymbol{C}$ where $\boldsymbol{C}$ is finite and nonsingular matrix. Further, $\boldsymbol{\beta}$ is a column vector of k unknown regression parameters, and $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)'$ are independent and identically distributed random variables with a distribution function F on real line $\Re = (-\infty,+\infty)$. Here we do not make any assumption about the functional form of the F. We assume that $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $E(\boldsymbol{\epsilon\epsilon}') = \sigma^2\boldsymbol{I_n}$, where σ^2 is unknown positive finite parameter.

We are primarily interested in the estimation of $\boldsymbol{\beta}$ when $\boldsymbol{\beta}$ is suspected to lie in the subspace defined by $\boldsymbol{H\beta} = \boldsymbol{h}$, where $\boldsymbol{H}$ is a given $q\times k$ matrix of rank $q \le k$ and $\boldsymbol{h}$ is a given $q\times 1$ vector of constants. The test statistic for the null hypothesis $H_o : \boldsymbol{H\beta} = \boldsymbol{h}$ is

$$D = (\boldsymbol{H\hat{\beta}} - \boldsymbol{h})'\left[\boldsymbol{H}\left(\boldsymbol{X_n'X_n}\right)^{-1}\boldsymbol{H}'\right]^{-1}(\boldsymbol{H\hat{\beta}} - \boldsymbol{h})/S^2,$$

$$S^2 = (\boldsymbol{Y_n} - \boldsymbol{X_n\hat{\beta}})'(\boldsymbol{Y_n} - \boldsymbol{X_n\hat{\beta}})/n-k,$$

where $\hat{\boldsymbol{\beta}}$ is the least square estimator (LSE) of $\boldsymbol{\beta}$. The restricted estimator of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{C_n^{-1}H'}\left(\boldsymbol{HC_n^{-1}}H'\right)^{-1}(\boldsymbol{H\hat{\beta}} - \boldsymbol{h}).$$

with $\boldsymbol{C}_n = \boldsymbol{X}_n'\boldsymbol{X}_n$. Consequently, placing $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ and D appropriately in (6) and (7), we can obtain shrinkage estimators for $\boldsymbol{\beta}$ respectively. Theses estimator preserve the dominance property over LSE. Recently, Nkurunziza and Ahmed (2010) extended shrinkage methodology to the matrix estimation. The literature on shrinkage estimation strategy is vast; and research on the statistical implications of these and other combining possibilities for a range of statistical models is ongoing and growing in statistical, econometric, and related scientific literatures.

About the Author

Ejaz Ahmed is Professor and Head of the Department of Mathematics and Statistics and Associate Faculty Member, Faculty of Engineering at the University of Regina. Before joining the University of Regina, he had a faculty position at the University of Western Ontario. Further,

he is a Senior Advisor to Sigma Analytics (Data Mining and Research), Regina. He developed the Master of Science and Doctoral Program in Statistics and produced the first Ph.D. student from this program. He has numerous published articles in scientific journals, both collaborative and methodological. Dr. Ahmed is an elected member of the International Statistical Institute and a Fellow of the Royal Statistical Society. He served as a Board of Director and Chairman of the Education Committee of the Statistical Society of Canada. Professor Ahmed was awarded the ISOSS (Islamic Countries Society of Statistical Sciences) Gold Medal (2001) in recognition of outstanding contribution to the development and promotion of statistical sciences "which has motivated Pakistanis, in particular, and Muslim statisticians, in general, to undertake research in this discipline" (Pakistan Journal of Statistics, *Special Issue in Honor of Professor S. Ejaz Ahmed*, Volume 18, No. 2, 2002, Forward).

Cross References

►Absolute Penalty Estimation
►Analysis of Variance
►Best Linear Unbiased Estimation in Linear Models
►Bootstrap Asymptotics
►Estimation
►James-Stein Estimator
►Likelihood
►Linear Regression Models
►Multivariate Statistical Analysis
►Optimal Shrinkage Preliminary Test Estimation
►Sufficient Statistics

References and Further Reading

Ahmed SE (2001) Shrinkage estimation of regression coefficients from censored data with multiple observations. In: Ahmed SE, Reid N (eds) Empirical Bayes and likelihood inference. Springer, NewYork, pp 103–120

Ahmed SE, Ullah B (1999) Improved biased estimation in an ANOVA model. Linear Algebra Appl 289:3–24

Ahmed SE, Hussein AA, Sen PK (2006) Risk comparison of some shrinkage M-estimators in linear models. J Nonparametric Stat 18:401–415

Ahmed SE, Doksum KA, Hossain S, You J (2007) Shrinkage, pretest and absolute penalty estimators in partially linear models. Aust N Z J Stat 49:435–454

James W, Stein C (1961) Estimation with quadratic loss. Proceeding of the fourth Berkeley symposium on Mathematical statistics and Probability. University of California Press, Berkeley, pp 361–379

Nkurunziza S, Ahmed SE (2010) Shrinkage drift parameter estimation for multi-factor Ornstein-Uhlenbeck processes. Appl Stochastic Models Bus Ind 26(2):103–124

Sclove SL (1968) Improved estimators for coefficients in linear regression. J Am Stat Assoc 63:596–606

Optimal Shrinkage Preliminary Test Estimation

S. Ejaz Ahmed, S. Chitsaz, S. Fallahpour
University of Windsor, Windsor, ON, Canada

Statistical models parameters are estimated in an effort to have knowledge about unknown quantities. In many situations, however, statisticians provide the estimation of the parameters by using not only information based on the sample, but other information as well. This information may be regarded as *nonsample information(NSI)* or *uncertain prior information (UPI)* about the parameter of interest. It is advantageous to utilize the *NSI* in the estimation procedure, especially when the information based on the sample may be rather limited or even the data quality is questionable. But in some experimental cases, it is not certain whether or not this information holds. Consider the data arising from tumor measurements in mice, for example, at various times following injection of carcinogens. Such data should be thought as coming from an *in vivo* experiment. Biologists are interested in estimating growth rate parameter θ when it suspected a priori that $\theta = \theta_0$. Such θ_0 can be obtained directly from *in vitro* experiments in which cell behavior may or may not be different because of different environmental conditions. Thus, biologists may have reason to suspect that θ_0 is the true value of the growth parameter for the *in vivo* experiment, but are not sure. Generally speaking, consequences of incorporating NSI depend on the *quality or reliability* of information introduced in the estimation process. This uncertain prior information in the form of the null hypothesis can be used in the estimation procedure. It is natural to perform a preliminary test on the validity of the *UPI* in the form of the parametric restrictions, and then choose between the restricted and unrestricted estimation procedure, depending upon the outcome of the preliminary test. This idea was initially conceived by T.A. Bancroft in 1944. However, this may be partly motivated by the remarks made by Berkson (1942). The *preliminary test estimators (PTE)* are widely used by researchers, as is evident from the extensive bibliographies and research articles. A standard reference on the PTE is Judge and Bock (1978).

For illustrative purposes, let us consider one sample problem. Suppose we observe $Y_1, \cdots, Y_n$ satisfying $Y_i = \theta + \epsilon_i, \quad i = 1, \cdots, n$, where the errors ϵ_i are independently and identically normally distributed with a mean of 0 and variance σ^2.

The statistical problem is to estimate θ when $\theta = \theta_0$ is suspected. To estimate θ, one need only to consider

the sufficient statistic $T(y) = \sum_{i=1}^{n} y_i$. The *benchmark or unrestricted estimator (UE)* θ is $\hat{\theta}^U = T(y)/n$, with variance $\frac{\sigma^2}{n}$.

Preliminary Test Estimation

The preliminary test estimation strategy involves a statistical test of the available NSI based on an appropriate test statistic and a decision is made on the outcome of the test. Thus, the *preliminary test estimator (PTE)* of θ denoted by $\hat{\theta}^P$ is defined as

$$\hat{\theta}^P = \theta_0 I(D_n < d_\alpha) + \hat{\theta}^U I(D_n \geq d_\alpha), \tag{1}$$

where D_n is any suitable test statistic and $I(A)$ is an indicator function of the set A, and d_α is the upper $100\alpha\%$ $(0 < \alpha < 1)$ point of the test statistic. For our estimation problem to test the null hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, we use the test statistic $D_n = \frac{n(\hat{\theta}^U - \theta_0)^2}{s^2}$, where s^2 is the sample estimate of σ^2. Under the null hypothesis, D_n follows an F-distribution. Thus, at the α-level of significance we reject the null hypothesis if $D_n \geq F_{(\alpha;\,1,\,n-1)}$, where $F_{(\alpha;\,1,\,n-1)}$ is the upper α-level critical values of the central F-distribution with $(1, n-1)$ degrees of freedom. Thus, we can replace d_α by $F_{(\alpha;\,1,\,n-1)}$ in (1) and the remaining discussion follows. By construction, $\hat{\theta}^P$ is a biased estimator and the mean squared error (MSE) of the $\hat{\theta}^P$ is given by

$$MSE(\hat{\theta}^P) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n}\left[H_{3,n-1}(F_a, \delta) - \delta\left\{2H_{3,n-1}(F_a, \delta) + H_{5,n-1}(F_b, \delta)\right\}\right], \tag{2}$$

where, $H_{k,\,n-1}(.\,,\,\delta/2,)$ is the cumulative distribution of a non-central F-distribution with k and $n-1$ degrees of freedom and non-centrality parameter $\delta/2$, where $\delta = n(\theta - \theta_0)^2/\sigma^2$. Further, $F_a = 1/3\ F_{(\alpha;\,1,\,n-1)}$ and $F_b = 1/5\ F_{(\alpha;\,1,\,n-1)}$.

The MSE expression reveals the typical characteristics of the preliminary test estimator. For small values of δ the performance of $\hat{\theta}^P$ is better than $\hat{\theta}^U$. Alternatively, for larger values of δ, the value of the $MSE(\hat{\theta}^P)$ increases, reaches its maximum after crossing the $MSE(\hat{\theta}^U) = \sigma^2/n$ and then monotonically decreases and approaches towards it. Further, as α, approaches to one, $MSE(\hat{\theta}^P)$ tends to $MSE(\hat{\theta}^U)$. Thus, the relative performance of $\hat{\theta}^P$ to $\hat{\theta}^U$ will also depend on the size of the preliminary test. It was recommended in the literature to use a level of significance of at least 0.25 or more for such preliminary testing to achieve reasonable MSE reduction. Use of such a large significance level helps maximizing the minimum efficiency of $\hat{\theta}^P$. Hence, the use of $\hat{\theta}^P$ was limited due to the large size of the preliminary test. To resolve this issue Ahmed (1992) introduced the shrinkage technique in the preliminary test estimation to overcome this difficulty. The proposed methodology remarkably improves upon the *PTE* with respect to the size of the preliminary test while maintaining the minimum efficiciecy.

A *shrinkage estimator (SE)* of θ is defined by $\hat{\theta}^{SR} = \pi\theta_0 + (1-\pi)\hat{\theta}^U$, where $\pi \in (0,1)$ is a coefficient reflecting the degree of confidence in the prior information. However, the key question in this type of estimator is how to select an optimal value for the shrinkage parameter π. In some situations, it may suffice to fix the parameter at some given value. The second choice, is to choose the parameter in a data-driven fashion by explicitly minimizing a suitable risk function. A common but also computationally intensive approach to estimate π is using cross-validation. On the other hand, from a Bayesian perspective one can employ the empirical Bayes technique. In this case π is treated as a hyper-parameter and may be estimated from the data by optimizing the marginal likelihood. For brevity sake, we assume that the value of π may be completely determined by the experimenter, depending upon the reliability of *NSI*. The SE performs better than $\hat{\theta}^U$ near the null hypothesis, but as the hypothesis error grows, $\hat{\theta}^{SR}$ may be considerably biased, inefficient and inconsistent, while the performance of $\hat{\theta}^U$ remains constant over such departures. Hence, this estimator is not preferable in its own right, however it can be used in constructing another estimators.

Shrinkage Preliminary Test Estimation

The *shrinkage preliminary test estimator (SPTE)* of θ denoted by $\hat{\theta}^{SP}$ is defined by replacing θ_0 by $\hat{\theta}^{SR}$ in the definition of $\hat{\theta}^P$ given in (1)

$$\hat{\theta}^{SP} = \{\pi\theta_0 + (1-\pi)\hat{\theta}^U\} I(D_n < d_\alpha) + \hat{\theta}^U I(D_n \geq d_\alpha). \tag{3}$$

The expression for the MSE of $\hat{\theta}^{SP}$ is

$$MSE(\hat{\theta}^{SP}) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n}\pi(2-\pi)H_{3,\,n-1}(F_a,\,\delta) + \frac{\sigma^2}{n}\pi\delta\{2H_{3,\,n-1}(F_a,\,\delta) - (2-\pi)H_{5,\,n-1}(F_b,\,\delta)\}. \tag{4}$$

It appears from the MSE expression of $\hat{\theta}^{SP}$ that as π increases, the MSE of $\hat{\theta}^{SP}$ becomes smaller than that of $\hat{\theta}^{SR}$ and $\hat{\theta}^P$, and approaches to MSE of $\hat{\theta}^U$ faster than that of $\hat{\theta}^P$. Further, $\hat{\theta}^{SP}$ dominates $\hat{\theta}^U$ over a wider range than $\hat{\theta}^P$. More importantly, $\hat{\theta}^{SP}$ has a good control on the maximum of MSE as compared with $\hat{\theta}^P$. Hence the $\hat{\theta}^{SP}$ provides much more meaningful size for the preliminary test than $\hat{\theta}^P$. The trick is in selecting the value of π. In this sense, $\hat{\theta}^{SP}$ can be considered as an optimal strategy.

Summary

When there is uncertainty concerning the appropriate statistical model-estimator to use in representing a data sampling process, the estimation rule based on preliminary test estimation provides a basis for optimally combining the estimation problems. This approach yields a well-defined data-based shrinkage preliminary test estimator that combines estimation problem by shrinking a benchmark estimator to plausible alternative quantity. Bearing in mind, neither $\hat{\theta}^{SP}$ nor $\hat{\theta}^{P}$ uniformly dominate $\hat{\theta}^{U}$. Thus, the performance of the estimators based on preliminary test procedures depends on the quality of the *NSI*. In this communication we considered formulations and evaluations for $\hat{\theta}^{SP}$ in one-sample-problems. The formulations have been extended to multiple estimation and regression problems in parametric, non-parametric and semi-parametric setup by a host of researchers including Ahmed (2001) and Ahmed et al. (2006, 2007). The literature on PTE is vast and research on the statistical implications of these and other combining possibilities for a range of statistical models is ongoing and growing in statistical and econometric literatures.

About the Author

For biography *see* entry ►Optimal Shrinkage Estimation.

Cross References

►Absolute Penalty Estimation
►Optimal Shrinkage Estimation
►Significance Testing: An Overview

References and Further Reading

Ahmed SE (1992) Shrinkage preliminary test estimation in multivariate normal distributions. J Stat Comput Sim 43:177–195

Ahmed SE (2001) Shrinkage estimation of regression coefficients from censored data with multiple observations. In: Ahmed SE, Reid N (eds) Empirical Bayes and likelihood inference. Springer, NewYork, pp 103–120

Ahmed SE, Hussein AA, Sen PK (2006) Risk comparison of some shrinkage M-estimators in linear models. J Nonparametric Stat 18:401–415

Ahmed SE, Doksum KA, Hossain S, You J (2007) Shrinkage, pretest and absolute penalty estimators in partially linear models. Aust NZ J Stat 49:435–454

Bancroft TA (1944) On biases in estimation due to the use of preliminary tests of significance. Ann Math Stat 15:190–204

Berkson J (1942) Test of significance considered as evidence. J Am Stat Assoc 37:325–335

Judge GG, Bock ME (1978) The statistical implication of pretest and Stein-rule estimators in econometrics. North-Holland, Amsterdam

Optimal Statistical Inference in Financial Engineering

Masanobu Taniguchi
Professor
Waseda University, Tokyo, Japan

The field of financial engineering has developed as a huge integration of economics, mathematics, probability theory, statistics, time series analysis, operation research etc. We describe financial assets as ►stochastic processes. Using stochastic differential equations, probabilists developed a highly sophisticated mathematical theory in this field. On the other hand empirical people in financial econometrics studied various numerical aspects of financial data by means of statistical methods.

Black and Scholes (1973) provided the modern option pricing theory assuming that the price process of an underlying asset follows a geometric Brownian motion (see ►Brownian Motion and Diffusions). But, a lot of empirical studies for the price processes of assets show that they do not follow the geometric Brownian motion. Concretely, we often observe that the sample autocorrelation function

$$\hat{\rho}_{X_t}(l) = \frac{\sum_{t=1}^{n-|l|}(X_{t+l}-\overline{X}_n)(X_t-\overline{X}_n)}{\sum_{t=1}^{n}(X_t-\overline{X}_n)^2}, \left(\overline{X}_n = n^{-1}\sum_{t=1}^{n} X_t\right)$$

of the return process $\{X_t\}$ of assets with time lag l becomes near 0 for $l \neq 0$, i.e., X_t's are almost uncorrelated, and that $\hat{\rho}_{X_t^2}(l)$, $l \neq 0$, of the square-transformed return X_t^2 are not near 0 for $l \neq 0$, i.e., X_t^2's are not uncorrelated. Hence we may suppose that

"return processes $\{X_t\}$ are non-Gaussian dependent." (1)

Based on this, it is important to investigate which stochastic models can describe the actual financial data sufficiently, and how to estimate the proposed models optimally, which lead to the theme of this section. Let $\{\mathbf{X}_t\}$ be an m-dimensional vector process whose components consist of the return of assets. A typical candidate for (1) is the following non-Gaussian linear process :

$$\mathbf{X}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} A_{\boldsymbol{\theta}}(j)\mathbf{U}_{t-j}, \quad (2)$$

where $\boldsymbol{\mu}$ is an m-dimensional non-random vector, $A_{\boldsymbol{\theta}}(j)$'s are $m \times m$-non-random matrices, $\boldsymbol{\theta}$ is a p-dimensional unknown parameter and $\mathbf{U}_s$'s are i.i.d. m-dimensional random vectors with probability density function $g(\mathbf{u})$, $\mathbf{u} \in \mathbf{R}^m$.

Lucien LeCam (e.g., LeCam 1986) established the most important and sophisticated foundation of the general statistical asymptotic theory. He introduced the concept of local asymptotic normality (LAN) for the likelihood ratio between contiguous hypotheses of general statistical models. Once LAN is proved, the asymptotic optimality of various statistical methods (estimators, tests and discriminant statistics etc.) is described in terms of the LAN property.

Under appropriate regularity conditions, Taniguchi and Kakizawa (2000) proved the LAN for (2), and showed the asymptotic optimality of the maximum likelihood estimator (MLE) for $\boldsymbol{\theta}$. The philosophy of optimal statistical inference in financial engineering is to develop the financial analysis (option pricing, VaR problem, portfolio construction and credit rating etc.) based on the optimally estimated statistical models. In what follows we explain a few topics in this stream.

Suppose that $\mathbf{X}_t$ is the random return on m assets at time t, and that $\mathbf{X}_t$ is generated by (2). Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ be the vector of portfolio weights. Then the return of portfolio is $\boldsymbol{\alpha}'\mathbf{X}_t$, and the mean and variance are, respectively, given by $\boldsymbol{\alpha}'\boldsymbol{\mu}$ and $\boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}$, where $\Sigma = Var(\mathbf{X}_t)$. Optimal portfolio weights have been proposed by various criteria. The most famous one is given by the solution of

$$\begin{cases} \max_{\boldsymbol{\alpha}} \{\boldsymbol{\alpha}'\boldsymbol{\mu} - \beta\boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}\} \\ \text{subject to } \mathbf{e}'\boldsymbol{\alpha} = 1 \end{cases} \tag{3}$$

where $\mathbf{e} = (1, \ldots, 1)'$ (m×1-vector), and β is a given positive number. Then the optimal portfolio for (3) becomes

$$\boldsymbol{\alpha}_I = \frac{1}{2\beta}\left\{\Sigma^{-1}\boldsymbol{\mu} - \frac{\mathbf{e}'\Sigma^{-1}\boldsymbol{\mu}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}}\Sigma^{-1}\mathbf{e}\right\} + \frac{\Sigma^{-1}\mathbf{e}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}}. \tag{4}$$

The criterion (3) is so called the "mean-variance method." If we use the other criterion, the optimal portfolio $\boldsymbol{\alpha}_{opt}$ is of the form

$$\boldsymbol{\alpha}_{opt} = g(\boldsymbol{\mu}, \Sigma), \tag{5}$$

i.e., a measurable function of $\boldsymbol{\mu}$ and Σ. A natural estimator of $\boldsymbol{\alpha}_{opt}$ is

$$\hat{\boldsymbol{\alpha}}_{opt} \equiv g(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) \tag{6}$$

where $\hat{\boldsymbol{\mu}}$ is the sample mean and $\hat{\Sigma}$ is the sample covariance matrix from the observations with length n. Shiraishi and Taniguchi (2008) gave a necessary and sufficient condition for $\hat{\boldsymbol{\alpha}}_{opt}$ to be asymptotically efficient in terms of the spectral density matrix of $\{\mathbf{X}_t\}$. This condition shows that if $\{\mathbf{X}_t\} \sim VARMA(p_1, p_2)$ with $p_1 < p_2$, then $\hat{\boldsymbol{\alpha}}_{opt}$ is not asymptotically efficient, which gives a strong warning for use of the classical mean-variance estimator. In view of the asymptotic efficiency, based on LAN, Shiraishi and Taniguchi (2008) showed that the MLE of $\boldsymbol{\alpha}_{opt}$ is asymptotically efficient.

So far we have assumed that the return process $\{\mathbf{X}_t\}$ is stationary. However, stationary models are not plausible to describe the real world. In fact, time series data with a long stretch often contain slow or rapid changes in the spectra. For this Dahlhaus (1997) introduced an important class of non-stationary processes, called locally stationary processes which have the time varying spectral density $f(u, \lambda)$ where u is a standardized time parameter and λ is the frequency. In the case of locally stationary return processes, Shiraishi and Taniguchi (2007) discussed the optimal estimation of $\boldsymbol{\alpha}_{opt}$.

Various approaches in financial engineering with optimal statistical properties can be found in Taniguchi et al. (2008).

About the Author

Masanobu Taniguchi joined the Department of Mathematics, Hiroshima University, and the Department of Mathematical Science, Osaka University, in 1983 and 1990, respectively. He was a Visiting Professor at the University of Bristol, UK, in 2000. He is currently a Professor in the Department of Applied Mathematics, Waseda University, Japan. His main contributions in time series analysis are collected in his book *Asymptotic Theory of Statistical Inference for Time Series* (New York, Springer, 2000). He is also a co-author (with Junichi Hirukawa and Kenichiro Tamaki) of the text *Optimal Statistical Inference in Financial Engineering* (Chapman and Hall, 2007). Professor Taniguchi is a Fellow of the Institute of Mathematical Statistics. He was awarded the Ogawa Prize (Japan, 1989), the Econometric Theory Award (USA, 2000), and the Japan Statistical Society Prize (2004). Currently, he is Editor of the *Journal of the Japan Statistical Society* (2006–).

Cross References

►Asymptotic Relative Efficiency in Estimation
►Brownian Motion and Diffusions
►Econometrics
►Portfolio Theory
►Quantitative Risk Management
►Statistical Inference
►Statistical Inference: An Overview
►Statistical Modeling of Financial Markets
►Stochastic Processes
►Stochastic Processes: Applications in Finance and Insurance

References and Further Reading

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–654

Dahlhaus R (1997) Fitting time series models to nonstationary processes. Ann Stat 25:1–37

LeCam L (1986) Asymptotic methods in statistical decision theory. Springer, New York

Shiraishi H, Taniguchi M (2007) Statistical estimation of optimal portfolios for locally stationary returns of assets. Int J Theor Appl Finance 10:129–154

Shiraishi H, Taniguchi M (2008) Statistical estimation of optimal portfolios for non-Gaussian dependent returns of assets. J Forecast 27:193–215

Taniguchi M, Kakizawa Y (2000) Asymptotic theory of statistical inference for time series. Springer, New York

Taniguchi M, Hirukawa J, Tamaki K (2008) Optimal statistical inference in financial engineering. Chapman and Hall/CRC, New York

Optimal Stopping Rules

Albert N. Shiryaev
Professor, Head of the Probability Theory Department at the Moscow State University
Moscow, Russia

The theory of *Optimal Stopping* was considerably stimulated by A. Wald (1947). He showed that – in contrast to the classical methods of the Mathematical Statistics, according to which the decision is taken in a fixed (and nonrandom) time – the methods of the sequential analysis take observations sequentially and the decision is taken, generally speaking, at a random time whose value is determined by the rule (strategy) of observation of a statistician. Wald discovered the remarkable advantage of the sequential methods in the problem of testing (from i.i.d. observations) two simple hypotheses. He proved that there is a sequential method (sequential probability-ratio test) which requires on average a smaller number of observations than any other method using fixed sample size (and the same probabilities of wrong decisions). It turned out that the problem of optimality of a sequential statistical decision can be reformulated as an "optimal stopping problem," and this was the essential step in constructing the General Optimal Stopping Theory.

The basic notions and results of the Optimal Stopping Theory are the following.

In the discrete-time case, one assumes that a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathsf{P})$ is given, where we interpret $\mathcal{F}_n$ as the information available up to time n. Let $G = (G_n)_{n\geq 0}$ be a sequence of random variables such that each G_n is $\mathcal{F}_n$-measurable. All our decisions in regard to optimal stopping at time n must be based on the information $\mathcal{F}_n$ only. In other words, no anticipation is allowed.

By definition, a *Markov time* is a random variable $\tau\colon \Omega \to \{0, 1, \ldots, \infty\}$ such that $\{\tau \leq n\} \in \mathcal{F}_n$ for all $n \geq 0$. A Markov time with the property $\tau < \infty$ is usually called a *stopping time* (the class of all stopping times will be denoted by $\mathfrak{M}$).

A basic problem of the Optimal Stopping Theory consists of finding the *value* function

$$V = \sup_{\tau\in\mathfrak{M}} \mathsf{E}G_\tau$$

and a stopping time τ^* (if it exists) such that $\mathsf{E}G_{\tau^*} = V$. (We assume that $\mathsf{E}\sup_{n\geq 0}|G_n| < \infty$.)

For solving this problem it is useful to introduce the value functions

$$V_n^N = \sup_{\tau\in\mathfrak{M}_n^N} \mathsf{E}G_\tau, \quad 0 \leq n \leq N < \infty, \quad \text{and} \quad V_n = \sup_{\tau\in\mathfrak{M}_n} \mathsf{E}G_\tau, \quad 0 \leq n < \infty,$$

where $\mathfrak{M}_n^N = \{\tau \in \mathfrak{M}\colon n \leq \tau \leq N\}$ and $\mathfrak{M}_n = \{\tau \in \mathfrak{M}\colon \tau \geq n\}$.

The usual method of finding the value functions V_n^N is the *method of backward induction*, which deals with a sequence of *random* variables $\left(S_n^N\right)_{0\leq n\leq N}$ defined recursively as follows:

$$S_n^N = G_N \qquad \text{for } n = N,$$

$$S_n^N = \max\left\{G_n, \mathsf{E}\left(S_{n+1}^N \mid \mathcal{F}_n\right)\right\} \qquad \text{for } n = N-1, \ldots, 0.$$

The method also suggests that we consider the time

$$\tau_n^N = \inf\left\{n \leq k \leq N\colon S_k^N = G_k\right\} \quad \text{for } 0 \leq n \leq N.$$

The following theorem (which we formulate for the case $0 \leq n \leq N < \infty$) plays a central role in the Optimal Stopping Theory.

Theorem 1 *The stopping time τ_n^N is optimal ($\mathsf{E}G_{\tau_n^N} = V_n^N$) in the class $\mathfrak{M}_n^N$. The value V_n^N equals $\mathsf{E}S_n^N$.*

To find the value functions $V_n = \sup_{\tau\geq n} \mathsf{E}G_\tau$, $n \geq 0$, we consider the sequence of random variables $(S_n)_{n\geq 0}$ defined by

$$S_n = \operatorname*{ess\,sup}_{\tau\geq n} \mathsf{E}(G_\tau \mid \mathcal{F}_n)$$

and the stopping time

$$\tau_n = \inf\{k \geq n\colon S_k = G_k\}.$$

Here ess sup is an abbreviation for essential supremum; the sequence $(S_n)_{n\geq 0}$ is often referred to as the *Snell envelope* of G.

Theorem 2 *The following recurrent relations hold:*

$$S_n = \max\{G_n, E(S_{n+1} \mid \mathcal{F}_n)\}.$$

If $\tau_n < \infty$, *then this stopping time* τ_n *is optimal:* $EG_{\tau_n} = V_n$ *and* $ES_{\tau_n} = V_n$.

Theorem 3 *Under assumption* $E \sup_{n\geq 0} |G_n| < \infty$ *we have* $S_n = \lim_{N\to\infty} S_n^N$ *and* $V_n = \lim_{N\to\infty} V_n^N$.

Theorems 1–3 cover the main results of the so-called *martingale approach* to the optimal stopping. A lot of attention in the Optimal Stopping Theory is paid to the so-called *Markovian approach*. This approach assumes that the gain functions G_n admit the *Markovian representation*, *i.e.*, there exist a Markov process $X = (X_n)_{n\geq 0}$ defined on $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathsf{P})$ and a measurable function $G = G(x)$ such that $G_n = G(X_n)$ for all $n \geq 0$. The assumption about the Markovian representation allows one to use for solving the optimal stopping problems the theory of ▶Markov processes, which has a well-developed tools, methods, and remarkable results.

To describe the Markovian approach to the Optimal Stopping, let us assume that $G = G(x)$ is a measurable function, $X = (X_n)_{n\geq 0}$ is a homogeneous Markov chain (see ▶Markov Chains) with values in $\mathbb{R}$. We denote by P_x the distribution of X under assumption $X_0 = x \in \mathbb{R}$ and by E_x the expectation with respect to P_x.

We assume $\mathsf{E}_x \sup_{n\geq 0} |G(X_n)| < \infty$, $x \in \mathbb{R}$, and use the following notation: for $n \geq 0$,

$$V^n(x) = \sup_{\tau\in\mathfrak{M}_0^n} \mathsf{E}_x G(X_\tau), \qquad V(x) = \sup_{\tau\in\mathfrak{M}} \mathsf{E}_x G(X_\tau);$$

for $N \geq 0$ fixed and $n \leq N$,

$$C^{N-n} = \{x \in \mathbb{R}: V^{N-n}(x) > G(x)\},$$
$$D^{N-n} = \{x \in \mathbb{R}: V^{N-n}(x) = G(x)\},$$

and

$$\tau_D^N = \inf\{0 \leq n \leq N: X_n \in D^{N-n}\}.$$

Denote also by T the transition operator of X:

$$TF(x) = \mathsf{E}_x F(X_1),$$

where $F(x)$ is a real function such that $\mathsf{E}_x F(|X_1|) < \infty$.

Theorem 4 *For all* $n \leq N$ *the function* $V^n(x)$ *satisfies the Wald–Bellman equation*

$$V^n(x) = \max\{G(x), TV^{n-1}(x)\}$$

with $V^0 = G$. *The stopping time* τ_D^N *is optimal:* $\mathsf{E}_x G\left(X_{\tau_D^N}\right) = V^N(x)$.

We see that optimal stopping time τ_D^N has a very transparent form: if $x_n \in C^{N-n}$, then we continue observations, and we stop, if $x_n \in D^{N-n}$. (The sets C^{N-n} and D^{N-n} are called sets of *continuation* and *stopping* of observations, respectively.)

For case of infinite horizon ($N = \infty$) the following theorem holds.

Theorem 5 *Under assumption* $E \sup_{n\geq 0} |G(X_n)| < \infty$, $x \in \mathbb{R}$, *the value function* $V(x)$ *solves the Wald–Bellman equation*

$$V(x) = \max\{G(x), TV(x)\}$$

for $x \in \mathbb{R}$. *If the stopping time* $\tau_D = \inf\{n \geq 0: X_n \in D\}$, *where* $D = \{x \in \mathbb{R}: V(x) = G(x)\}$, *is finite* ($\tau_D < \infty$), *then this stopping time is optimal:* $\mathsf{E}_x G(X_{\tau_D}) = V(x)$, $x \in \mathbb{R}$.

Remark 1. In Theorems 2 and 5 it suffices to assume that $\tau_n < \infty$ (P-a.s.) and $\tau_D < \infty$ (P-a.s.), respectively. For case $\tau_n = \infty$ and $\tau_D = \infty$ we put $G_{\tau_n} = 0$ and $G(X_{\tau_D}) = 0$.

The results formulated above give the basis for solving different concrete optimal stopping problems in the theory of probability, mathematical statistics, financial mathematics.

Above we presented the optimal stopping theory only for the discrete-time case. For the case of continuous time, the formulations of problems are very similar (one should consider the continuous-time filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathsf{P})$ instead of the discrete-time filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 0}, \mathsf{P})$). The corresponding general theory is exposed in Chow et al. (1971), DeGroot (1970), Peskir and Shiryaev (2006) and Shiryaev (2007), where one can also find solutions to many concrete optimal stopping problems from different fields of probability theory, mathematical statistics, financial mathematics.

About the Author

Albert Nikolayevich Shiryaev was born on 12 October 1934. He has been a professor at the department of Mechanics and Mathematics of Moscow State University, since 1971. He has also been working in Steklov Mathematical Institute. He earned his candidate degree in 1961 (Andrey Kolmogorov was his advisor) and a doctoral degree in 1967. He was elected a corresponding member of the Russian Academy of Sciences in 1997. Professor Shiryaev was awarded the A.N. Markov Prize of USSR Academy of Sciences (1974), the A.N. Kolmogorov Prize of Russian Academy of Sciences (1994), and Humboldt Research Award (1996). He was President of the Bernoulli Society (1989–1991), President of the Russian Actuarial Society (1994–1998) and President of the Bachelier Finance Society (1998–1999). Professor Shiryaev is a Honorary Fellow of the Royal Statistical Society (1985), Member of the Academia Europea (1990) and Member of the New York

Academy of Science (1997). He holds two Honorary Doctorates. He is the author of more than 150 research papers in probability theory, mathematical and applied statistics, stochastic control, and financial mathematics. Professor Shiryaev is internationally known for his work in probability theory, statistics and financial mathematics. He is one of the leading experts in the field of Optimal stopping rules, and has published the text *Optimal Stopping Rules* (Springer, 3rd Rev. ed., 2007), and *Optimal Stopping and Free-Boundary Problems* (with G. Peskir; Lectures in Mathematics, ETH Zürich, Birkhäuser, 2005). He was advisor of more than 50 Ph.D. students (including one of our contributors, Professor Boris Rozovsky).

Professor Shiryaev is the third head in the history of the department (since 1996), the first and founding head was A.N. Kolmogorov in 1936, and the second Kolmogorov's student B.V. Gnedenko (from 1966 to 1995).

Cross References

►Markov Chains
►Martingales
►Sequential Probability Ratio Test
►Sequential Sampling
►Surveillance

References and Further Reading

Chow YS, Robbins H, Siegmund D (1971) Great expectation : the theory of optimal stopping. Houghton Mifflin, Boston

DeGroot MH (1970) Optimal statistical decisions. McGraw-Hill, New York

Peskir G, Shiryaev AN (2006) Optimal stopping and free-boundary problems. Birkhäuser, Basel

Shiryaev AN (1999) Essentials of stochastic finance: facts, models, theory. World Scientific, River Edge, NJ

Shiryaev AN (2007) Optimal stopping rules. Springer, New York (Soft cover reprint of the 1978 edition, with the new preface by the author)

Wald A (1947) Sequential analysis. Wiley, New York; Chapman and Hall, London

Optimality and Robustness in Statistical Forecasting

Yuriy S. Kharin
Professor, Head of the Department of Mathematical Modeling and Data Analysis, Director
Belarus State University, Minsk, Belarus

Introduction

Many applied problems in engineering, economics, finance, and medicine lead to the important problem of statistical data analysis – forecasting of time series. The mathematical substance of the forecasting problem is quite simple: to estimate the future value $x_{T+\tau} \in \mathbb{R}^d$ of the d-variate time series in $\tau \in \mathbb{N}$ steps ahead by $T \in \mathbb{N}$ successive observations $\{x_1, \ldots, x_T\} \subset \mathbb{R}^d$.

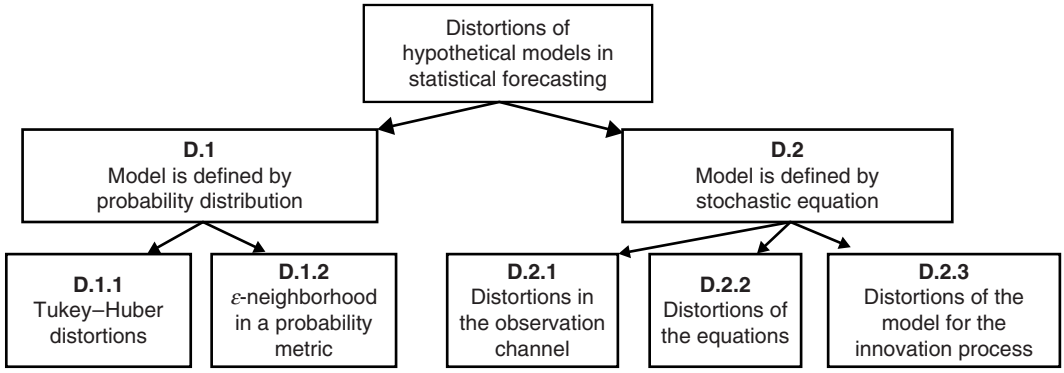
We can distinguish two stages in the history of attacks on the forecasting problem. The research on the first stage (before the year 1974) was oriented to the development of forecasting statistics and algorithms that minimize the mean square risk (error) of forecasting for a set of basic mathematical models, e.g., stationary time series with some known spectral density, time series with a trend from some known parametric family, and autoregressive integrated moving-average (ARIMA) time series (Bowerman and O'Connel 1993).

It was detected by many researchers that the "optimal" forecasting algorithms on the real statistical data have the risk values that are much more than the expected theoretical values. In his lecture at the World Congress of Mathematicians in 1974, Peter Huber explained the reason for this strange situation (Huber 1974): "Statistical inferences (including statistical forecasts) depend only in part upon the observations. An equally important base is formed by prior assumptions about the underlying situation." The system of prior assumptions is called the hypothetical model of the data M_0. In applied problems, the hypothetical model assumptions M_0 are often distorted, and this fact leads to the unstability of the "optimal" forecasting statistics that are optimal only under M_0. Huber has proposed to construct robust statistical inferences that are "weak-sensitive w.r.t. small distortions of the hypothetical model M_0"; this event has opened the second stage in statistical forecasting of time series. The present-day state of the research in ►robust statistics is displayed in analytic reviews (Davies and Gather 2004; Maronna et al. 2006; Morgenthaler 2007).

Distortions of Hypothetical Models

The probability model of the observed time series under distortions is determined by the family of probability distributions $\{P^{\varepsilon}_{T,\theta^0}(A),\ A \in B^{Td}\colon\ T \in \mathbb{N},\ \theta^0 \in \Theta \subseteq \mathbb{R}^m,\ \varepsilon \in [0, \varepsilon_+]\}$, where B^{Td} is the Borel σ-algebra in $\mathbb{R}^{Td}$, θ^0 is an unknown true value of the model parameters, ε is the distortion level, $\varepsilon_+ \geq 0$ is its maximal admissible value. If $\varepsilon_+ = 0$, then the distortions are absent, and we have the hypothetical model M_0.

A short scheme of classification for typical distortions of hypothetical models is presented in Fig. 1; a more detailed scheme of classification is given in Kharin (2008). Let us describe briefly the distortions indicated on Fig. 1.



Optimality and Robustness in Statistical Forecasting. Fig. 1 Classification for types of distortions

With respect to (w.r.t.) the form of presentation of the hypothetical model M_0, the set of all types of distortions can be split into two classes: the model in the explicit form (**D.1**), i.e., in the form of some probability distribution $P^0(\cdot)$; the model in the implicit form (**D.2**) determined by a stochastic equation

$$x_t = G(x_{t-1}, \ldots, x_{t-s}, u_t, u_{t-1}, \ldots, u_{t-L}; \theta^0),\ t \in \mathbb{Z},$$

where $u_t \in \mathbb{R}^v$ is an innovation process (usually the white noise), $s, L \in \mathbb{N}$ are some natural numbers indicating the memory depth, $\theta^0 \in \Theta \subseteq \mathbb{R}^m$ is the vector of model parameters.

Tukey–Huber distortions (**D.1.1**) for the observation vector X are described by the mixture: $p(X) = (1 - \varepsilon)p^0(X) + \varepsilon h(X)$, where $p^0(\cdot)$ is some "non-distorted" (hypothetical) p.d.f., $h(\cdot)$ is the so-called contaminating p.d.f., $\varepsilon \in [0, \varepsilon_+)$ is the distortion level. If $\varepsilon = 0$, then $p(\cdot) = p^0(\cdot)$, and distortions are absent.

Distortions of the type **D.1.2** are described by ε-neighborhoods in any probability metric: $0 \le \rho(p(\cdot), p^0(\cdot)) \le \varepsilon$, where $\rho(\cdot)$ is some probability metric.

The class **D.2** consists of three subclasses. The subclass **D.2.1** describes distortions in the observation channel: $X = H(X^0, V)$, where $X^0 = \left(x_k^0\right) \in \mathbb{R}^{Td}$ is the "non-observable prehistory" of the process, $X \in \mathbb{R}^{Td}$ is the vector of observations, that is the "observable prehistory," $V = \left(v_1', \ldots, v_T'\right) \in \mathbb{R}^{Tl}$ is the non-observable random vector of distortions (errors in the observation channel), $H(\cdot)$ is a function that describes the registration algorithm.

The subclass (**D.2.1**) includes five types of distortions.

Additive (**D.2.1.1**) and multiplicative (**D.2.1.2**) distortions in the observation channel are described by the equations

$$x_t = x_t^0 + \varepsilon v_t,\ x_t = (1 + \varepsilon v_t)x_t^0,\ t \in \mathbb{N},$$

respectively, where $\{v_t\}$ are i.i.d. random variables, $\mathbb{E}\{v_t\} = 0$, $\mathbb{D}\{v_t\} = \sigma^2 < +\infty$.

The subclass **D.2.1.3** (ε-non-homogeneities) includes the cases where the random vectors of distortions $v_t \in \mathbb{R}^l$ are nonidentically distributed, but their probability distributions differ not more than on ε in some probability metric.

The subclass **D.2.1.4** describes the ▸outliers in the data. The replacement outliers (RO) and the additive outliers (AO) in the observation channel are described by the equations:

$$x_t = (1 - \xi_t)x_t^0 + \xi_t v_t,\ x_t = x_t^0 + \xi_t v_t,\ t \in \mathbb{N},$$

respectively, where $\{\xi_t\}$ are i.i.d. Bernoulli random variables, $\mathbb{P}\{\xi_t = 1\} = 1 - \mathbb{P}\{\xi_t = 0\} = \varepsilon$, $\{v_t\}$ are random variables describing outliers, ε is the probability of the outlier appearance, and $\mathbb{E}\{v_t\}, \mathbb{D}\{v_t\}$ characterize the level of outliers.

The subclass **D.2.1.5** considers the missing values in $X^0 \in \mathbb{R}^{Td}$. To describe the missing values, it is convenient to define the binary $(T \times d)$-matrix $O = (o_{ti})$: $o_{ti} = \{1,$ if x_{ti}^0 is observed; $0,$ if x_{ti}^0 is a missing value$\}$.

The subclass **D.2.2** describes distortions of the generating stochastic equation ("misspecification errors") and includes two types of distortions:

- Parametric distortions (**D.2.2.1**), when instead of the true parameter value θ^0 we get (or estimate by statistical data) a different value $\tilde{\theta}$, with $|\tilde{\theta} - \theta^0| \le \varepsilon$
- Functional distortions (**D.2.2.2**), when instead of the true function $G(\cdot)$ we get a different function $\tilde{G}(\cdot)$, and in some metric $\|\tilde{G}(\cdot) - G(\cdot)\| \le \varepsilon$

The subclass **D.2.3** describes distortions of the innovation process $u_t \in \mathbb{R}^v$ in the generating stochastic

equation and includes distortions of three types: ε-non-homogeneities (**D.2.3.1**), probabilistic dependence (**D.2.3.2**), "innovation outliers" (**D.2.3.3**).

Characteristics of Optimality and Robustness in Forecasting

Let $\hat{x}_{T+\tau} = f(X)$: $\mathbb{R}^{Td} \to \mathbb{R}^d$ be any forecasting statistic. Its performance is evaluated by the mean square risk of forecasting:

$$r_\varepsilon = r_\varepsilon(f) = \mathbb{E}_\varepsilon\{\|\hat{x}_{T+\tau} - x_{T+\tau}\|^2\} \geq 0,\ \varepsilon \in [0, \varepsilon_+],$$

where $\mathbb{E}_\varepsilon\{\cdot\}$ is the expectation symbol w.r.t. the probability distribution $P^\varepsilon_{T,\theta^0}(\cdot)$. If the distortions are absent ($\varepsilon = 0$) and the hypothetical model M_0 is valid, this functional is called the hypothetical risk $r_0 = r_0(f)$. The guaranteed (upper) risk is

$$r_+ = r_+(f) = \sup_{0 \leq \varepsilon \leq \varepsilon_+} r_\varepsilon(f),$$

where the supremum is taken on all admissible distortions of the hypothetical model M_0.

Let further $\hat{x}^0_{T+\tau} = f^0(X; \theta^0)$ be the optimal forecasting statistic under the known hypothetical model M_0 ($\varepsilon = 0$) that gives the minimal value to the hypothetical risk:

$$r_0 = r_0(f^0) = \inf_{f(\cdot)} r_0(f).$$

In practice, the family of the so-called plug-in forecasting statistics is often used: $\hat{x}_{T+\tau} = f(X) := f^0(X; \hat{\theta})$, where $\hat{\theta} \in \mathbb{R}^m$ is some consistent statistical estimator of the unknown parameter θ^0 based on the observed time series X. A forecasting statistic $\hat{x}_{T+\tau} = f(X)$ is called the asymptotically optimal forecasting statistic if $\lim_{T\to\infty}(r_0(f) - r_0(f^0)) = 0$.

The risk unstability coefficient κ is the relative increment of the guaranteed risk w.r.t. the minimal hypothetical risk

$$\kappa = \kappa(f) = (r_+(f) - r_0)/r_0 \geq 0.$$

For any $\delta > 0$ define another characteristic of robustness that is quite useful in applications:

$$\varepsilon^* = \varepsilon^*(\delta) = \sup\{\varepsilon \in [0, \varepsilon_+] :\ \kappa(f) \leq \delta\},$$

that is called the δ-admissible distortion level. It indicates the maximal level of distortions for which the relative increment of the risk is yet not greater than the fixed value $\delta \cdot 100\%$.

The smaller the value κ is and the greater the value ε^* is, the more robust the forecasting statistic is. The minimax robust forecasting statistic $\hat{x}^*_{T+\tau} = f^*(X)$ minimizes the risk unstability coefficient

$$\kappa(f^*) = \inf_{f(\cdot)} \kappa(f).$$

In Hampel et al. (1986), a characteristic of the "qualitative robustness" is introduced – the Hampel breakdown point ε^{**}; it is the maximal fraction of "arbitrary large outliers" in the sample X such that the considered forecasting statistic $f(X)$ is bounded: $\varepsilon^{**} = \max\{\varepsilon \in [0,1] : \sup|f(X)| \leq C < +\infty\}$.

Some Approaches to Construct Robust Forecasting Statistics

There are three main approaches to construct robust forecasting statistics.

The first approach is based on minimization of the risk unstability coefficient $\kappa(f)$ in some family of forecasting statistics $\mathcal{F}$: $\kappa(f) \to \min_{f\in\mathcal{F}}$. This approach is theoretically complicated. Because of the computational complexity, the minimization problem can be solved for some cases only; relevant examples are given in Kharin (2008).

The second approach is the "plug-in" approach: $\tilde{x}_{T+\tau} = f^0(X; \tilde{\theta})$, where $\tilde{\theta}$ is some robust estimator of parameters θ^0 by the distorted data X; the methods for robust estimation of parameters can be found in Davies and Gather (2004), Hampel et al. (1986), Kharin (2008), Maronna et al. (2006), Morgenthaler (2007), and Rousseeuw and Leroy (1987). Robustness of the "plug-in" forecasting statistics can be evaluated by the characteristics indicated in the previous section ("►Characteristics of Optimality and Robustness in Forecasting").

The third approach is based on the idea of preliminary filtering ("cleaning") of the observed time series X. Some methods of forecasting based on filtering are discussed in Maronna et al. (2006) and Rousseeuw and Leroy (1987).

About the Author

Yuriy S. Kharin is Director of the Research Institute for Applied Problems of Mathematics and Informatics at the Belarusian State University, Minsk, Republic of Belarus. He is also Head of the Department of Mathematical Modeling and Data Analysis at the Belarusian State University, Correspondent Member of the National Academy of Sciences (from 2004), Past President of the Belarusian Statistical Association. He has authored and coauthored more than 350 papers, 15 monographs and textbooks, including the monograph *Robustness in Statistical Pattern Recognition* (Kluwer 1996). Professor Kharin has received the National State Prize in Science of the Republic of Belarus (2002).

Cross References

►Forecasting with ARIMA Processes
►Forecasting: An Overview
►Robust Statistics

►Time Series
►Time Series Regression

References and Further Reading

Bowerman B, O'Connel RT (1993) Forecasting and time series. Wadworth, Belmont
Davies PL, Gather U (2004) Robust statistics. In: Handbook of computational statistics. Springer, Berlin, pp 655–695
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York
Huber PJ (1974) Some mathematical problems arising in robust statistics. In: Proceedings of the international congress of mathematics. Vancouver, pp 821–824
Kharin Yu S (2008) Optimality and robustness in statistical forecasting. BSU, Minsk (in Russian)
Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, Chichester
Morgenthaler S (2007) A survey of robust statistics. Stat Meth Appl 15:271–293
Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Chapman and Hall, London

Optimum Experimental Design

Anthony C. Atkinson
Professor of Statistics
London School of Economics, London, UK

Introduction

Experimental design is concerned with the allocation of treatments to units. The methods of optimum design were originally developed for the choice of those values of the explanatory variables x in a regression model at which observations should be taken (Smith 1918). For example, in a chemical experiment there may be several factors, such as time of reaction, temperature, pressure and catalyst concentration, that affect the response which is a smooth function of these variables (*see* ►Response Surface Methodology). At what combination of variables should measurements be taken in order to obtain good estimates of the dependence of responses, such as yield or purity of product, on these variables? More recent developments include the design of experiments for the nonlinear models occurring in pharmacokinetic experiments in drug development (Gagnon and Leonov 2005).

Perhaps after data transformation (*see* ►Box-Cox Transformations), efficient analysis of regression experiments requires the use of least squares parameter estimation (*see* ►Least Squares). In a good experiment the variances and covariances of the estimated parameters will be small. Optimal experimental designs minimize functions of these variances and so provide good estimates of the parameters.

Since optimal designs focus on the variances of the estimated parameters, it is necessary to specify a model or models. Also needed is an experimental region $\mathcal{X}$ that specifies the range of values of the experimental variables. The modern statistical theory of optimum experimental design was developed in a series of papers by Kiefer (Brown et al. 1985). A succinct introduction to the theory is given by Fedorov and Hackl (1997), which Atkinson et al. (2007) fleshes out with examples and SAS programs.

Least Squares and the Information Matrix

The linear regression model (see ►Linear Regression Models) is written $E(y) = F\beta$ with y the $N \times 1$ vector of responses, β a vector of p unknown parameters and F the $N \times p$ *extended* design matrix that may contain functions of the explanatory variables such as powers and interactions. The ith row of F is $f^T(x_i)$, a known function of the m explanatory variables. It is usual to assume that the observational errors are independent with constant variance σ^2. This value does not affect the design.

The least squares estimator of the parameters is $\hat{\beta}$ with information matrix F^TF. The "larger" F^TF, the greater is the information in the experiment. Depending on the scientific purpose, different optimality criteria are chosen which maximize different functions of the information matrix.

Criteria of Optimality

D-optimality. The volume of the confidence region for all p elements of β is proportional to the square root of $\sigma^2/|F^TF|$, the generalized variance of $\hat{\beta}$. Designs which maximize $|F^TF|$ minimize this generalized variance and are called D-optimum (for Determinant).

G-optimality. The prediction from the fitted model at a point x is $\hat{y}(x) = \hat{\beta}^T f(x)$, with variance var $\hat{y}(x)$. Designs which minimize the maximum over $\mathcal{X}$ of var $\hat{y}(x)$ are called G-optimum. A famous result, the "General Equivalence Theorem" (Kiefer and Wolfowitz 1960), shows that as N increases, D-optimality and G-optimality give increasingly similar designs. The designs may also be identical for some specific smaller values of N.

I or **V-optimality.** An alternative to this minimax approach to var $\hat{y}(x)$ is to find designs that minimize the average value of the variance over $\mathcal{X}$. Such designs are variously called I-optimum or V-optimum from Integrated Variance.

c-optimality. Another criterion of importance in applications, particularly for ▶nonlinear models, is that of C-OPTIMALITY in which the variance of the linear combination $c^T\hat{\beta}$ is minimized, where c is a $p \times 1$ vector of constants.

T-optimality. If there is uncertainty about the true model, T-optimality provides powerful designs for discriminating between models.

Details of these and other design criteria are in Chap. 10 of Atkinson et al. (2007). Compound designs in which provide good designs for several criteria are in Chap. 21. The optimum designs consist of a list of N values of x, often with replication. The indicated sets of conditions should be run in random order.

Some D-optimum Designs

Many standard designs which have been derived over the years by a variety of methods share the property that they are D- and G-optimum. One example is the series of 2^m factorial designs and their symmetric fractions forming the 2^{m-f} fractional factorial designs, which can also be used to construct D-optimally blocked 2^m designs. However, if N is not a power of 2, numerical methods for the construction of optimal designs will be needed. In the case of the blocked 2^m factorial, we might be forced by the size of batches of raw materials, to have blocks which were not all of size 2^{m-f}.

The dependence of the optimum design on the value of N is illustrated by D-optimum designs for the second-order response surface, that is a quadratic model in two factors including an interaction (x_1x_2) term. For $N = 9$, the optimum design is the 3^2 factorial. For $N = 13$, the points of the 2^2 factorial are added to this design. Increasing N to 14 duplicates the centre point. Addition of a few further design points destroys the symmetry of the designs and produces designs that have a lower efficiency per observation, although of course the overall information increases. The choice of design size as well as the values of the x_i are both important.

Advantages of Optimum Designs

Here are some advantages of optimum experimental design:

1. The availability of algorithms for the construction of designs
2. The calculation of good designs for a specified number of trials
3. The provision of simple, but incisive, methods for the comparison of designs
4. The ability to divide designs into blocks of a specified size
5. The determination of good response surface designs over non-regular regions
6. The construction of designs for ▶mixture models, perhaps again over non-regular regions

Further Models

The methods of optimum design, sketched here for linear models, can be extended to several important classes of model.

Nonlinear models. These models, nonlinear in the parameters, are particularly important in chemical and pharmaco-kinetics. Informative experiments are often concentrated in a small part of $\mathcal{X}$ and it is easy to waste experimental effort. Because of the non-linearity of the models, techniques incorporating prior information are important. See Chaps. 17 and 18 of Atkinson et al. (2007) and ▶Nonlinear Models.

Generalized Linear models. Particularly if the responses are binomial, the models need extending, although the principles of optimum design remain the same. See Chap. 22 of Atkinson et al. (2007) and ▶Designs for Generalized Linear Models.

Discrete Designs. In the designs appropriate in agriculture, many of the factors are discrete, rather than continuous as in regression, and analysis of variance models are appropriate (*see* ▶Analysis of Variance). The experimental units, for example plots in a field, may be highly variable and require division into blocks (*see* ▶Statistical Design of Experiments, ▶Agriculture, Statistics in and ▶Incomplete Block Designs). In addition to D-optimality, such designs are often compared using A-OPTIMALITY, in which the Average variance of treatment contrasts is minimized and E-OPTIMALITY in which the variance of the least well-estimated, or Extreme, linear combination $a^T\hat{\beta}$ is minimized subject to $a^Ta = 1$. Design construction is often facilitated by the use of combinatorial methods.

About the Author

Professor Atkinson has been elected a Fellow of the American Statistical Association "for outstanding contributions to data analysis methods, including variable selection in regression and multivariate data, and methods of optimum experimental design and for service to the profession." Professor Atkinson has served as editor of *The Journal of the Royal Statistical Society*, Series B and as associate editor of *Biometrika* and *Technometrics*. He has published five books and 200 articles in these and other journals including *Biometrics*, *The Journal of the American Statistical Association*, and *Statistics and Computing*.

Cross References

►Clinical Trials: An Overview
►Data Quality (Poor Quality Data: The Fly in the Data Analytics Ointment)
►Design of Experiments: A Pattern of Progress
►Designs for Generalized Linear Models
►Factorial Experiments
►Nonlinear Models
►Optimal Regression Design
►Randomization
►Response Surface Methodology
►Statistical Design of Experiments (DOE)
►Uniform Experimental Design

References and Further Reading

Atkinson AC, Donev AN, and Tobias RD (2007) Optimum experimental designs, with SAS. Oxford University Press, Oxford
Brown LD, Olkin I, Sacks J, Wynn HP (eds) (1985) Jack Carl Kiefer collected papers III. Wiley, New York
Fedorov VV, Hackl P (1997) Model-oriented design of experiments. Lecture Notes in Statistics 125. Springer, New York
Gagnon RC, Leonov SL (2005) Optimum population designs for PK models with serial sampling. J Biopharm Stat 15:143–163
Kiefer and Wolfowitz J (1960) The equivalence of two extremum problems. Canadian J Math 12:363–366
Smith K (1918) On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. Biometrika 12:1–85

Order Statistics

Herbert A. David
Distinguished Professor Emeritus
Iowa State University, Ames, IA, USA

Consider a *life test* in which n like items are tested until they fail. The failure times may be denoted by

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(r)} \le \ldots \le x_{(n)}.$$

Here $x_{(r)}$ is the *r-th order statistic*, also denoted by $x_{r:n}$ if the sample size needs to be emphasized. The duration of the test, $x_{(n)}$, may be unduly long, so that it becomes desirable to *censor* the test after, say, the r-th failure.

More commonly, the observations come to us unordered as $x_1, x_2, \ldots, x_n$, say the diameters of n mass-produced items. Here n is usually small, e.g., $n = 5$, but such samples are taken frequently and charts of both sample means and sample *ranges* $x_{(n)} - x_{(1)}$ plotted. The range is more convenient than the standard deviation and almost as efficient in small samples from the underlying normal populations generally assumed.

Basic Distribution Theory

Let $X_1, X_2, \ldots, X_n$ be independent random variables drawn from a population with cumulative distribution function (cdf) $F(x)$. Then the cdf of $X_{(r)}$ is given by

$$\begin{aligned} F_{X_{(r)}}(x) &= Pr(X_{(r)} \le x) \\ &= Pr\ (\text{at least } r \text{ of } X_1, \ldots, X_n \text{ are} \le x) \\ &= \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1-F(x)]^{n-i} \end{aligned} \quad (1)$$

since the term in the summand is the binomial probability that exactly i of $X_1, \ldots, X_n$ are less than or equal to x.

This result holds whether X is a continuous or discrete variate. In the former case differentiation gives the probability density function (pdf) of $X_{(r)}$ in terms of the underlying $f(x) = F'(x)$ as

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!} F^{r-1}(x) f(x) [1-F(x)]^{n-r}. \quad (2)$$

If X is discrete, taking integral values, we have simply

$$f_{X_{(r)}}(x) = Pr\left(X_{(r)} = x\right) = F_{X_{(r)}}(x) - F_{X_{(r)}}(x-1).$$

In the continuous case the joint pdf of $X_{(r)}$ and $X_{(s)}$ $(1 \le r < s \le n)$ is

$$\begin{aligned} f_{X_{(r)},\, X_{(s)}}(x,y) = {} & \frac{n!}{(r-1)!(s-r-1)!(n-s)!} F^{r-1}(x) f(x) \\ & \times [F(y) - F(x)]^{s-r-1} f(y) [1-F(y)]^{n-s}. \end{aligned} \quad (3)$$

The distribution of functions of two order statistics may now be obtained from (3) by standard transformation of variables methods. For example, the pdf of the range $W_n = X_{(n)} - X_{(1)}$ is

$$f_{W_n}(x) = n(n-1) \int_{-\infty}^{\infty} f(x) [F(x+w) - F(x)]^{n-2} f(x+w)\, dx,$$

giving after integration the useful formula

$$F_{W_n}(x) = n \int_{-\infty}^{\infty} f(x) [F(x+w) - F(x)]^{n-1}\, dx.$$

For a $N(\mu, \sigma^2)$ population, both $E(W_n/\sigma)$ and percentage points of W_n/σ have been widely tabulated (e.g., Pearson and Hartley 1970), providing the basis for ►control charts on μ and σ.

Confidence Intervals and Tolerance Intervals

The population qauntile of order p, denoted by ξ_p, is defined by $F(\xi_p) = p, 0 < p < 1$. In particular, $\xi_{\frac{1}{2}}$ is the

population median. Since, for $r < s$,

$$Pr\left(X_{(r)} < \xi_p < X_{(s)}\right) = Pr\left(X_{(r)} < \xi_p\right) - Pr\left(X_{(s)} < \xi_p\right)$$

it follows from (1) that the confidence interval $(X_{(r)},\ X_{(s)})$ covers ξ_p with probability

$$P_1 = \sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}.$$

We can choose r and s to make P_1 satisfactorily large.

A tolerance interval $(X_{(r)}, X_{(s)})$ covers a proportion γ of the underlying population with probability P_2. Thus

$$P_2 = Pr\left\{F\left(X_{(s)}\right) - F\left(X_{(r)}\right) \geq \gamma\right\}.$$

Since $F(X)$ is a uniform variate, U, over the interval $(0,1)$

$$P_2 = Pr\left(U_{(s)} - U_{(r)} \geq \gamma\right) = 1 - I_\gamma\left(n-r, n-s+r+1\right),$$

with the help of (3), where $I_\gamma(a,b)$ is the incomplete beta function.

Note that both intervals require only that the underlying distribution is absolutely continuous, i.e., these procedures are distribution-free or nonparametric. In this they are exceptional in the theory of order statistics, which is largely distribution-dependent. Order statistics are an aid to but not a branch of nonparametric statistics.

About the Author

Herbert Aron David was born in Berlin, Germany, on December 19, 1925. He earned a Ph.D. (1953) in statistics from University College London. He is Emeritus Distinguished Professor in Liberal Arts and Sciences at Iowa State University. He is a Fellow of the American Statistical Association, the American Association for the Advancement of Science and the Institute of Mathematical Statistics. He is an elected member of the International Statistical Institute. Professor David served as Department Head (1972–1984). He was Editor of *Biometrics* (1967–1972) and President of Biometric Society (1982–1983). He has authored/coauthored over 100 articles on order statistics, paired comparisons, experimental designs, competing risks, and the history of statistics. Professor David is author of *Order Statistics* (Wiley 1970, 3*rd* edition with H.N. Nagaraja, 2003).

"I am convinced that Order Statistics will continue to be the most valuable source of reference for students and researchers alike." (Paul Embrechts, September 2004, *Journal of the American Statistical Association*, Vol. 99, No. 467, p. 907).

Cross References

►Central Limit Theorems
►Normal Scores
►Normality Tests: Power Comparison
►Ordered Statistical Data: Recent Developments
►Permanents in Probability Theory
►Ranked Set Sampling
►Record Statistics
►Robust Statistical Methods

References and Further Reading

Arnold BC, Balakrishnan N, Nagaraja HN (1992) A first course in order statistics. Wiley, New York

David HA, Nagaraja HN (2003) Order statistics, 3rd edn. Wiley, Hoboken, NJ

Pearson ES, Hartley HO (eds) (1970/1972) Biometrika tables for statisticians, vols 1 and 2, 3rd edn. Cambridge University Press, Cambridge, UK

Ordered Statistical Data: Recent Developments

Mohammad Z. Raqab
Professor of Statistics, Acting Dean, Faculty of Science
University of Jordan, Amman, Jordan

The term "order statistics" (see ►Order Statistics) was introduced by Wilks in 1942. The history of the ordered variates goes back to older years. Throughout the last thirty years, order statistics and other related statistics have changed considerably and moved to be involved in statistical modeling, inferences, decision procedures and the study of the reliability systems.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution with probability density function (pdf) f and cumulative distribution function (cdf) F. If the random variables are arranged in ascending order of magnitude, then the ordered quantities $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ are called order statistics. Since there are $n!$ equally orderings of the x_i's, the joint pdf of $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ is

$$f_{1,2,\ldots,n}(x_1, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i), \quad x_1 \leq \ldots \leq x_n.$$

Let $F_{k:n}(x)(k = 1, \ldots, n)$ denote the cdf of the kth order statistic $X_{k:n}$. Its form can be derived as follows:

$$\begin{aligned} F_{k:n}(x) &= \mathbb{P}(X_{k:n} \le x) \\ &= \mathbb{P}(\text{at least k of the } X_i \text{ are less than or equal to x}) \\ &= \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j \, [1 - F(x)]^{n-j}. \end{aligned} \tag{1}$$

The binomial probability in the summand represents the probability that exactly k of the X_i's are less than or equal to x. From the relationship between the binomial sum of probabilities and the incomplete beta, we write the cdf of $X_{k:n}$ in (1) as

$$F_{k:n}(x) = \frac{n!}{(k-1)!(n-k)!} \int_0^{F(x)} u^{k-1}(1-u)^{n-k} \, du. \tag{2}$$

The expression (2) holds for continuous and discrete variates. If we assume that X_i is continuous with pdf $f(x) = F'(x)$, then the pdf of $X_{k:n}$ will have the form

$$f_{k:n}(x) = \frac{n!}{(k-1)!(n-k)!} [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x).$$

In the last few decades, the ordered statistics are naturally engaged in reliability theory. In many applications, a $(n - k + 1)$-out-of-n technical system of n identical components with independent and identically distributed (iid) life-lengths, functions if at least $(n - k + 1)$ of the components function. In this context, the life-length of the system is the same as the life-length of $X_{k:n}(1 \le k \le n)$. For $k = 1$, the system is a series system where any component failure is disastrous. When $k = n$, the system is also known as a parallel system. Statistical inference on the model of order statistics can be found in David and Nagaraja (2003). Recent years have been seen a rapid growth on techniques and applications involving order statistics. Its connection with reliability theory, survival studies and biostatistical sciences is the main reason for its wide applications.

Kaminsky and Rhodin (1985) and Raqab and Nagaraja (1995) have examined the maximum likelihood prediction (MLP) of the future failure time $X_{s:n}$ based on the observed failure times $X_{1:n}, X_{2:n}, \ldots, X_{r:n}, 1 \le r < s \le n$. The corresponding prediction likelihood function (PLF) is

$$L \propto \prod_{i=1}^{r} f(x_{i:n}) [F(x_{s:n}) - F(x_{r:n})]^{s-r-1} [1 - F(x_{s:n})]^{n-s} f(x_{s:n}).$$

It is observed that the MLP is biased in general and it may lead to a smaller mean square error when compared with best linear unbaised predictor (Kaminsky and Nelson 1975). Raqab (1997) has approximated the PLF and obtained some approximate MLPs. It is found that the approximate MLPs perform well in some cases compared to the MLPs.

Applications of the ordered statistical data can be observed in the analysis of the two most common censoring schemes termed as Type-I and Type-II censoring. Let us assume that n items are placed on a life-testing experiment. In a Type-I censoring scheme, the experiment continues up to a specified time T and the failures occurring after T are not observed. The Type-II scheme requires the experiment to continue until a pre-specified number of failures $m \le n$ occurs. If a practitioner desires to remove experimental units at points other than the terminal point of the experiment, then a generalization of the above mentioned schemes can be put forward in a similar manner. This generalization is referred to as progressive Type-II censoring. Under this general censoring scheme, n units are placed on a life-testing experiment and only $m(< n)$ are completely observed until failure. The censoring occurs progressively in m stages. These m stages offer failure times of the m completely observed units. At the time of the first failure (the first stage), r_1 of the $n - 1$ surviving units are randomly withdrawn (censored) from the experiment, r_2 of the $n - 2 - r_1$ surviving units are withdrawn (censored) at the time of the second failure (the second stage), and so on. Finally, at the time of the *mth* failure (the *mth* stage), all the remaining $r_m = n - m - r_1 - \ldots - r_{m-1}$ surviving units are withdrawn. This scheme includes the conventional Type II right censoring scheme $(r_1 = r_2 = \ldots = r_{m-1} = 0, r_m = n - m)$ and the complete sampling scheme $(r_1 = r_2 = \ldots = r_{m-1} = 0, n = m)$. In the approximately ten years since the publication of the Progressive Censoring, Theory, Methods and Applications (Balakrishnan and Aggrawala 2000), various optimal techniques of estimation have appeared in the literature (Balakrishnan 2007).

Predictions of order statistics are of natural interest. Prediction problems can be generally classified into two types. In the first type, the variable to be predicted comes from the same sequence of variables observed and is therefore correlated with the observed data. This is referred to as the *one-sample prediction problem*. In the second type, referred to as the *two-sample prediction problem*, the variable to be predicted comes from another independent future sample. Considerable work has been done on the one-sample prediction problem, and both parametric and nonparametric inferential methods have been developed in this regard. Interested readers may refer to Gulati and Padgett (2003) for details on these developments. In contrast, the two-sample prediction problem has not received much attention.

Let $Y_{1:m:n}, Y_{2:m:n}, \ldots, Y_{m:m:n}$, denote the above mentioned progressively type II right censored observed sample. It is of interest to predict the life-lengths $Z_{s:r_i}$

$(s = 1, 2, \ldots, r_i;\ i = 1, 2, \ldots, m)$ of all censored units in all m stages of censoring. Here $Z_{s:r_i}$ $(s = 1, 2, \ldots, r_i;\ i = 1, 2, \ldots, m)$ denotes the sth order statistic from a sample of size r_i removed at stage $i = 1, 2, \ldots, m$. Basak et al. (2006) used different classical prediction methods such as best linear unbiased, maximum likelihood and conditional median to predict the times to failure $Z_{s:r_i}$ $(s = 1, 2, \ldots, r_i;\ i = 1, 2, \ldots, m)$. The prediction of times to failure of the last r_m units still surviving at the observation Y_m has been discussed by Balakrishnan and Rao (1997). Madi and Raqab (2009) have discussed the prediction of unobserved failure times from the generalized exponential (GE) distribution using the Gibbs and Metropolis samplers.

Other related statistics are the record values, which are introduced by Chandler (1952). Record statistics arise naturally in many practical problems, and there are several situations pertaining to meteorology, hydrology, sporting and athletic events wherein only record values may be recorded. Let $\{X_i, i \geq 1\}$ be a sequence of iid random variables with common absolutely continuous cdf F and pdf f. Define a sequence of record times $U(n), n = 1, \ldots,$ as follows:

$$U(1) = 1,$$

$$U(n) = \min\left\{j : j > U(n-1),\ X_j > X_{U(n-1)}\right\}, \quad n \geq 2.$$

Then the r.v.'s $X_{U(n)}, n \geq 0$ are called upper records. That is; an observation X_j is an upper record value (or simply a record) if its value exceeds that of all previous observations. The joint pdf of $X_{U(1)}, \ldots, X_{U(n)}$ is

$$f_{U(1),\ldots,U(n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n-1} \frac{f(x_i)}{1-F(x_i)} f(x_n), \quad -\infty < x_1 < \ldots < x_n < \infty. \tag{3}$$

By integrating the expression in (3) out of $x_1, \ldots, x_{n-1}$, we obtain the pdf of the nth upper record value as

$$f_{U(n)}(x) = \frac{[-log(1-F(x))]^{n-1}}{(n-1)!} f(x), \quad -\infty < x < \infty.$$

It is clear that the upper record values are the largest values observed to date. If the second or third largest values are observed to date then the model of kth record values is adequate when k is a positive integer (cf. Dziubdziela and Kopociński 1976). Record values are closely connected with the occurrence times of some corresponding non-homogeneous Poisson processes (see ►Poisson Processes) and used in so-called shock models. The successive shocks may be considered as realizations of record values from a sequence of identically independent random variables. In the context of statistical inference, maximum likelihood, best linear unbiased and best linear invariant are used to estimate the model parameters based on record data (Arnold et al. 1998). A Bayesian parametric approach via the Gibbs and Metropolis samplers is also used to predict the behavior of further future records (Madi and Raqab 2007).

The order statistics and record values are included in a more general model called generalized order statistics (cf. Kamps 1995). The concept of generalized order statistics allows us to unify the models and examine the similarities and analogies. Through the last twenty years, it was observed that many well-known distributional properties of order statistics and record values are also valid for generalized order statistics. The best linear unbiased estimation is applied by Burkschat et al. (2007) to estimate location and scale parameters of generalized Pareto distribution based on generalized order statistics.

About the Author

Dr. Mohammad Raqab is Professor of Statistics and Acting Dean, Faculty of Science, University of Jordan. He completed his B.Sc. (1981) in mathematics at University of Jordan, M. Sc. (1989) and Ph.D. (1992) in statistics at Ohio State University, Ohio State, USA. He has received five Scientific Awards, as a distinguished researcher in statistics. Raqab is an elected member of the International Statistical Institute (2007), a Fellow of American Association for the Advancement of Science (2003–present), New York Academy of Sciences (1998–present), American Statistical Association (1992–present), and International Biometric Society, USA (2000). Professor Raqab is currently an Associate Editor of the *Journal of Applied Statistical Science* (2000–present), *Journal of Statistical Theory and Applications* (2002–present), and *Journal of Probability and Statistical Science* (2003–present). He has (co-)authored more than 80 refereed articles and five books including, *Recent Development in Ordered Random Variables* (with M. Ahsanullah, Nova Science, New York, USA, 2007).

Cross References

►Best Linear Unbiased Estimation in Linear Models
►Binomial Distribution
►Censoring Methodology
►Order Statistics
►Parametric and Nonparametric Reliability Analysis
►Permanents in Probability Theory
►Ranked Set Sampling
►Record Statistics

References and Further Reading

Arnold BC, Balakrishnan N, Nagaraja HN (1998) *Records*. Wiley, New York

Balakrishnan N (2007) Progressive methodology: an appraisal, (with discussions). Test 16(2):211–259

Balakrishnan N, Aggrawala R (2000) Progressive censoring, theory, methods and applications. Birkhäuser, Boston
Basak I, Basak P, Balakrishnan N (2006) On some predictors of times to failure of censored items in progressively censored samples. Comput Stat Data Anal 50:1313–1337
Balakrishnan N, Rao CR (1997) Large sample approximations to best linear unbiased estimation and best linear unbiased prediction based on progressively censored samples and some applications. In: Panchapakesan S, Balakrishnan N (eds) Advances in statistical decision theory and applications. Birkhäuser, Boston, pp 431–444
Burkschat M, Cramer E, Kamps U (2007) Linear estimation of location and scale parameters based on generalized order statistics from generalized pareto distribution. In: Ahsanullah M, Raqab MZ (eds) Recent developments in ordered random variables. Nova Science, New York, pp 253–261
Chandler KN (1952) The distribution and frequency of record values. J R Stat Soc B 14:220–228
David HA, Nagaraja HN (2003) Order statistics. Wiley, New York
Dziubdziela W, Kopociński B (1976) Limiting properties of the k-th record values. Appl Math (Warsaw) 15:187–190
Gulati S, Padgett WJ (2003) Parametric and nonparametric inference from record-breaking data. Springer, New York
Kaminsky KS, Nelson PI (1975) Best linear unbiased prediction of order statistics in location and scale families. J Am Stat Assoc 70:145–150
Kaminsky KS, Rhodin LS (1985) Maximum likelihood prediction. Ann Inst Stat Math 37:507–517
Kamps U (1995) A concept of generalized order statistics. J Stat Plann Infer 48:1–23
Madi MT, Raqab MZ (2007) Bayesian Prediction of Rainfall Records Using the Generalized Exponential Distribution. Environmetrics 18:541–549
Madi MT, Raqab MZ (2009) Bayesian Inference for the generalized exponential distribution based on progressively censored data. Commun Stat-Theor M 38(12):2016–2029
Raqab MZ (1997) Modified maximum likelihood predictors of future order statistics from normal samples. Comput Stat Data Anal 25:91–106
Raqab MZ, Nagaraja HN (1995) On some predictors of future order statistics. Metron 53:185–204
Wilks SS (1942) Statistical prediction with special reference to the problem of tolerance limits. Ann Math Stat 13:400–409

Outliers

Toby Lewis
Professor
University of East Anglia, Norwich, UK

A familiar problem in analyzing data is the occurrence of one or more values in a data set which appear to the analyst, (who thus makes an individual subjective judgment, see Collett and Lewis (1976)), to be inconsistent or out of line with the rest of the data, relative to the probability model – call it F – assumed for the data. Such a surprising value is called an *outlying observation* or *outlier*. The term 'outlier' also applies to a value in the data which would have appeared surprising if the analyst had observed it. The damaging effect of such an unobserved outlier is illustrated by Ho and Naugher (2000); see the discussion of *Accommodation* below.

In what follows, some references are given to particular topics, but the reader is otherwise referred to the comprehensive treatment of outliers and outlier problems in Barnett and Lewis (1994).

Outliers are encountered in a variety of situations, e.g., an outlying value or a subset of 2 or more outlying values (the multiple outlier situation) in a univariate sample, an outlying point (x, y) in the regression of a variable y on a regressor variable x, an outlying value in a time series or in a contingency table, and so on. The general principles and the range of procedures for dealing with outliers are first discussed in the case of an upper outlier in a univariate sample of n observations $x_1, x_2, \ldots, x_n$ of a variable X, denoted in ascending order $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$. If the greatest observation $x_{(n)}$ is not only higher than the rest of the sample but appears surprisingly high, it is an outlier, relative to the assumed model F for the data. If an appropriate statistical test is carried out of the null hypothesis H_0 that all n values belong to F against an alternative H_1 that $n-1$ values belong to F and $x_{(n)}$ does not, and H_0 is rejected, the value $x_{(n)}$ is judged to be not credible (at the level of the test) as the highest value in the sample of size n from F. This value is then said to be *discordant*; the *discordancy test* has shown it to be a *discordant outlier*. The value $x_{(n)}$ is taken not to come from the distribution modeled, and it is conventionally called a *contaminant*.

"This apparently pejorative term . . . does not necessarily have any undesirable implications, sometimes quite the contrary; for instance, in a study of performance in examinations a phenomenally high mark by a student of exceptional ability might be called a 'contaminant'!" (Langford and Lewis 1998).

Examples of a discordancy test statistic for an upper outlier $x_{(n)}$ in a univariate sample of size n are $(x_{(n)}/\bar{x})/s$ for an assumed normal sample and $x_{(n)}/n\bar{x}$ for an exponential sample.

Note that some writers use the word "outlier" for a discordant outlier, and refer to an outlier (as said above, a value which looks unusual but may or may not be discordant) as a "suspicious" or "suspect" value.

What action is called for with an outlier? In some cases the value is deleted from the data set without the need for statistical assessment, e.g., because it is known to have been misrecorded or miscalculated. Again there are cases when a discordancy test is carried out on an outlier, say the greatest value $x_{(n)}$ in a univariate sample, the null hypothesis

H_0 is not rejected and $x_{(n)}$ is declared non-discordant. All the n sample values are then taken to belong to F, and the outlier is retained in the data set although it has appeared to the analyst to be somewhat unusual.

When, on the other hand, H_0 is rejected and the outlier is declared discordant on the assumption of model F, three possibilities arise, depending on the aim of the investigation. If its concern is with characteristics of the population that has been sampled, e.g., to obtain a confidence interval for the population mean, the discordant outlier, judged not to belong to F, should be rejected from the data set. (A valuable alternative strategy, to use a robust method of estimating the population parameter of interest and dispense with testing the outlier for discordancy, is discussed below – see ►Accommodation.) If on the other hand the concern of the investigation is with the outlying value ($x_{(n)}$, say) and the information conveyed by it (showing, for example, that a duration of pregnancy as long as $x_{(n)}$ years has been known to occur), further examination of this information and its implications would naturally follow.

These two possibilities assume acceptance of the model F, and conclude that the outlying value is discordant. There is a third possibility, that the value belongs to the same distribution as the rest of the data, but that the appropriate model for this distribution is not F but some other distribution G. A fresh analysis is then called for with a suitable choice of model G (e.g., a gamma distribution G with unknown shape parameter instead of an exponential distribution F), on the basis of which the discordant outlier relative to F is a non-discordant member of the data set, and the outlier problem disappears.

The *multiple outlier* situation is now considered. The choice of procedures for dealing with data sets with two or more outliers is first discussed in the context of two upper outliers $x_{(n-1)}, x_{(n)}$, in a univariate sample of n observations $x_1, \ldots, x_n$, of a continuous variable X. In testing the two outliers for discordancy the analyst has a three-way choice between a *block procedure*, an *inward consecutive procedure,* and an *outward consecutive procedure.*

Block procedure: $x_{(n-1)}$ and $x_{(n)}$ are tested together as an outlying pair in the data set of n values, with alternative hypothesis that $x_{(1)}, \ldots, x_{(n-2)}$. belong to F while $x_{(n-1)}$ and $x_{(n)}$ do not. There are two possible outcomes: (1) $x_{(n)}$ and $x_{(n-1)}$ are both judged non-discordant, (2) they are both judged discordant.

Inward consecutive procedure: first, $x_{(n)}$ is tested as an outlier in the complete data set of n values. If $x_{(n)}$ is judged discordant, $x_{(n-1)}$ is then tested as an outlier in the data set of $n-1$ values $x_{(1)}, \ldots, x_{(n-1)}$. There are three possible outcomes: (l) $x_{(n)}$ and therefore also $x_{(n-1)}$ are judged non-discordant, (2) $x_{(n)}$ is judged discordant, $x_{(n-1)}$ non-discordant, (3) $x_{(n)}$ and $x_{(n-1)}$ are both judged discordant.

Outward consecutive procedure: first, $x_{(n-1)}$ is tested as an outlier in the data set of $n-1$ values $x_{(1)}, \ldots, x_{(n-1)}$. If $x_{(n-1)}$ is judged non-discordant, $x_{(n)}$ is then tested as an outlier in the complete data set of n values. There are clearly three possible outcomes: (1) $x_{(n-1)}$ and therefore also $x_{(n)}$ are judged discordant, (2) $x_{(n-1)}$ is judged non-discordant, $x_{(n)}$ discordant, (3) $x_{(n-1)}$ and $x_{(n)}$ are both judged non-discordant.

Block testing may well be appropriate when the outliers form, from non-statistical characteristics, a natural subset of the data, for example when some measure of an individual's performance is being studied in an analysis of a random sample of individuals, two of whom happen to be siblings. However, a possible danger to the use of a block test arises when *swamping* occurs. This is when a non-discordant value $x_{(n-1)}$ is much nearer to $x_{(n-2)}$ than to $x_{(n)}$, but when $x_{(n)}$ is extreme enough to make $x_{(n)}$ and $x_{(n-1)}$ jointly declared discordant in a block test. $x_{(n-1)}$ is then wrongly judged, having been *swamped* by $x_{(n)}$ in the block test.

Conversely, a possible danger to the use of an *inward consecutive procedure* arises when *masking* occurs. The procedure starts with the test for discordancy of an outlier $x_{(n)}$ which is in fact discordant. If $x_{(n-1)}$ is near in value to $x_{(n)}$ its presence in the test may cause the outlier $x_{(n)}$ to be judged non-discordant. $x_{(n)}$ is then wrongly judged, having been *masked* by $x_{(n-1)}$. In an outward consecutive procedure, masking does not occur.

Accommodation

In analyzing a data set, say a sample of values of a variable X, with the aim of making inferences about the population being sampled, the approach described so far for dealing with any outlier in the data set has been to test it for discordancy and remove it from the data set if it is judged discordant. As already mentioned, an alternative approach is to dispense with discordancy testing and to *accommodate* the outlier by giving extreme values in the data set reduced weight in the analysis whether they are discordant or not; in short, to use a procedure *robust against the presence of outliers*. For example, a well known robust procedure for accommodating possible outliers in a sample $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ when estimating the population mean $E(X)$ is *trimming* – working with a "trimmed sample" from which the r lowest and s highest values, with an appropriate choice of r and s, have been

removed. This replaces the sample mean $\bar{x}$ by the (r,s)-fold trimmed mean $(x_{(r+1)} + \cdots + x_{(n-s)})/(n-r-s)$ when estimating $E(X)$.

Clearly, an important advantage of accommodation as against discordancy testing is the protection it gives against the danger of unobserved outliers Ho and Naugher (2000).

There is a range of relevant robust procedures available in the literature; see Huber (1981), Rousseeuw and Leroy (1987).

Directional data: The mean of a number of directions is not given by the simple arithmetic average, so different test and accommodation procedures are needed for outliers in data sets where the observations are directions of occurrences in two or in three dimensions, represented respectively by points on the circumference of a circle or on the surface of a sphere. Examples of "circular data" are vanishing directions of homing pigeons and arrival times on a 24-hour clock of patients at a hospital's accident and emergency department; examples of "spherical data" are arrival directions of showers of cosmic rays and wind directions. See Fisher (1993) and Fisher et al. (1987) for treatment of directional outliers.

The discussion has, for convenience, mainly been in the context of outliers in univariate samples; however, as stated earlier, outliers are encountered in a variety of more complex situations.

Much work has been done on the treatment of outliers in the following data situations:
Multivariate data, Regression, Designed experiments, Contingency tables, Time series, and some work also on Multilevel data and Sample surveys.

For discussion of outliers in these data situations [except Multilevel data – but see Langford and Lewis (1998)] see Barnett and Lewis (1994).

About the Author

Toby Lewis was Honorary Professor at the University of East Anglia from 1986 and Director of the Centre for Statistics there from 1989, both to 2010. He was previously Professor of Statistics at the Open University (1980–1986) and Professor of Mathematical Statistics at the University of Hull (1968–1979), at each university as first occupant of a newly created chair. He is a Fellow of the Royal Statistical Society, a Chartered Statistician, and a member of the International Statistical Institute. He is the co-author (with V. Barnett) of the first comprehensive text on outlier methods in statistical analysis: *Outliers in Statistical Data*, Wiley (1st edition 1978; 3rd edition 1994). He is 91 and is still-active professor. "I leave it to others to work out exactly how much of an outlier that will make him in the world of statisticians" (Julian Champkin (2008). Toby Lewis. *Significance* **5**: p. 132).

Cross References

►Adaptive Methods
►Cook's Distance
►Detecting Outliers in Time Series Using Simulation
►Estimation
►Exploratory Data Analysis
►Intervention Analysis in Time Series
►Median Filters and Extensions
►Misuse of Statistics
►Multivariate Outliers
►Multivariate Technique: Robustness
►Nonparametric Rank Tests
►Optimality and Robustness in Statistical Forecasting
►Preprocessing in Data Mining
►Regression Diagnostics
►Residuals
►Robust Statistical Methods
►Robust Statistics
►Summarizing Data with Boxplots

References and Further Reading

Barnett V, Lewis T (1994) Outliers in statistical data, 3rd edn. Wiley, Chichester
Collett D, Lewis T (1976) The subjective nature of outlier rejection procedures. Appl Stat 25:228–237
Fisher NI (1993) Statistical analysis of circular date. Cambridge University Press, Cambridge
Fisher NI, Lewis T, Embleton BJJ (1987, paperback edn 1993). Statistical analysis of spherical data. Cambridge University Press, Cambridge
Ho K, Naugher JR (2000) Outlier lies: an illustrative example of identifying outliers and applying robust models. Multiple Linear Regression Viewpoints 26(2):2–6
Huber PJ (1981) Robust Statistics. Wiley, New York
Langford IH, Lewis T (1998) Outliers in multilevel data. J R Stat Soc A 161:121–160
Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York

P

Panel Data

Markku Rahiala
Professor (emeritus) of Econometrics
University of Oulu, Oulu, Finland

Panel data (or longitudinal data) are data sets, where information on a number of observational units have been collected at several time points. These observational units are usually called *individuals* or *subjects*. In economic applications they might be households, firms, individual persons, countries, investors or other economic agents. In medical, biological and social applications the subjects might be patients, test animals, individual persons etc.

Panel data are most often used to study unidirectional relationships between some explanatory variables $X_{it} = (x_{1,i,t} \ \ldots \ x_{m,i,t})'$ and a continuous dependent variable y_{it}, where i ($i = 1, \ldots, N$) refers to individual and t ($t = t_{oi}, \ldots, T_i$) refers to time or period. (To simplify notation, we will assume $t_{oi} \equiv 1$ and $T_i \equiv T$.) Panel data have many advantages over cross sectional data or single aggregated time series. One can for instance study the dynamic effects of the covariates on a micro level and one can explore heterogeneities among the individuals. Regression models of the linear type (for suitably transformed variables)

$$y_{it} = \alpha_i + \beta' X_{it} + \varepsilon_{it} \tag{1}$$

are especially popular as model frameworks. The error terms ε_{it} are assumed to be independent of the explanatory variables X_{it}, the processes $\{\varepsilon_{it}\}$ are assumed stationary with zero means and the data from different individuals are assumed independent of each other. The length of the data T is often fairly short, whereas the number of subjects N might be large. The levels of any dependent variable usually show considerable individual variation, which makes it necessary to allow for individually varying intercept terms α_i in model (1). If these intercepts are taken as fixed parameters, the model is called a *fixed effects* model. When N is large, this will however lead to some inferential problems. For instance ML estimators of the variance–covariance structure of the error processes would be inconsistent for fixed T and increasing N. This is why the intercept terms are often treated as mutually independent random variables $\alpha_i \sim \mathrm{IID}(\alpha, \tau^2)$ (α_i also independent of $\{\varepsilon_{it}\}$), whenever the observed subjects can be interpreted as a sample from a larger population of potential subjects. These *random effects* models are special cases of so-called *mixed* models or *variance components* models. If the effects of the covariates X_{it} vary over individuals, the regression coefficients $\beta_{(i)}$ can similarly be interpreted as random variables, $\begin{pmatrix}\alpha_i & \beta_{(i)}'\end{pmatrix}' \sim \mathrm{IID}_{m+1}\left(\begin{pmatrix}\alpha & \beta'\end{pmatrix}', G\right)$. If all the random elements in the model were assumed normally distributed, the whole model for subject i could be written as

$$\begin{aligned} Y_i &= (y_{i1} \ \ldots \ y_{iT})' \\ &= \alpha + X_i\beta + Z_i U_i + \varepsilon_i, \quad \varepsilon_i \text{ independent of } U_i, \\ U_i &\sim \mathrm{NID}_{m+1}(0, G), \quad \varepsilon_i = (\varepsilon_{i1} \ \ldots \ \varepsilon_{iT})' \nu \\ &\sim \mathrm{NID}_T(0, R), \end{aligned} \tag{2}$$

where $U_i = \begin{pmatrix}\alpha_i & \beta_{(i)}'\end{pmatrix}' - \begin{pmatrix}\alpha & \beta'\end{pmatrix}'$, $X_i = (X_{i1} \ \ldots \ X_{iT})'$, $Z_i = (\mathbf{1}_T \quad X_i)$ and $\mathbf{1}_T = (1 \ \ldots \ 1)'$. This is a standard form of a mixed model leading to GLS estimators for α and β once the parameters incorporated in the matrices G and R have been estimated.

To take account of possible unobserved changes in the general (either economic or biological) environment, one can include an additive term $\gamma_{(i)}' F_t$ in model (1), where F_t denotes an r-dimensional vector of common factors and $\gamma_{(i)}$ is a vector containing the factor loadings for individual i. These common factors will induce dependencies between the y_{it}-observations from different individuals. (See e.g., Pesaran 2006.)

In model (1), all the explanatory variables were assumed *exogenous*. However, econometric models quite often include also lagged values of the dependent variable as regressors. Much of economic theory starts from the assumption that the economic agents are optimizing an *intertemporal* utility function, and this assumption often induces an autoregressive, dynamic model

$$\begin{aligned} y_{it} &= \alpha_i + \phi_1 y_{i,t-1} + \ldots + \phi_p y_{i,t-p} + \beta' X_{it} + \varepsilon_{it}, \\ \varepsilon_{it} &\sim \mathrm{IID}(0, \sigma^2) \end{aligned} \tag{3}$$

for the dependent variable y_{it}. In case of random intercepts α_i, the lagged values of the dependent variable and

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

the combined error terms $(\alpha_i - \alpha) + \varepsilon_{it}$ will be correlated. This would lead to inconsistent least squares estimators for the ϕ-parameters. The problem can be circumvented by the GMM estimation method (Generalized Method of Moments). (See e.g., the Appendices in Arellano 2003.) Once the order p of the autoregressive model (3) has been correctly specified, it will be easy to find valid instruments for the GMM estimationt among the lagged differences of the dependent variable. Model (3) can be straightforwardly extended to the vector-valued case. (See e.g., Hsiao 2003, Chap. 4.7.)

If the dependent variable y_{it} is *discrete* (either a count variable or measured on a nominal or ordinal scale), one can combine the basic idea of model (1) and the concept of ►*generalized linear models* by assuming that the covariates X_{it} and the heterogeneity terms α_i affect the so-called *linear predictors* analogously to (1),

$$\eta_{it} = g\left(E\left(y_{it} \mid X_{it}, \alpha_i\right)\right) = \alpha_i + \beta' X_{it} \qquad (4)$$

and by assuming that conditionally on X_{it} and α_i , y_{it} follows a distribution belonging to the exponential family of distributions.1 (See e.g. Diggle et al. 2001, Chap. 11, or Fitzmaurice et al. 2004, Chap. 12.) Function g is called the *link* function, and models (4) are called *generalized linear mixed models* (GLMM). If for instance y_{it} would be dichotomous obtaining the values 0 or 1, *logit* link function would lead to the model

$$P\left(y_{it} = 1 \mid X_{it}, \alpha_i\right) = \exp(\alpha_i + \beta' X_{it}) / \left(1 + \exp(\alpha_i + \beta' X_{it})\right).$$

The resulting likelihood function contains complicated integral expressions, but they can be effectively approximated by numerical techniques.

About the Author

Markku Rahiala received his Ph.D. in statistics in 1985 at the University of Helsinki. He is Professor of econometrics, University of Oulu since 1998. He was Associate editor of the *Scandinavian Journal of Statistics* (1989–1995). He is Member of the Econometric Society since 1985.

Cross References

►Data Analysis
►Event History Analysis
►Linear Mixed Models
►Medical Statistics
►Multilevel Analysis
►Nonsampling Errors in Surveys
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Repeated Measures
►Sample Survey Methods
►Social Network Analysis
►Statistical Analysis of Longitudinal and Correlated Data
►Testing Variance Components in Mixed Linear Models

References and Further Reading

Arellano M (2003) Panel data econometrics. Oxford University Press, Oxford

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL (2001) Analysis of longitudinal data, 2nd edn. Oxford University Press, Oxford

Fitzmaurice GM, Laird NM, Ware JH (2004) Applied longitudinal analysis. Wiley, Hoboken

Hsiao C (2003) Analysis of panel data, 2nd edn. Cambridge University Press, Cambridge

Pesaran MH (2006) Estimation and inference in large heterogeneous panels with multifactor error structure. Econometrica 74: 967–1012

Parametric and Nonparametric Reliability Analysis

Chris P. Tsokos
Distinguished University Professor
University of South Florida, Tampa, FL, USA

Introduction

Reliability is "the probability that a piece of equipment (component, subsystem or system) successfully performs its intended function for a given period of time under specified (design) conditions" (Martz and Waller 1982). Failure means that an item does not perform its required functions. To evaluate the performance of an item, to predict its failure time and to find its failure pattern is the subject of Reliability.

Mathematically we can define reliability, $R(t)$, as follows:

$$R(t) = \Pr(T > t) = 1 - \int_0^t f(\tau)\, d\tau, \quad (t \geq 0)$$

where T denotes the failure time of the system or component, and $f(t)$ the failure probability distribution.

The main entity in performing accurate reliability analysis depends on having properly identified a classical discrete or continuous probability distribution that will characterize the behavior of the failure data. In practice, scientists and engineers either assume one, such as the exponential, Weibull, Poisson, etc., or a perform goodness of fit test to properly identify the failure distribution and then proceed with the reliability analysis. It is possible that the assumed failure distribution is not the correct one and

furthermore, the goodness of fit test methodology failed to identify a classical probability distribution. Thus, proceeding with the reliability analysis will result in misleading and incorrect results.

In this brief document we discuss a nonparametric reliability procedure when one cannot identify a classical failure distribution, $f(t)$, to characterize the failure data of the system. The method is based on estimating the failure density through the concept of distribution-free kernel density method. Utilizing such methods on the subject area offers significant computation difficulties. Therefore, in order to use this method, one must be able to obtain the optimal bandwidth for the kernel density estimate. Here, we recommend a six-step procedure which one can apply to compute the optimal nonparametric probability distribution that characterizes the failure times. Some useful references on the subject matter are Bean and Tsokos (1980, 1982), Liu and Tsokos (2001, 2002a, b), Qiao and Tsokos (1994, 1995), Rust and Tsokos (1981), Silverman (1986), and Tsokos and Rust (1980). First we briefly discuss the parametric approach to reliability using the popular three-parameter Weibull probability distribution as the failure model. Some additional useful failure models can be found in Tsokos (1998, 1995).

Parametric Approach to Reliability

The two-parameter Weibull probability distribution is universally used to characterize the failure times of a system or component to study its reliability behavior instead of the three-parameter Weibull model. Recently, methods have been developed along with effective algorithms for which one can obtain estimates of the three-parameter Weibull probability distribution. Here we will use the three-parameter Weibull failure model.

The three-parameter Weibull failure model is given by

$$\hat{f}(t) = \frac{\hat{c}\,(t-\hat{a})^{\hat{c}-1}}{\hat{b}^{\hat{c}}} \exp\left\{-\left(\frac{t-\hat{a}}{\hat{b}}\right)^{\hat{c}}\right\},$$

where $\hat{a}$, $\hat{b}$ and $\hat{c}$ are the maximum likelihood estimates to the location, scale and shape parameters, respectively. For calculation of the estimates of a, b and c, see Qiao and Tsokos (1994, 1995).

Thus, the parametric reliability estimation $\hat{R}_p(t)$ of the three-parameter Weibull model is given by

$$\hat{R}_p(t) = \exp\left\{-\left(\frac{t-\hat{a}}{\hat{b}}\right)^{\hat{c}}\right\}.$$

The goodness of fit criteria that one can use in identifying the appropriate classical failure probability distribution to characterize the failure times is the popular ►Kolmogorov–Smirnov test.

Briefly, it tests the null hypothesis that the data $\{t_j\}_{j=1}^n$ is from some specified classical probability distribution against the alternative hypothesis that it is from another probability distribution. That is,

$$\begin{cases} H_0 : \{t_j\}_{j=1}^n \sim F(t), \\ H_1 : \{t_j\}_{j=1}^n \not\sim F(t). \end{cases}$$

Let $F_n^*(t)$ be the empirical distribution function for the failure data. The Kolmogorov–Smirnov statistic is defined by

$$D_n = \sum_t \left|F_n^*(t) - F(t)\right|.$$

The statistic D_n can be easily calculated from the following formula:

$$D_n = \max\left\{\max_{1\le i\le n}\left[\frac{i}{n} - F\left(t_{(i)}\right)\right], \max_{1\le i\le n}\left[\frac{i-1}{n} - F\left(t_{(i)}\right)\right]\right\},$$

where $t_{(1)} \le t_{(2)} \le \cdots \le t_{(n)}$ is the order statistic of $\{t_j\}_{j=1}^n$.

Let $D_{n,\alpha}$ be the upper α-percent point of the distribution of D_n, that is,

$$P\{D_n > D_{n,\alpha}\} \le \alpha.$$

Tables for the exact critical values $D_{n,\alpha}$ are available. See Miller (1956) and Owen (1962), among others, to make the appropriate decision.

Nonparametric Approach to Reliability

Let $\{t_j\}_{j=1}^n$ be the failure data characterized by the probability density function $f(t)$. Then the nonparametric probability density estimation $\hat{f}$ can be written as

$$\hat{f}_{\hat{h}}(t) = \frac{1}{n\hat{h}} \sum_{j=1}^{n} K\left(\frac{t-t_j}{\hat{h}}\right)$$

where $K(t)$ is the kernel and assumed to be Gaussian given by

$$K(t) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}t^2}$$

and $\hat{h}$ is the estimate of the optimal bandwidth. There are several other choices for the kernel, namely, Epanechnikov, Corine, biweight, triweight, triangle and uniform. For further information regarding the selection process, see Silverman (1986). The most important element in $\hat{f}_{\hat{h}}(t)$ being effective to characterize the failure data is the bandwidth, $\hat{h}$. Given below is a procedure that works fairly well in obtaining optimal $\hat{h}$ and then $\hat{f}_{\hat{h}}(t)$ for the failure data,

(Bean and Tsokos 1982; Silverman 1986). This procedure is summarized below.

- Calculate S^2 and T_4 using the failure data $t_1, t_2, \ldots, t_n$:

$$S^2 = (n-1)^{-1} \sum_{j=1}^{n} (t_j - \bar{t})^2 \text{ and } T_4 = \frac{1}{n} \sum_{j=1}^{n} (t_j - \bar{t})^4.$$

- Determine a value for U_2 and U_4 as defined below:

$$U_2 \approx S^2 \text{ and } U_4 \approx \frac{n^3}{(n-1)(n^2-2n+3)} \left(T_4 - \frac{3(n-1)(2n-3)}{n^3} S^2 \right).$$

- Find estimates of the parameters μ and σ:

$$\hat{\mu} = \sqrt[4]{\frac{3U_2^2 - U_4}{2}} \text{ and } \hat{\sigma} = \sqrt{U_2 - \hat{\mu}^2}.$$

- Calculate $\int_{-\infty}^{\infty} f''^2(t)\, dt$ from the following:

$$\int_{-\infty}^{\infty} f''^2(t)\, dt = \frac{3}{16\sqrt{\pi}\hat{\sigma}^5} + \frac{1}{4\sqrt{\pi}\hat{\sigma}^5} e^{-\frac{\hat{\mu}^2}{\hat{\sigma}^2}} \left(\frac{3}{4} - \frac{3\hat{\mu}^2}{\hat{\sigma}^2} + \frac{\hat{\mu}^4}{\hat{\sigma}^4} \right).$$

- Find h_{opt} from the following:

$$h_{opt} = 2^{-\frac{1}{5}} \pi^{-\frac{1}{10}} n^{-\frac{1}{5}} \left\{ \int_{-\infty}^{\infty} f''^2(t) dt \right\}^{-\frac{1}{5}}.$$

- Obtain the estimate of the nonparametric failure distribution of the data:

$$\hat{f}(t) = \frac{1}{n h_{opt}} \sum_{j=1}^{n} K\left(\frac{t - t_j}{h_{opt}} \right).$$

There are other methods for dealing with the optimal bandwidth selection, see Bean and Tsokos (1982) and Silverman (1986).

The nonparametric estimate of the failure probability distribution $\hat{R}_{np}(t)$ can be obtained,

$$\begin{aligned} \hat{R}_{np}(t) &= \int_t^{\infty} \hat{f}(\tau)\, d\tau \\ &= \int_t^{\infty} \frac{1}{n\hat{h}} \sum_{j=1}^{n} K\left(\frac{\tau - t_j}{\hat{h}} \right) d\tau \\ &= \frac{1}{n\hat{h}} \sum_{j=1}^{n} \int_t^{\infty} K\left(\frac{\tau - t_j}{\hat{h}} \right) d\tau \\ &= \frac{1}{n} \sum_{j=1}^{n} \int_{-\infty}^{\frac{t-t_j}{\hat{h}}} K(\tau)\, d\tau. \end{aligned}$$

To evaluate the integral in the above equation, let

$$\Phi(t) = \int_{-\infty}^{t} K(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}} dx.$$

Then we have

$$\begin{aligned} \hat{R}_{np}(t) &= \frac{1}{n} \sum_{j=1}^{n} \left[1 - \Phi\left(\frac{t - t_j}{\hat{h}} \right) \right] \\ &= 1 - \frac{1}{n} \sum_{j=1}^{n} \Phi\left(\frac{t - t_j}{\hat{h}} \right). \end{aligned}$$

To calculate $\Phi(t)$, let

$$\int_0^u e^{-x^2} dx = e^{-u^2} \sum_{k=0}^{\infty} \frac{2^k \cdot u^{2k+1}}{(2k+1)!!}.$$

It follows that we can write

$$\begin{aligned} \Phi(t) &= \int_0^t K(x)\, dx + 0.5 \\ &= \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{x^2}{2}} dx + 0.5 \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{2} \int_0^{\frac{t}{\sqrt{2}}} e^{-\tau^2} d\tau + 0.5 \\ &= \frac{1}{\sqrt{\pi}} \cdot e^{-\frac{t^2}{2}} \cdot \sum_{k=0}^{\infty} \frac{2k \left(\frac{t}{\sqrt{2}} \right)^{2k+1}}{(2k+1)!!} + 0.5 \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} \cdot \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!!} + 0.5. \end{aligned}$$

Note that $\Phi(t) \approx 0$ when $t < -4$, and $\Phi(t) \approx 1$ when $t > 4$. Then we need to carry out the summation in the interval $|t| \le 4$.

Since the sum converges quite fast, when $|t| < 4$, we have overcome the numerical difficulty in the calculation of the nonparametric reliability. An efficient numerical procedure is given below to evaluate the above summation.

Step 1 Construct a subroutine for calculating $\Phi(t)$ as follows:

1. Notations: At input, t stores the point; at output, p stores $\Phi(t)$. Other values for the computation: cc, each term of the sum; tt stores the value $t * t$, to save computer time.
2. Let $p = t$, $cc = t$, $tt = t * t$.
3. For $k = 1, 2, 3, \ldots$, perform (4) through (5) that follows.
4. If $|cc| <$ tolerance (we use 10^{-4} for tolerance), then

$$p = \frac{1}{\sqrt{2\pi}} e^{-\frac{tt}{2}} \cdot p + 0.5,$$

output p and exit. Otherwise continue with (5).

5. $cc = cc \cdot tt/(2k+1)$, $p = p + cc$.

Step 2 Find the optimal bandwidth $\hat{h}$ from the six-step procedure introduced in section "►Parametric Approach to Reliability".

Step 3 The reliability function at any given point t is given by

$$\hat{R}_{np}(t) = 1 - \frac{1}{n}\sum_{j=1}^{n} \Phi\left(\frac{t - t_j}{\hat{h}}\right).$$

Several applications of real data, along with Monte Carlo simulations and the nonparametric kernel probability estimate of reliability, give very good results in comparison with the parametric version of the reliability function.

About the Author

For biography *see* the entry ►Mathematical and Statistical Modeling of Global Warming.

Cross References

►Bayesian Reliability Modeling
►Degradation Models in Reliability and Survival Analysis
►Imprecise Reliability
►Kolmogorov-Smirnov Test
►Nonparametric Density Estimation
►Ordered Statistical Data: Recent Developments
►Tests of Fit Based on The Empirical Distribution Function
►Weibull Distribution

References and Further Reading

Bean S, Tsokos CP (1980) Developments in nonparametric density estimation. Int Stat Rev 48(3):267–287
Bean S, Tsokos CP (1982) Bandwidth selection procedures for kernel density estimates. Comm Stat A Theor 11(9):1045–1069
Liu K, Tsokos CP (2001) Kernel estimates of symmetric density function. Int J Appl Math 6(1):23–34
Liu K, Tsokos CP (2002a) Nonparametric reliability modeling for parallel systems. Stochastic Anal Appl 20(1):185–197
Liu K, Tsokos CP (2002b) Optimal bandwidth selection for a nonparametric estimate of the cumulative distribution function. Int J Appl Math 10(1):33–49
Martz HF, Waller RA (1982) Bayesian reliability analysis. Wiley Series in probability and mathematical statistics: applied probability and statistics. Wiley, Chichester
Miller LH (1956) Table of percentage points of Kolmogorov statistics. J Am Stat Assoc 51:111–121
Owen DB (1962) Handbook of statistical tables. Addison-Wesley, Reading, MA
Qiao H, Tsokos CP (1994) Parameter estimation of the Weibull probability distribution. Math Comput Simulat 37(1):47–55
Qiao H, Tsokos CP (1995) Estimation of the three parameter Weibull probability distribution. Math Comput Simulat 39(1–2):173–185
Rust A, Tsokos CP (1981) On the convergence of kernel estimators of probability density functions. Ann Inst Stat Math 33(2):233–246
Silverman BW (1986) Density estimation for statistics and data analysis. Monographs on statistics and applied probability. Chapman and Hall, London
Tsokos CP (1995) Reliability growth: nonhomogeneous Poisson process. In: Balakrishnan N, and Cohen AC (eds) Recent advances in life-testing and reliability. CRC, Boca Raton, pp 319–334
Tsokos CP (1998) Ordinary and Bayesian approach to life testing using the extreme value distribution. In: Basu AP, Basu SK, Mukhopadhyay S (eds) Frontiers in reliability analysis volume 4 of Series on Quality, Reliability and Engineering Statistics, World Scientific, Singapore, pp 379–395
Tsokos CP, Rust A (1980) Recent developments in nonparametric estimation of probability density. In: Applied stochastic processes (Proc. Conf., Univ. Georgia, Athens, Ga., 1978), Academic, New York, pp 269–281

Parametric Versus Nonparametric Tests

David J. Sheskin
Professor of Psychology
Western Connecticut State University, Danbury, CT, USA

A common distinction made with reference to statistical tests/procedures is the classification of a procedure as *parametric* versus *nonparametric*. This distinction is generally predicated on the number and severity of assumptions regarding the population that underlies a specific test. Although some sources use the term *assumption free* (as well as *distribution free*) in reference to nonparametric tests, the latter label is misleading, in that nonparametric tests are not typically assumption free. Whereas *parametric statistical tests* make certain assumptions with respect to the characteristics and/or parameters of the underlying population distribution upon which a test is based, *nonparametric tests* make fewer or less rigorous assumptions. Thus, as Marascuilo and McSweeney (1977) suggest, nonparametric tests should be viewed as *assumption freer* tests. Perhaps the most common assumption associated with parametric tests that does not apply to nonparametric tests is that data are derived from a normally distributed population.

Many sources categorize a procedure as parametric versus nonparametric based on the level of measurement a set of data being evaluated represents. Whereas parametric tests are typically employed when interval or ratio data are evaluated, nonparametric tests are used with rank-order (ordinal) and categorical data. Common example of parametric procedures employed with interval or ratio data are ►*Student's t tests*, *analysis of variance* procedures (see ►Analysis of Variance), and the *Pearson product moment coefficient of correlation* (see ►Correlation Coefficient). Examples of commonly

described nonparametric tests employed with rank-order data are the *Mann–Whitney U test*, *Wilcoxon's signed–ranks and matched–pairs signed ranks tests*, the *Kruskal–Wallis one-way analysis of variance by ranks*, the *Friedman two–way analysis of variance by ranks*, and *Spearman's rank order correlation coefficient*. Examples of commonly described nonparametric tests employed with categorical data are ►*chi-square tests* such as the *goodness-of-fit test*, *test of independence*, and *test of homogeneity* and the *McNemar test*.

Researchers are in agreement that since ratio and interval data contain a greater amount of information than rank-order and categorical data, if ratio or interval data are available it is preferable to employ a parametric test for an analysis. One reason for preferring a parametric test is that the latter type of test generally has greater *power* than its nonparametric analog (i.e., a parametric test is more likely to reject a false null hypothesis). If, however, one has reason to believe that one or more of the assumptions underlying a parametric test have been saliently violated (e.g., the assumption of underlying normal population distributions associated with the *t test for independent samples*), many sources recommend that a nonparametric test (e.g., a rank-order test that does not assume population normality such as the *Mann–Whitney U test*, which can also be used to evaluate data involving two independent samples) will provide a more reliable analysis of the data. Yet other sources argue it is still preferable to employ a parametric test under the latter conditions, on the grounds that parametric tests are, for the most part, *robust*. A *robust test* is one that still allows a researcher to obtain reasonably reliable conclusions even if one or more of the assumptions underlying the test are violated. As a general rule, when a parametric test is employed in circumstances when one or more of its assumptions are thought to be violated, an adjusted probability distribution is employed in evaluating the data. Sheskin (2007, p. 108) notes that in most instances, the debate concerning whether a researcher should employ a parametric or nonparametric test for a specific experimental design turns out to be of little consequence, since in most cases data evaluated with both a parametric test and its nonparametric analog will result in a researcher reaching the same conclusions.

Cross References

►Analysis of Variance
►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Asymptotic Relative Efficiency in Testing
►Chi-Square Test: Analysis of Contingency Tables
►Explaining Paradoxes in Nonparametric Statistics
►Fisher Exact Test
►Frequentist Hypothesis Testing: A Defense
►Kolmogorov-Smirnov Test
►Measures of Dependence
►Mood Test
►Multivariate Rank Procedures: Perspectives and Prospectives
►Nonparametric Models for ANOVA and ANCOVA Designs
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Permutation Tests
►Randomization Tests
►Rank Transformations
►Robust Inference
►Scales of Measurement and Choice of Statistical Methods
►Sign Test
►Significance Testing: An Overview
►Significance Tests: A Critique
►Statistical Inference
►Statistical Inference: An Overview
►Student's t-Tests
►Validity of Scales
►Wilcoxon–Mann–Whitney Test
►Wilcoxon-Signed-Rank Test

References and Further Reading

Marascuilo LA, McSweeney M (1977) Nonparatmetric and distribution-free methods for the social sciences. Brooks/Cole, Monterey, CA

Sheskin DJ (2007) Handbook of parametric and nonparametric statistical procedures, 4th edn. Chapman and Hall/CRC, Boca Raton

Pareto Sampling

Lennart Bondesson
Professor Emeritus
Umeå University, Umeå, Sweden

Introduction

Let $\mathcal{U} = \{1, 2, \ldots, N\}$ be a population of units. To get information about the population total Y of some interesting y-variable, unequal probability sampling is a widely applied method. Often a random sample of a fixed number, n, of distinct units is to be selected with prescribed inclusion probabilities π_i, $i = 1, 2, \ldots, N$, for the units in $\mathcal{U}$. These π_i should sum to n and are usually chosen to be proportional to some auxiliary variable. In this form unequal

probability sampling is called fixed size πps sampling (ps = proportional to size). By the help of a πps sample and recorded y-values for the units in the sample, the total Y can be estimated unbiasedly by the ►Horvitz–Thompson estimator $\hat{Y}_{HT} = \sum_{i\in s} y_i/\pi_i$, where s denotes the sample. There are many possible fixed size πps sampling designs, see, e.g., Brewer and Hanif (1983) and Tillé (2006).

This article treats Rosén's (1997a,b) Pareto order πps sampling design. Contrary to many other fixed size πps designs, it is very easy to implement. However, it is only approximate. Independently of Rosén, Saavedra (1995) suggested the design. Both were inspired by unpublished work of Ohlsson, cf. Ohlsson (1998).

The Pareto Design

Let U_i, $i = 1, 2, \ldots, N$, be independent $U(0,1)$-distributed random numbers. Further, let

$$Q_i = \frac{U_i/(1-U_i)}{p_i/(1-p_i)}, \quad i = 1, 2, \ldots, N,$$

be so-called ranking variables. Put $p_i = \pi_i$, $i = 1, 2, \ldots, N$, and select as sample those n units that have the smallest Q-values. The factual inclusion probabilities π_i^* do not equal the desired π_i but approximately. For $d = \sum_{i=1}^{N} \pi_i(1-\pi_i)$ not too small ($d > 1$), the agreement is surprisingly good and better the larger d is.

The main advantages of the method are its simplicity, its high ►entropy, and that the U_is can be used as permanent random numbers, i.e., can be reused when at a later occasion the population, more or less altered, is re-sampled. In this way changes can be better estimated.

The name of the method derives from the fact that $u/(1-u) = F^{-1}(u)$, where F^{-1} is the inverse of the special Pareto distribution function $F(x) = x/(1+x)$ on $(0, \infty)$. A general order sampling procedure uses instead $Q_i = F^{-1}(U_i)/F^{-1}(\pi_i)$ for any specified F. As $d \to \infty$, correct inclusion probabilities are obtained but Rosén showed that the Pareto choice gives smallest possible asymptotic bias for them. Ohlsson (1998) uses $Q_i = U_i/\pi_i$, i.e., the distribution function $F(x)$ of a uniform distribution over (0,1).

Probabilistically the Pareto design is very close to Sampford's (1967) design for which the factual inclusion probabilities agree with the desired ones. Let $\mathbf{x}$ be a binary N-vector such that $x_i = 1$ if unit i is sampled and otherwise 0. The probability function $p(\mathbf{x})$ of the Sampford design is given by

$$p(\mathbf{x}) = C \prod_{i=1}^{N} \pi_i^{x_i}(1-\pi_i)^{1-x_i} \times \sum_{k=1}^{N}(1-\pi_k)x_k, \quad \sum_{i=1}^{N} x_i = n,$$

where C is a constant. For the Pareto design the probability function has the same form but the factor $1-\pi_k$ is replaced by c_k, where c_k is given by an integral that is closely proportional to $1-\pi_k$ if d is large (Bondesson et al. 2006).

For the Pareto design the factual inclusion probabilities π_i^* can be calculated in different ways (Aires 1999; Bondesson et al. 2006). By an iterative procedure based on recalculated factual inclusion probabilities, it is possible to adjust the parameters $p_i = \pi_i$ for the Pareto procedure, so that the desired inclusion probabilities π_i are exactly obtained (Aires 2000). The iterative procedure is time consuming. It is also possible to get good improvement by a simple adjustment. The ranking variables Q_i are replaced by the adjusted ranking variables

$$Q_i^{Adj} = Q_i \exp\left(\pi_i(1-\pi_i)\left(\pi_i - \frac{1}{2}\right)/d^2\right).$$

For $N = 6$ and $n = 3$ the following table illustrates the improvement:

	π_i	0.1	0.3	0.4	0.5	0.75	0.95
Pareto	π_i^*	0.0963	0.2916	0.3952	0.5040	0.7610	0.9519
AdjPar	π_i^*	0.0987	0.2987	0.3993	0.5018	0.7510	0.9505

Restricted Pareto Sampling

Pareto sampling can be extended to cases where there are further restrictions on the sample than just fixed sample size (Bondesson 2010). Such restrictions appear if the population is stratified in different ways. The restrictions are usually of the form $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is an $M \times N$ matrix. Then

$$\sum_{i=1}^{N} x_i \log Q_i = \sum_{i=1}^{N} x_i \left(\log \frac{U_i}{1-U_i} - \log \frac{\pi_i}{1-\pi_i}\right)$$

is minimized with respect to $\mathbf{x}$ given the linear restrictions. This minimization can be done by using a program for combinatorial optimization but it usually also suffices to use linear programming and the simplex algorithm with the additional restrictions $0 \le x_i \le 1$ for all i. Under some conditions asymptotically correct inclusion probabilities are obtained. However, in practice some adjustment is often needed. A simple adjustment is to replace Q_i by

$$Q_i^{Adj} = Q_i \exp\left(\pi_i(1-\pi_i)\left(\pi_i - \frac{1}{2}\right)\left(\mathbf{a}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{a}_i\right)^2\right),$$

where $\boldsymbol{\Sigma} = \mathbf{ADA}^T$ with $\mathbf{D} = \mathrm{diag}(\pi_1(1-\pi_1), \ldots, \pi_N(1-\pi_N))$, and $\mathbf{a}_i$ is the ith column vector in $\mathbf{A}$. This method

is suggested by Bondesson (2010), who also presents another method for improvement, conditional Pareto sampling. For the latter method, the random numbers are conditioned to satisfy $\mathbf{AU} = \frac{1}{2}\mathbf{A1}$.

About the Author

Lennart Bondesson (academically a grand-son of Harald Cramér) received his Ph.D. in 1974. In 1983 he became professor of mathematical statistics in forestry at the Swedish University of Agricultural Sciences in Umeå, and in 1999 professor of mathematical statistics at Department of Mathematics and Mathematical Statistics, Umeå University. He has published more than 80 papers and one research book *Generalized Gamma Convolutions and Related Classes of Distributions and Densities* (Lecture Notes in Statistics 76, Springer, 1992). Professor Bondesson was Editor of *Scandinavian Journal of Statistics* (2001–2003).

Cross References

►Horvitz–Thompson Estimator
►Sampling Algorithms
►Uniform Random Number Generators

References and Further Reading

Aires N (1999) Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs. Meth Comput Appl Probab 4:457–469

Aires N (2000) Comparisons between conditional Poisson sampling and Pareto πps sampling designs. J Stat Plann Infer 88:133–147

Bondesson L (2010) Conditional and restricted Pareto sampling; two new methods for unequal probability sampling. Scand J Stat 37, doi: 10.1111/j.1467-9469.2010.000700.x

Bondesson L, Traat I, Lundqvist A (2006) Pareto sampling versus Sampford and conditional Poisson sampling. Scand J Statist 33: 699–720

Brewer KRW, Hanif M (1983) Sampling with unequal probabilities. Lecture Notes in Statistics, No. 15. Springer, New York

Ohlsson E (1998) Sequential Poisson sampling. J Official Stat 14: 149–162

Rosén's B (1997a) Asymptotic theory for order sampling. J Stat Plann Infer 62:135–158

Rosén's B (1997b) On sampling with probability proportional to size. J Stat Plann Infer 62:159–191

Saavedra P (1995) Fixed sample size PPS approximations with a permanent random number. Joint Statistical Meetings American Statistical Association, Orlando, Florida

Sampford MR (1967) On sampling without replacement with unequal probabilities of selection. Biometrika 54:499–513

Tillé Y (2006) Sampling algorithms. Springer series in statistics. Springer science + Business Media, New York

Partial Least Squares Regression Versus Other Methods

Smail Mahdi
Professor of Mathematical Statistics
Mathematics and Physics, University of The West Indies, Cave Hill Campus, Barbados

Introduction

The concept of regression originated from genetics and the word *regression* was introduced into statistical literature in the published paper by Sir Francis Galton (1886) on the relationship between the stature of children and their parents. The relationship was found to be approximately linear with an approximate gradient of 2/3, and this suggested that very tall parents tend to have children shorter than themselves and vice versa. This phenomena was referred to as regression or return to mediocrity or to an average value. The general framework of the linear regression model (see ►Linear Regression Models)

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ is an $n \times q$ matrix of observations on q dependent variables, $\mathbf{X}$ is an $n \times p$ explicative matrix on p variables, $\boldsymbol{\beta}$ is a $p \times q$ matrix of unknown parameters, and $\boldsymbol{\epsilon}$ is $n \times q$ matrix of errors. The rows of $\boldsymbol{\epsilon}$ are independent and identically distributed, often assumed to be Gaussian. The justification for the use of a linear relationship comes from the fact that the conditional mean of a Gaussian random vector given the value of another Gaussian random vector is linear when the joint distribution is Gaussian. Without loss of generality, we will assume throughout that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are mean centered and scaled. The aim is to estimate the unknown matrix $\boldsymbol{\beta}$. The standard approach is to solve in L2 the optimization problem:

$$L(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{\beta}\boldsymbol{X}\|_2.$$

If $\boldsymbol{X}$ has a full rank, then a unique solution exists and is given by

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

When the errors are assumed Gaussian, $\hat{\boldsymbol{\beta}}_{OLS}$ is also the maximum likelihood estimator and therefore, the best unbiased linear estimator (BLUE) of $\boldsymbol{\beta}$. Unfortunately, when some of the dependent variables are collinear, the matrix $\boldsymbol{X}$ does not have full rank and the ordinary least squares estimator may not exist or

becomes inappropriate. To overcome this multicollinearity problem (see ►Multicollinearity), many other estimators have been proposed in literature. This includes ridge regression estimator (RR), principal component regression estimator (PCR), and more recently the partial least squares regression estimator (PLSR).

Ridge Regression

The slightest multicollinearity of the independent vectors may make the matrix $\boldsymbol{X}^T\boldsymbol{X}$ ill conditioned and increase the variance of the components of $\hat{\boldsymbol{\beta}}_{OLS}$, which can lead to an unstable estimation of $\boldsymbol{\beta}$. Hoerl (1962) proposed the ridge regression method (see ►Ridge and Surrogate Ridge Regressions) that consists in adding a small positive constant λ to the diagonal of the standardized matrix $\boldsymbol{X}^T\boldsymbol{X}$ to obtain the RR estimator as

$$\hat{\boldsymbol{\beta}}_{RR} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix. The matrix $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}$ is always invertible since it is positive definite. Several techniques are available in literature for finding the optimum value of λ. Another solution for the collinearity problem is given by the principal component regression (PCR) technique that is outlined below.

Principal Component Regression

Principal component regression is a ►principal component analysis (PCA) followed by a linear regression. In this case, the response variables are regressed on the leading principal components of the matrix $\boldsymbol{X}$, which can be obtained from its singular value decomposition (SVD). PCA is a compressing data technique that consists of finding a k-rank, $k = 1, \ldots, p$, projection matrix $\boldsymbol{P}_k$ such that the variance of the projected data $\boldsymbol{X} \times \boldsymbol{P}_k$ is maximized. The columns of the matrix $\boldsymbol{P}_k$ consist of the k unit-norm leading eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$. Using the matrix $\boldsymbol{P}_k$, we regress $\boldsymbol{Y}$ on the orthogonal score matrix $\boldsymbol{T}_k$ satisfying $\boldsymbol{X} = \boldsymbol{T}_k\boldsymbol{P}_k$; this yields the k latent-component regression coefficients matrix

$$\hat{\boldsymbol{\beta}}_{PCR} = \boldsymbol{P}_k\left(\boldsymbol{T}_k^T\boldsymbol{T}_k\right)^{-1}\boldsymbol{T}_k^T\boldsymbol{Y}.$$

However, two major problems may occur: firstly the choice of k and secondly, how, if the ignored principal components that correspond to the small eigenvalues are in fact relevant for explaining the covariance structure of the $\boldsymbol{Y}$ variables. For this reason, PLS comes into play to better deal with the multicollinearity problem by creating $\boldsymbol{X}$ latent components for explaining $\boldsymbol{Y}$, through the maximization of the covariance structure between the $\boldsymbol{X}$ and $\boldsymbol{Y}$ spaces.

Partial Least Squares Regression

Partial least squares (PLS), also known as projection method to latent structures, is applied to a broad area of data analysis including ►generalized linear models, regression, classification, and discrimination. It is a multivariate technique that generalizes and combines ideas from principal component analysis (PCA) and ordinary least squares (OLS) regression methods. It is designed to not only confront situations of correlated predictors but also relatively small samples and even the situation where the number of dependent variables exceeds the number of cases. The original idea came in the work of the Swedish statistician Herman Wold (1966) in the 1960s and became popular in computational chemistry and sensory evaluation by the work of his son Svante who developed the popular software soft independent modeling of class analogies (SIMCA) (Wold 1976). PLS find first two sets of weights $\boldsymbol{\omega} = (w_1, \ldots, w_m)$ and $\boldsymbol{U} = (u_1, \ldots, u_m)$ for $\boldsymbol{X}$ and $\boldsymbol{Y}$ such that $Cov\left(\boldsymbol{t}_l^X, \boldsymbol{t}_l^Y\right)$, where $\boldsymbol{t}_l^X = \boldsymbol{X} \times \boldsymbol{\omega}_l$, $\boldsymbol{t}_l^Y = \boldsymbol{Y} \times \boldsymbol{U}_l$ and $l = 1, \ldots, m$, is maximal. Then, the $\boldsymbol{Y}$ variables are regressed on the latent matrix $\boldsymbol{T}$ whose columns are the latent vectors $\boldsymbol{t}_i$, see, e.g., (Vinzi et al. 2005) for more details. Classically, the ordinary least squares regression (OLS) is used but other methods have been considered along with the corresponding algorithms. We describe the PLS technique through the algorithm, outlined, for instance in Mevik and Wehrens (2007), which is based on the singular value decomposition (SVD) of the cross product $\boldsymbol{X}^T\boldsymbol{Y}$ types. First, we set $\boldsymbol{E} = \boldsymbol{X}$ and $\boldsymbol{F} = \boldsymbol{Y}$. Then, we perform the singular decomposition of $\boldsymbol{E}^T\boldsymbol{F}$ and take the first left-singular vector $\boldsymbol{\omega}$ and the right-singular vector $\boldsymbol{q}$ of $\boldsymbol{E}^T\boldsymbol{F}$ to obtain the scores $\boldsymbol{t}$ and $\boldsymbol{u}$ as follows:

$$\boldsymbol{t} = \boldsymbol{E}\boldsymbol{\omega}$$

and

$$\boldsymbol{u} = \boldsymbol{F}\boldsymbol{q}.$$

The first score then is obtained as $\boldsymbol{t} = \boldsymbol{t}_1 = \frac{\boldsymbol{t}}{(\boldsymbol{t}^T\boldsymbol{t})^{1/2}}$. The effect of this score $\boldsymbol{t}_1$ on $\boldsymbol{E}$ and $\boldsymbol{F}$ is obtained by regressing $\boldsymbol{E}$ and $\boldsymbol{F}$ on $\boldsymbol{t}_1$. This gives $\boldsymbol{p} = \boldsymbol{E}^T\boldsymbol{t}$ and $\boldsymbol{q} = \boldsymbol{F}^T\boldsymbol{t}$. This effect is then subtracted from $\boldsymbol{E}$ and $\boldsymbol{F}$ to obtain the deflated matrices, with one rank less, $\boldsymbol{E}'$ and $\boldsymbol{F}'$, see, for example, Wedderburn (1934). The matrices $\boldsymbol{E}'$ and $\boldsymbol{F}'$ are then substituted for $\boldsymbol{E}$ and $\boldsymbol{F}$ and the process is reiterated again from the singular value decomposition of $\boldsymbol{E}^T\boldsymbol{F}$ to obtain the second normalized latent component $\boldsymbol{t}_2$. The process is continued until the required number m of latent components is obtained. The optimal value of m is often obtained by cross-validation or bootstrap, see, for example, Tenenhaus (1998). The obtained latent components $\boldsymbol{t}_l$, $l = 1, \ldots, m$ are

saved after each iteration as column of the latent matrix $\boldsymbol{T}$. Finally, the original Y variables are regressed on the latent components $\boldsymbol{T}$ to obtain the regression estimator

$$\hat{\boldsymbol{\beta}}_{PLS} = (\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{Y}$$

and the estimator

$$\hat{\boldsymbol{Y}} = \boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{Y}$$

of $\boldsymbol{Y}$. Note that several of the earlier PLS algorithms, available in literature, lack robustness. Recently, Simonetti et al. (2006) substituted systematically the least median of squares regression, Rousseeuw (1984), for the least squares regression in Garthwaite (1994) PLS setup to obtain a robust estimation for the considered data set.

Cross References

- ►Chemometrics
- ►Least Squares
- ►Linear Regression Models
- ►Multicollinearity
- ►Multivariate Statistical Analysis
- ►Principal Component Analysis
- ►Ridge and Surrogate Ridge Regressions

References and Further Reading

Garthwaite PH (1994) An interpretation of partial least squares. J Am Stat Assoc 89(425):122–127

Galton F (1886) Regression toward mediocrity in hereditary stature. Nature 15:246–263

Hoerl AE (1962) Application of ridge analysis to regression problems. Chem Eng Prog 58:54–59

Mevik B, Wehrens R (2007) The pls package: principal component and partial least squares regression in R. J Stat Softw 18(2):1–24

Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–888

Simonetti B, Mahdi S, Camminatiello I (2006) Robust *PLS* regression based on simple least median squares regression. MTISD'06 Conference, Procida, Italy

Tenenhaus M (1998) La Régréssion *PLS*: Théorie et Pratique. Technip, Paris

Vinzi VE, Lauro CN, Amato S (2005) *PLS* typological regression: algorithms, classification and validation issues. New developments in classification and data analysis: Vichi M, Monari P, Mignani S, Montanari A (eds) Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Bologna, Springer, Berlin, Heidelberg

Wedderburn JHM (1934) Lectures on matrices. AMS, vol 17. Colloquium, New York

Wold H (1966) Estimation of principal components and related models by iterative least squares. In: Krishnaiah PR (ed) Multivariate analysis. Academic, New York, pp 391–420

Wold S (1976) Pattern recognition by means of disjoint principal components models. Pattern Recogn 8:127–139

Pattern Recognition, Aspects of

Dražen Domijan[1], Bojana Dalbelo Bašić[2]
[1]Associate Professor
University of Rijeka, Rijeka, Croatia
[2]Professor
University of Zagreb, Zagreb, Croatia

Introduction

Recognition is regarded as a basic attribute of human beings and other living organisms. A pattern is a description of an object (Tou and Gonzalez 1974). Pattern recognition is a process of assigning category labels to a set of patterns. For instance, visual patterns "A," "a," and "***A***" are members of the same category, which is labeled as "letter A" and can easily be distinguished from the patterns, "B," "b," and "***B***," which belong to another category labeled as "letter B." Humans perform pattern recognition very well and the central problem is how to design a system to match human performance. Such systems find practical applications in many domains such as medical diagnosis, image analysis, face recognition, speech recognition, handwritten character recognition, and more (Duda et al. 2001; Fukunaga 1990).

The problem of pattern recognition can be tackled using handcrafted rules or heuristics for distinguishing the category of objects, though in practice such an approach leads to proliferation of the rules and exceptions to the rules, and invariably gives poor results (Bishop 2006). Some classification problems can be tackled by syntactic (linguistic) pattern recognition methods, but most real-word problems are tackled using the machine learning approach. Pattern recognition is an interdisciplinary field involving statistics, probability theory, computer science, machine learning, linguistics, cognitive science, psychology, etc. Pattern recognition systems involve the following phases: sensing, feature generation, feature selection, classifier design, and system performance evaluation (Tou and Gonzalez 1974).

In the previous example, the pattern was a set of black and white pixels. However, patterns might be a set of continuous variables. In statistical pattern recognition, features are treated as random variables. Therefore, patterns are random vectors that are assigned to a class or category with certain probability. In this case, patterns could be conceived as points in a high-dimensional feature space. Pattern recognition is the task of estimating a function that divides the feature space into regions, where each region corresponds to one of the categories (classes).

Such a function is called the decision or discriminant function and the surface that is realized by the function is known as the decision surface.

Feature Generation and Feature Selection

Before the measurement data obtained from the sensor could be utilized for the design of the pattern classifier, sometimes it is necessary to perform several preprocessing steps such as outlier removal, data normalization, and treatment of the missing data. After preprocessing, features are generated from measurements using data reduction techniques, which exploits and removes redundancies in the original data set. A popular way of generating features is to use linear transformations such as the Karhunen–Loeve transformation (▶principal component analysis), independent component analysis, discrete Fourier transform, discrete cosine and sine transforms, and Hadamard and Haar transforms. Important consideration in using these transformations is that they should preserve as much of the information that is crucial for classification task as possible, while removing as much redundant information as possible (Theodoridis and Koutroumbas 2009).

After the features are generated, they could be independently tested for their discriminatory capability. We might select features based on the statistical hypothesis testing. For instance, we may employ t-test or Kruskall–Wallis test to investigate the difference in mean feature values for two classes. Another approach is to construct the characteristic receiver operating curve and to explore how much overlap exists between distributions of feature values for two classes. Furthermore, we might compute class separability measures that take into account correlations between features such as Fisher's discriminant ratio and divergence. Besides testing individual features, we might ask what is the best feature vector or combination of features that gives the best classification performance. There are several searching techniques such as sequential backward or forward selection that can be employed in order to find an answer (Theodoridis and Koutroumbas 2009).

Design of a Pattern Classifier

The principled way to design a pattern classifier would involve characterization of the class probability density functions in the feature space and finding an appropriate discriminant function to separate the classes in this space. Every classifier can make an error by assigning the wrong class label to the pattern. The goal is to find the classifier with the minimal probability of classification error. The best classifier is based on the Bayes decision rule (Fukunaga 1990). The basic idea is to assign a pattern to the class having the highest a posteriori probability for a given pattern. A posteriori probabilities for a given pattern are computed from a priori class probabilities and conditional density functions. However, in practice, it is often difficult to compute a posteriori probabilities as a priori class probabilities are not known in advance.

Therefore, although the Bayesian classifiers are optimal they are rarely used in practice. Instead, classifiers are designed directly from the data. A simple and computationally efficient approach to the design of the classifier is to assume that the discriminant function is linear. In that case, we can construct a decision hyperplane through the feature space defined by

$$g(x) = w^T x + w_0 = 0$$

where $w = [w_1, w_2, \ldots]^T$ are unknown coefficients or weights, $x = [x_1, x_2, \ldots]$ is a feature vector, and w_0 is a threshold or bias. Finding unknown weights is called learning or training. In order to find an appropriate value for the weights, we can use iterative procedures such as the perceptron learning algorithm. The basic idea is to compute error or cost function, which measures the difference between actual classifier output and desired output. Error function is used to adjust current values of the weights. This process is repeated until perceptron converges, that is, until all patterns are correctly classified. This is possible if patterns are linearly separable. After the perceptron converges, new patterns could be classified according to the simple rule:

$$\text{If } w^T x + w_0 > 0 \text{ assign } x \text{ to class } \omega_1$$
$$\text{If } w^T x + w_0 < 0 \text{ assign } x \text{ to class } \omega_2$$

where $\omega 1$ and $\omega 2$ are class labels.

The problem with linear classifiers is that they lead to suboptimal performance when classes are not linearly separable. An example of pattern recognition problem that is not linearly separable is the logical predicate XOR where the patterns (0,1) and (1,0) belong to class $\omega 1$, and the patterns (0,0) and (1,1) belong to $\omega 2$. It is not possible to draw a straight line (linear decision boundary) in two-dimensional feature space that will discriminate between these two classes. One approach to deal with such problems is to build a linear classifier that will minimize the mean square error between the desired and the actual output of the classifier. This can be achieved using least mean square (LMS) or Widrow–Hoff algorithm for weight

adjustment. Another approach is to design a nonlinear classifier (Bishop 1995). Examples of nonlinear classifiers are multilayer perceptron trained with error backpropagation, radial basis functions network, k-nearest neighbor classifier, and decision trees. In practice, it is possible to combine outputs from several different classifiers in order to achieve better performance. Classification is an example of so-called supervised learning in which each feature vector has a preassigned target class.

Performance Evaluation of the Pattern Classifier

An important task in the design of a pattern classifier is how well it will perform when faced with new patterns. This is an issue of generalization. During learning, the classifier builds a model of the environment to which it is exposed. The model might vary in complexity. A complex model might offer a better fit to the data, but might also capture more noise or irrelevant characteristics in the data, and thus be poor in the classification of new patterns. Such a situation is called over-fitting. On the other hand, if the model is of low complexity, it might not fit the data well. The problem is how to select an appropriate level of complexity that will enable the classifier to fit the observed data well, while preserving enough flexibility to classify unobserved patterns. This is known as a bias-variance dilemma (Bishop 1995).

The performance of the designed classifier is evaluated by counting the number of errors committed during a testing with a set of feature vectors. Error counting provides an estimation of classification error probability. The important question is how to choose a set of feature vectors that will be used for building the classifier and a set of feature vectors that will be used for testing. One approach is to exclude one feature vector from the sample, train the classifier on all other vectors, and then test the classifier with the excluded vector. If misclassification occurs, error is counted. This procedure is repeated N times with different excluded feature vectors every time. The problem with this procedure is that it is computationally demanding. Another approach is to split the data set into two subsets: (1) a training sample used to adjust (estimate) classifier parameters and (2) a testing sample that is not used during training but is applied to the classifier following completion of the training. The problem with this approach is that it reduces the size of the training and testing samples, which reduces the reliability of the estimation of classification error probability (Theodoridis and Koutroumbas 2009).

Acknowledgment

We would like to thank Professor Sergios Theodoridis for helpful comments and suggestions that significantly improved our presentation.

Cross References

►Data Analysis
►Fuzzy Sets: An Introduction
►ROC Curves
►Significance Testing: An Overview
►Statistical Pattern Recognition Principles

References and Further Reading

Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, Oxford, UK

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York

Fukunaga K (1990) Introduction to statistical pattern recognition, 2nd edn. Academic, San Diego, CA

Theodoridis S, Koutroumbas K (2009) Pattern recognition, 4th edn. Elsevier

Tou JT, Gonzalez RC (1974) Pattern recognition Principles. Addison-Wesley, Reading, MA

Permanents in Probability Theory

Ravindra B. Bapat
Professor, Head New Delhi Centre
Indian Statistical Institute, New Delhi, India

Preliminaries

If A is an $n \times n$ matrix, then the permanent of A, denoted by per A, is defined as

$$\text{per } A = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i\sigma(i)},$$

where S_n is the set of permutations of $1, 2, \ldots, n$. Thus the definition of the permanent is similar to that of the determinant except that all the terms in the expansion have a positive sign.

Example: Consider the matrix

$$A = \begin{bmatrix} 2 & -1 & 3 \\ 1 & 2 & 3 \\ -2 & 4 & 1 \end{bmatrix}.$$

Then

$$\text{per } A = 4 + 6 + 12 - 12 + 24 - 1 = 33.$$

Permanents find numerous applications in probability theory, notably in the theory of discrete distributions and ▶order statistics. There are two main advantages of employing permanents in these areas. Firstly, permanents serve as a convenient notational device, which makes it feasible to write complex expressions in a compact form. The second advantage, which is more important, is that some theoretical results for the permanent lead to statements of interest in probability theory. This is true mainly of the properties of permanents of entrywise nonnegative matrices.

Although the definition of the permanent is similar to that of the determinant, many of the nice properties of the determinant do not hold for the permanent. For example, the permanent is not well behaved under elementary transformations, except under the transformation of multiplying a row or column by a constant. Similarly, the permanent of the product of two matrices does not equal the product of the permanents in general. The Laplace expansion for the determinant holds for the permanent as well and is a convenient tool for dealing with the permanent. Thus, if $A(i,j)$ denotes the submatrix obtained by deleting row i and column j of the $n \times n$ matrix A, then

$$\text{per } A = \sum_{k=1}^{n} a_{ik} \text{per } A(i,k) = \sum_{k=1}^{n} a_{ki} \text{per } A(k,i), \quad i = 1, 2, \ldots, n.$$

We refer to van Lint and Wilson (2001) for an introduction to permanents.

Combinatorial Probability

Matrices all of whose entries are either 0 or 1, the so called $(0,1)$-matrices, play an important role in combinatorics. Several combinatorial problems can be posed as problems involving counting certain permutations of a finite set of elements and hence can be represented in terms of $(0,1)$-matrices. We give two well-know examples in combinatorial probability.

Consider n letters and n envelopes carrying the corresponding addresses. If the letters are put in the envelopes at random, what is the probability that none of the letters goes into the right envelope? The probability is easily seen to be

$$\frac{1}{n!}\text{per}\begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix},$$

which equals $1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!}$. The permanent in the above expression counts the number of *derangements* of n symbols.

Another problem, which may also be posed as a probability question, is the *problème des ménages* (Kaplansky and Riordan 1946). In how many ways can n couples be placed at a round table so that men and women sit in alternate places and no one is sitting next to his or her spouse? This number equals $2n!$ times the permanent of the matrix $J_n - I_n - P_n$, where J_n is the $n \times n$ matrix of all ones, I_n is the $n \times n$ identity matrix, and P_n is the full cycle permutation matrix, having ones at positions $(1,2), (2,3), \ldots, (n-1,n), (n,1)$ and zeroes elsewhere. The permanent can be expressed as

$$\text{per } (J_n - I_n - P_n) = \sum_{i=0}^{n} (-1)^i \frac{2n}{2n-i} \binom{2n-i}{i} (n-i)!.$$

Discrete Distributions

The densities of some discrete distributions can be conveniently expressed in terms of permanents. We illustrate by an example of multiparameter ▶multinomial distribution.

We first consider the multiparameter binomial. Suppose n coins, not necessarily identical, are tossed once, and let X be the number of heads obtained. Let p_i be the probability of heads on a single toss of the i-th coin, $i = 1, 2, \ldots, n$. Let $\mathbf{p}$ be the column vector with components $p_1, \ldots, p_n$, and let $\mathbf{1}$ be the column vector of all ones. Then it can be verified that

$$Prob(X = x) = \frac{1}{x!(n-x)!} \text{per } [\underbrace{\mathbf{p}, \ldots, \mathbf{p}}_{r} \underbrace{(\mathbf{1}-\mathbf{p}), \ldots, (\mathbf{1}-\mathbf{p})}_{n-r}]. \tag{1}$$

Now consider an experiment which can result in any of r possible outcomes, and suppose n trials of the experiment are performed. Let p_{ij} be the probability that the experiment results in the j-th outcome at the i-th trial, $i = 1, 2, \ldots, n; j = 1, 2, \ldots, r$. Let P denote the $n \times r$ matrix (p_{ij}) which is row stochastic. Let $P_1, \ldots, P_r$ be the columns of P. Let X_j denote the number of times the j-th outcome is obtained in the n trials, $j = 1, 2, \ldots, r$. If $k_1, \ldots, k_r$ are nonnegative integers summing to n, then as a generalization of (1) we have (Bapat 1990)

$$Prob(X_1 = k_1, \ldots, X_r = k_r) = \frac{1}{k_1! \cdots k_r!} \text{per } [\underbrace{P_1, \ldots, P_1}_{k_1}, \ldots, \underbrace{P_r, \ldots, P_r}_{k_r}].$$

Similar representations exist for the multiparameter negative binomial.

Order Statistics

Permanents provide an effective tool in dealing with order statistics corresponding to independent random variables, which are not necessarily identically distributed. Let $X_1, \ldots, X_n$ be independent random variables with distribution functions $F_1, \ldots, F_n$ and densities $f_1, \ldots, f_n$ respectively. Let $Y_1 \leq \cdots \leq Y_n$ denote the corresponding order statistics. We introduce the following notation. For a fixed y, let

$$\hat{f}(y) = \begin{bmatrix} f_1(y) \\ \vdots \\ f_n(y) \end{bmatrix} \text{ and } \hat{F}(y) = \begin{bmatrix} F_1(y) \\ \vdots \\ F_n(y) \end{bmatrix}.$$

For $1 \leq r \leq n$, the density function of Y_r is given by Vaughan and Venables (1972)

$$g_r(y) = \frac{1}{(r-1)!(n-r)!} \text{per} \left[\underbrace{\hat{f}(y)}_{1} \underbrace{\hat{F}(y)}_{r-1} \underbrace{\mathbf{1} - \hat{F}(y)}_{n-r}\right], \infty < y < \infty$$

For $1 \leq r \leq n$, the distribution function of Y_r is given by Bapat and Beg (1989)

$$Prob(Y_r \leq y) = \sum_{i=r}^{n} \frac{1}{i!(n-i)!} \text{per} \left[\underbrace{\hat{F}(y)}_{i} \underbrace{\mathbf{1} - \hat{F}(y)}_{n-i}\right], \quad \infty < y < \infty$$

The permanental representation can be used to extend several recurrence relations for order statistics from the i.i.d. case to the case of nonidentical, independent random variables. Using the Alexandroff inequality for the permanent of a nonnegative matrix, it can be shown (Bapat 1990) that for any y, the sequence $Prob(Y_i \leq y | Y_{i-1} \leq y)$, $i = 2, \ldots, n$ is nonincreasing.

For additional material on applications of permanents in order statistics we refer to Balakrishnan (2007) and the references contained therein.

Acknowledgments

The support of the JC Bose Fellowship, Department of Science and Technology, Government of India, is gratefully acknowledged.

About the Author

Past President of the Indian Mathematical Society (2007–2008), Professor Bapat joined the Indian Statistical Institute, New Delhi, in 1983, where he holds the position of Head, Delhi Centre at the moment. He held visiting positions at various Universities in the U.S., including University of Connecticut and Oakland University, and visited several Institutes abroad in countries including France, Holland, Canada, China and Taiwan for collaborative research and seminars. The main areas of research interest of Professor Bapat are nonnegative matrices, matrix inequalities, matrices in graph theory and generalized inverses. He has published more than 100 research papers in these areas in reputed national and international journals. He has written books on Linear Algebra, published by Hindustan Book Agency, Springer and Cambridge University Press. He has recently written a book on Mathematics for the general readers, in Marathi, which won the state government award for best literature in Science for 2004. He is Elected Fellow of the Indian Academy of Sciences, Bangalore and Indian National Science Academy, Delhi.

Cross References

►Multinomial Distribution
►Order Statistics
►Probability Theory: An Outline
►Random Permutations and Partition Models
►Univariate Discrete Distributions: An Overview

References and Further Reading

Balakrishnan N (2007) Permanents, order statistics, outliers, and robustness. Rev Mat Complut 20(1):7–107
Bapat RB (1990) Permanents in probability and statistics. Linear Algebra Appl 127:3–25
Bapat RB, Beg MI (1989) Order statistics for nonidentically distributed variables and permanents. Sankhya A 51(1):79–93
Kaplansky I, Riordan J (1946) The probléme des ménages. Scripta Math 12:113–124
van Lint JH, Wilson RM (2001) A course in combinatorics, 2nd edn. Cambridge University Press, Cambridge
Vaughan RJ, Venables WN (1972) Permanent expressions for order statistic densities J R Stat Soc B 34:308–310

Permutation Tests

Markus Neuhäuser
Professor
Koblenz University of Applied Sciences, Remagen, Germany

A permutation test is illustrated here for a two-sample comparison. The notation is as follows: Two independent groups with sample sizes n and m have independently and

identically distributed values $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively, $n + m = N$. The means are denoted by $\overline{X}$ and $\overline{Y}$, and the distribution functions by F and G. These distribution functions of the two groups are identical with the exception of a possible location shift: $F(t) = G(t - \theta)$ for all $t, -\infty < \theta < \infty$. The null hypothesis states $H_0 : \theta = 0$, whereas $\theta \neq 0$ under the alternative H_1.

In this case Student's t test (see ▶Student's t-Tests) can be applied. However, if F and G were not normal distributions, it may be better to avoid using the t distribution. An alternative method is to use the permutation null distribution of the t statistic.

In order to generate the permutation distribution all possible permutations under the null hypothesis have to be generated. In the two-sample case, each permutation is a possible (re-)allocation of the N observed values to two groups of sizes n and m. Hence, there are $\binom{N}{n}$ possible permutations. The test statistic is calculated for each permutation. The null hypothesis can then be accepted or rejected using the permutation distribution of the test statistic, the p-value being the probability of the permutations giving a value of the test statistic as or more supportive of the alternative than the observed value. Thus, inference is based upon how extreme the observed test statistic is relative to other values that could have been obtained under the null hypothesis.

Under H_0 all permutations have the same probability. Hence, the p-value can simply be computed as the proportion of the permutations with a test statistic's value as or more supportive of the alternative than the observed value.

The order of the permutations is important, rather than the exact values of the test statistic. Therefore, modified test statistics can be used (Manly 2007, pp. 16–17). For example, the difference $\overline{X} - \overline{Y}$ can be used instead of the t statistic

$$t = \frac{\overline{X} - \overline{Y}}{S \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where S is the estimated standard deviation.

This permutation test is called Fisher–Pitman permutation test or randomization test (see ▶Randomization Tests), it is a nonparametric test (Siegel 1956; Manly 2007). However, at least an interval measurement is required for the Fisher-Pitman test because the test uses the numerical values $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ (Siegel 1956).

The permutation distribution depends on the observed values, therefore a permutation test is a conditional test, and a huge amount of computing is required. As a result, the Fisher–Pitman permutation test was hardly ever applied before the advent of fast PCs, although it was proposed in the 1930s and its high efficiency was known since decades. Nowadays, the test is often recommended and implemented in standard software such as SAS (for references see Lehmann 1975, or Neuhäuser and Manly 2004).

Please note that the randomization model of inference does not require randomly sampled populations. For a permutation test it is only required that the groups or treatments have been assigned to the experimental units at random (Lehmann 1975).

A permutation test can be applied with other test statistics, too. Rank-based statistics such as Wilcoxon's rank sum can also be used as test statistic (see the entry about the Wilcoxon–Mann–Whitney test). Rank tests can also be applied for ordinal data. Moreover, rank tests had advantages in the past because the permutation null distribution can be tabulated. Nowadays, with modern PCs and fast algorithms, permutation tests can be carried out with any suitable test statistic. However, rank tests are relatively powerful in the commonly-occurring situation where the underlying distributions are non-normal (Higgins 2000). Hence, permutation tests on ranks are still useful despite the fact that more complicated permutation tests can be carried out (Neuhäuser 2005).

When the sample sizes are very large, the number of possible permutations can be extremely large. Then, a ▶simple random sample of M permutations can be drawn in order to estimate the permutation distribution. A commonly used value is $M = 10,000$. Please note that the original observed values must be one of the M selected permutations (Edgington and Onghena 2007, p. 41).

It should be noted that permutation procedures do have some disadvantages. First, they are computer-intensive, although the computational effort seems to be no longer a severe objection against permutation tests. Second, conservatism is often the price for exactness. For this reason, the virtues of permutation tests continue to be debated in the literature (Berger 2000).

When a permutation test is performed for other situations than the comparison of two samples the principle is analogue. The Fisher-Pitman test can also be carried out with the ANOVA F statistic. Permutation tests can also be applied in case of more complex designs. For example, the residuals can be permuted (ter Braak 1992; Anderson 2001).

Permutation tests are also possible for other situations than the location-shift model. For example, Janssen (1997) presents a permutation test for the ▶Behrens-Fisher problem where the population variances may differ.

About the Author

For biography *see* the entry ►Wilcoxon-Mann-Whitney Test.

Cross References

►Behrens–Fisher Problem
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Randomization
►Randomization Tests
►Statistical Fallacies: Misconceptions, and Myths
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Anderson MJ (2001) Permutation tests for univariate or multivariate analysis of variance and regression. Can J Fish Aquat Sci 58: 626–639

Berger VW (2000) Pros and cons of permutation tests in clinical trials. Stat Med 19:1319–1328

Edgington ES, Onghena P (2007) Randomization tests, 4th edn. Chapman and Hall/CRC, Boca Raton

Higgins JJ (2000) Letter to the editor. Am Stat 54, 86

Janssen A (1997) Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens–Fisher problem. Stat Probab Lett 36:9–21

Lehmann EL (1975) Nonparametrics: statistical methods based on ranks. Holden-Day, San Francisco

Manly BFJ (2007) Randomization, bootstrap and Monte Carlo methods in biology, 3rd edn. Chapman and Hall/CRC, London

Neuhäuser M (2005) Efficiency comparisons of rank and permutation tests. Stat Med 24:1777–1778

Neuhäuser M, Manly BFJ (2004) The Fisher-Pitman permutation test when testing for differences in mean and variance. Psychol Rep 94:189–194

Siegel S (1956) Nonparametric statistics for the behavioural sciences. McGraw-Hill, New York

ter Braak CJF (1992) Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Jöckel KH, Rothe G, Sendler W (eds) Bootstraping and related techniques. Springer, Heidelberg, pp 79–85

Pharmaceutical Statistics: Bioequivalence

Francis Hsuan
Professor Emeritus
Temple University, Philadelphia, PA, USA

In the terminology of pharmacokinetics, the *bioavailability* (BA) of a drug product refers to the rate and extent of its absorbed active ingredient or active moiety that becomes available at the site of action. A new/test drug product (T) is considered *bioequivalent* to an existing/reference product (R) if there is no significant difference in the bioavailabilities between the two products when they are administered at the same dose under similar conditions in an appropriately designed study. Studies to demonstrate either *in-vivo* or *in-vitro* bioequivalence are required for government regulatory approvals of generic drug products, or new formulations of existing products with known chemical entity.

Statistically an *in-vivo* bioequivalence (BE) study employs a crossover design with T and R drug products administered to healthy subjects on separate occasions according to a pre-specified randomization schedule, with ample washout between the occasions. The concentration of the active ingredient of interest in blood is measured over time per subject in each occasion, resulting in multiple concentration-time profiles curves for each subject. From which a number of bioavailability measures, such as area under the curve (AUC) and peak concentration (C_{max}) in either raw or logarithmic scale, are then computed, statistically modeled, and analyzed for bioequivalence. Let Y_{ijkt} be a log-transformed bioavailability measure of subject j in treatment sequence i at period t, having the treatment k. In a simplest 2×2 crossover design with two treatment sequences T/R ($i=1$) and R/T ($i=2$), is assumed to follow a mixed-effects ANOVA model

$$Y_{ijkt} = \mu_0 + \omega_i + \phi_k + \pi_t + S_{ij} + \varepsilon_{ijkt}$$

where μ_0 is an overall constant, ω_i, ϕ_k and π_t are, respectively, (fixed) effects of sequence i, formulation k and period t, S_{ij} is the (random) effect of subject j in sequence i, and ε_{ijkt} the measurement error. In this model, the between-subject variation is captured by the random component $S_{ij} \sim N\left(0, \sigma_B^2\right)$ and the within-subject variation $\varepsilon_{ijkt} \sim N\left(0, \sigma_{Wk}^2\right)$ is allowed to depend on formulation k. Bioequivalence of T and R is declared when the null hypothesis $H_0 : \phi_T \le \phi_R - \varepsilon$ or $\phi_T \ge \phi_R + \varepsilon$ can be rejected against the alternative $H_1 : -\varepsilon \le \phi_T - \phi_R \le \varepsilon$ at some significance level α, where $\varepsilon = \log(1.25) = 0.2231$ and $\alpha = 0.05$ are set by regulatory agencies. This last sentence is referred to as the criterion for *average BE*. A common method to establish average BE is to calculate the shortest $(1-2\alpha)\times 100\%$ confidence interval of $\delta = \phi_T - \phi_R$ and show that it is contained in the equivalence interval $(-\varepsilon, \varepsilon)$.

Establishing average BE using any nonreplicated $k \times k$ crossover design can be conducted along the same line as described above. For certain types of drug products, such as those with Narrow Therapeutic Index (NTI), questions were raised regarding the adequacy of the average BE method (e.g., Blakesley et al. 2004). To address this issue,

other criteria and/or statistical methods for bioequivalence have been proposed and studied in the literature. In partic-h ular, *population BE* (PBE) assesses the difference between T and R in both means and variances of bioavailability measures, and *individual BE* (IBE) assesses, in addition, the variation in the average T and R difference among individuals. Some of these new concepts, notably individual bioequivalence, would require high-order crossover designs such as TRT/RTR. Statistical designs and analyses for population BE and individual BE are described in detail in a statistical guidance for industry (USA FDA, 2001) and several monographs (Chow and Shao 2002; Wellek 2003). Hsuan and Reeve (2003) proposed a unified procedure to establish IBE using any high-order crossover design and a multivariate ANOVA model. Recently the USA FDA (2007) proposes a process of making available the public guidance(s) on how to design bioequivalence studies for specific drug products.

About the Author

Dr. Francis Hsuan has been a faculty member in the Department of Statistics at the Fox School, Temple University, for more than 25 years. He received his Ph.D. in Statistics from Cornell University and his B.S. in Physics from the National Taiwan University. He was also a visiting scholar at the Harvard School of Public Health from 1998 to 1999, working on projects related to the analysis of longitudinal categorical data with informative missingness.

Cross References

►Biopharmaceutical Research, Statistics in
►Equivalence Testing
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences

References and Further Reading

Blakesley V, Awni W, Locke C, Ludden T, Granneman GR, Braverman LE (2004) Are bioequivalence studies of levothyroxine sodium formulations in euthyroid volunteers reliable? Thyroid 14:191–200

Chow SC, Shao J (2002) Statistics in drug research: methodologies and recent developments. Marcell Dekker, New York

Hsuan F, Reeve R (2003) Assessing individual bioequivalence with high-order crossover designs: a unified procedure. Stat Med 22:2847–2860

FDA (2001) Guidance for industry on statistical approaches to establishing bioequivalence. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland, USA, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm070244.pdf

FDA (2007) Guidance for industry on bioequivalence recommendations for specific products. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, Maryland, USA, http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072872.pdf

Wellek S (2003) Testing statistical hypotheses of equivalence. Chapman and Hall/CRC, Florida, USA

Philosophical Foundations of Statistics

Inge S. Helland
Professor
University of Oslo, Oslo, Norway

The philosophical foundations of statistics involve issues in theoretical statistics, such as goals and methods to meet these goals, and interpretation of the meaning of inference using statistics. They are related to the philosophy of science and to the ►philosophy of probability.

As with any other science, the philosophical foundations of statistics are closely connected to its history, which again is connected to the men with whom different philosophical directions can be associated. Some of the most important names in this connection are Thomas Bayes (1702–1761), Ronald A. Fisher (1890–1962) and Jerzy Neyman (1894–1981).

The standard statistical paradigm is tied to the concept of a statistical model, an indexed family of probability measures $P^{\theta}(\cdot)$ on the observations, indexed by the parameters θ. Inference is done on the parameter space. This paradigm was challenged by Breiman (2001), who argued for an algorithmical, more intuitive model concept. Breiman's tree models are still much in use, together with other algorithmical bases for inference, for instance within chemometry. For an attempt to explain some of these within the framework of the ordinary statistical paradigm, see Helland (2010).

On the other hand, not all indexed families of distributions lead to sensible models. McCullagh (2002) showed that several absurd models can be produced.

The standard statistical model concept can be extended by implementing some kind of model reduction (Wittgenstein 1961: "The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience") or by, e.g., adjoining a symmetry group to the model (Helland 2004, 2010).

To arrive at methods of inference, the model concept must be supplemented by certain principles. In this connection, an experiment is ordinarily seen as given by a statistical model together with some focus parameter. Most

statisticians agree at least to some variants of the following three principles: The conditionality principle (When you choose an experiment randomly, the information in this large experiment, including the ►randomization, is not more than the information in the selected experiment.), the sufficiency principle (Experiments with equal values of a sufficient statistics have equal information.) and the likelihood principle (All the information about the parameter is contained in the likelihood for the parameter, given the observations.). Birnbaum's famous theorems says that likelihood principle follows from the conditionality principle together with the sufficiency principle (for some precisely defined version of these principles). This, and the principles themselves are discussed in detail in Berger and Wolpert (1984).

Berger and Wolpert (1984) also argue that the likelihood principle "nearly" leads to Bayesian inference, as the only mode of inference which really satisfies the likelihood principle. This whole chain of reasoning has been countered by Kardaun et al. (2003) and by leCam (1984), who states that he prefers to be a little "unprincipled."

To arrive at statistical inference, whether it is point estimates, confidence intervals (credibility intervals for Bayesians) or hypothesis testing, we need some decision theory (see ►Decision Theory: An Introduction, and ►Decision Theory: An Overview). Such decision theory may be formulated differently by different authors. The foundation of statistical inference from a Bayesian point of view is discussed by Good (1988), Lindley (2000) and Savage (1972). From the frequentist point of view it is argued by Efron (1986) that one should be a little more informal; but note that a decision theory may be very useful also in this setting.

The philosophy of foundations of statistics involves many further questions which have direct impact on the theory and practice of statistics: Conditioning, randomization, shrinkage, subjective or objective priors, reference priors, the role of information, the interpretation of probabilities, the choice of models, optimality criteria, nonparametric versus parametric inference, the principle of an axiomatic foundation of statistics etc. Some papers discussing these issues are Cox (1997) with discussion, Kardaun et al. (2003) and Efron (1978, 1979). The last paper takes up the important issue of the relationship between statistical theory and the ongoing revolution of computers.

The struggle between ►Bayesian statistics and frequentist statistics is not so hard today as it used to be some years ago, partly since it has been realized that the two schools often lead to similar results. As hypothesis testing and confidence intervals are concerned, the frequentist school and the Bayesian school must be adjoined by the Fisherian or fiducian school, although largely out of fashion today. The question of whether these three schools in some sense can agree on testing, is addressed by Berger (2003).

The field of design of experiments also has its own philosophical foundation, touching upon practical issues like randomization, blocking and replication, and linked to the philosophy of statistical inference. A good reference here is Cox and Reid (2000).

About the Author

Inge Svein Helland is Professor, University of Oslo (1996–present). He was Head of Department of Mathematical Sciences, Agricultural University of Norway (1987–1989) and Head of Statistics Division, Department of Mathematics, University of Oslo (1999–2001). He was Associate Editor, *Scandinavian Journal of Statistics* (1994–2001). Professor Helland has (co-)authored about 65 papers, reports, manuscripts, including the book *Steps Towards a Unified Basis for Scientific Models and Methods* (World Scientific, Singapore, 2010).

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Decision Theory: An Introduction
►Decision Theory: An Overview
►Foundations of Probability
►Likelihood
►Model Selection
►Philosophy of Probability
►Randomization
►Statistical Inference: An Overview

References and Further Reading

Berger JO (2003) Could fisher, jeffreys and neyman have agreed on testing? Stat Sci 18:1–32
Berger JO, Wolpert RL (1984) The likelihood principle. Lecture Notes, Monograph Series, vol 6, Institute of Mathematical Statistics, Hayward
Breiman L (2001) Statistical modelling: the two cultures. Stat Sci 16:199–231
Cox DR (1997) The current position of statistics: a personal view. Int Stat Rev 65:261–290
Cox DR, Reid N (2000) The theory of design of experiments. Chapman and Hall/CRC, Boca Raton, FL
Efron B (1978) Controverses in the foundations of statistics. Am Matem Month 85:231–246
Efron B (1979) Computers and the theory of statistics: thinking the unthinkable. Siam Rev 21:460–480
Efron B (1986) Why isn't everyone Bayesian? Am Stat 40:1–5
Good IJ (1988) The interface between statistics and philosophy of science. Stat Sci 3:386–412
Helland IS (2004) Statistical inference under symmetry. Int Stat Rev 72:409–422

Helland IS (2010) Steps towards a unified basis for scientific models and methods. World Scientific, Singapore
Kardaun OJWF, Salomé D, Schaafsma W, Steerneman AGM, Willems JC, Cox DR (2003) Reflections on fourteen cryptic issues concerning the nature of statistical inference. Int Stat Rev 71: 277–318
leCam L (1984) Discussion in Berger and Wolpert. 182–185
Lindley DV (2000) The philosophy of statistics. Statistician 49: 293–337
McCullagh P (2002) What is a statistical model? Ann Stat 30: 1225–1310
Savage LJ (1972) The foundations of statistics. Dover, New York
Wittgenstein L (1961) Tractatus Logico-Philosophicus (tr. Pears and McGuinnes). Routledge Kegan Paul, London

Philosophy of Probability

Martin Peterson
Associate Professor
Eindhoven University of Technology, Eindhoven, Netherlands

The probability calculus was created in 1654 by Pierre de Fermat and Blaise Pascal. The philosophy of probability is the philosophical inquiry into the semantic and epistemic properties of this mathematical calculus. The question at the center of the philosophical debate is *what it means to say* that the probability of an event or proposition equals a certain numerical value, or in other words, what the *truth-conditions* for a probabilistic statement are. There is significant disagreement about this, and there are two major camps in the debate, viz., *objectivists* and *subjectivists*.

Objectivists maintain that statements about probability refer to some features of the external world, such as the relative frequency of some event. Probabilistic statements are thus objectively true or false, depending on whether they correctly describe the relevant features of the external world. For example, some objectivists maintain that it is true that the probability that a coin will land heads is ½ if and only if the relative frequency of this type of event is ½ (Other objective interpretations will be discussed below).

Subjectivists disagree with this picture. They deny that statements about probability should be understood as claims about the external world. On their view, they should rather be understood as claims about the speaker's degree of belief that an event will occur. Consider, for example, Mary's probability that her suitor will propose. What is Mary's probability that this event will occur? It seems rather pointless to count the number of marriage proposals that other people get, because this does not tell us anything about the probability that *Mary* will be faced with a marriage proposal. If it is true that Mary's probability is, say, 1/2 then it is true because of her mental state, i.e., her degree of belief that her suitor will propose. Unfortunately, nothing follows from this about whether her suitor actually will propose or not. Mary's probability that her suitor will propose may be high even if the suitor feels that marriage is totally out of the question.

Both the objective and subjective interpretations are compatible with Kolmogorov's axioms. These axioms come out as true, irrespective of whether we interpret them along the lines suggested by objectivists and subjectivists. However, more substantial questions about what the probability of an event is may depend on which interpretation is chosen. For example, Mary's subjective belief that her suitor will propose might be low, although the objective probability is quite high.

The notion of subjective probability is closely related to ►Bayesian statistics, presented elsewhere in this book. (In recent years some authors have, however, also developed objectivist accounts of Bayesian statistics.)

In what follows we shall give a more detailed overview of some of the most well-known objective and subjective interpretations, viz. the relative-frequency view, the propensity interpretation, the logical interpretation, and the subjective interpretation. We shall start, however, with the classical interpretation. It is strictly speaking neither an objective nor a subjective interpretation, since it is salient about many of the key questions discussed by objectivists and subjectivists.

The *classical interpretation*, advocated by Laplace, Pascal, Bernoulli and Leibniz, holds the probability of an event to be a fraction of the total number of possible ways in which the event can occur. Hence, the probability that you will get a six if you roll a six-sided die is 1/6. However, it takes little reflection to realize that this interpretation presupposes that all possible outcomes are equally likely. This is not always a plausible assumption, as Laplace and others were keen to stress. For an extreme example, consider the weather in the Sahara desert. It seems to be much more probable that it will be sunny in the Sahara desert tomorrow than not, but according to the classical interpretation the probability is ½ (since there are two possible outcomes, sun or no sun). Another problem with the classical interpretation is that it is not applicable when the number of possible outcomes is infinite. Then the probability of every possibility would be zero, since the ratio between any finite number and infinity is zero.

The *frequency interpretation*, briefly mentioned above, holds that the probability of an event is the ratio between the numbers of times the event has occurred divided by the total number of observed cases. Hence, if you toss a coin 1,000 times and it lands heads up 508 times then the relative frequency, and thus the probability,

would be 508/1000 = 0.508. A major challenge for anyone seeking to defend the frequency interpretation is to specify which reference class is the relevant one and why. For example, suppose I toss the coin another 1,000 times, and that it lands heads up on 478 occasions. Does this imply that the probability has changed from 0.508 to 0.486? Or is the "new" probability 508 + 486/2.000? The physical constitution of the coin is clearly the same.

The problem of identifying the relevant reference class becomes particularly pressing as the frequency interpretation is applied to unique events, i.e., events that only occur once, such as the US presidential election in 2000 in which George W Bush won over Al Gore. The week before the election the probability was – according to many political commentators – about 50% that Bush would win. Now, to which reference class does this event belong? If this was a *unique* event the reference class has, by definition, just one element, viz., the event itself. So according to the frequency interpretation the probability that Bush was going to win was 1/1, since Bush actually won the election. This cannot be the right conclusion.

Venn famously argued that the frequency interpretation makes sense only if the reference class is taken to be infinitely large. More precisely put, he pointed out that one should distinguish sharply between the underlying *limiting* frequency of an event and the frequency *observed* so far. The limiting frequency is the proportion of successful outcomes *would* get if one were to repeat *one and the same* experiment infinitely many times. So even though the US presidential election in 2000 did *actually* take place just once, we can nevertheless *imagine* what would have happened had it been repeated infinitely many times.

Of course, we cannot actually toss a coin infinitely many times, but we could imagine doing so. Therefore, the limiting frequency is often thought of as an abstraction, rather than as a series of events that take place in the real world. This point has some important philosophical implications. First, it seems that one can never be sure that a limiting relative frequency exists. When tossing a coin, the relative frequency of heads will perhaps never converge towards a specific number. In principle, it could oscillate forever. Moreover, the limiting relative frequency seems to be inaccessible from an epistemic point of view, even in principle. If you observe that the relative frequency of a coin landing heads up is close to 1/2 in a series of ten million tosses, this does not exclude that the true long-run frequency is much lower or higher than 1/2. No finite sequence of observations can prove that the limiting frequency is even close to the observed frequency.

According to the *propensity interpretation*, probabilities should be identified with another feature of the external world, namely the propensity (or disposition or tendency) of an object to give rise to a certain effect. For instance, symmetrical coins typically have a propensity to land heads up about every second time they are tossed, which means that their probability of doing so is about one in two.

The propensity interpretation was developed by Popper in the 1950s. His motivation for developing this view was that it avoids the problem of assigning probabilities to unique events faced by the frequency view. Even an event that cannot take place more than once can nevertheless have a certain propensity (or tendency) to occur. However, Popper's version of the theory also sought to connect propensities with long-run frequencies whenever the latter existed. Thus, his theory is perhaps best thought of as a hybrid between the two views. Contemporary philosophers have proposed "pure" versions of the propensity interpretation, which make no reference what so ever to long-run frequencies.

A well-known objection to the propensity interpretation is Humphreys' paradox. To state this paradox, recall that conditional probabilities can be "inverted" by using ▶Bayes' theorem. Thus, if we know the probability of A given B we can calculate the probability of B given A, given that we know the priors. The point made by Humphreys is that propensities cannot be inverted in this sense. Suppose, for example, that we know the probability that a train will arrive on time at its destination given that it departs on time. Then it makes sense to say that if the train departs on time, it has a propensity to arrive on time at its destination. However, even though it makes sense to speak of the inverted probability, i.e., the probability that the train departed on time given that it arrived on time, it makes no sense to speak of the corresponding inverted propensity. No one would admit that the on-time arrival of the train has a propensity to make it depart on time a few hours earlier.

The thrust of Humphreys' paradox is thus the following: Even though we may not know exactly what a propensity (or disposition or tendency) is, we do know that propensities have a temporal direction. If A has a propensity to give rise to B, then A cannot occur after B. In this respect, propensities function very much like causality; if A causes B, then A cannot occur after B. However, probabilities lack this temporal direction. What happens now can tell us something about the probability of past events, and reveal information about past causes and propensities, although the probability in itself is a non-temporal concept. Hence, it seems that it would be a mistake to identify probabilities with propensities.

The *logical interpretation* of probability was developed by Keynes and Carnap. Its basic idea is that probability is a logical relation between a hypothesis and the evidence

supporting it. More precisely put, the probability relation is best thought of as a generalization of the principles of deductive logic, from the deterministic case to the indeterministic one. For example, if an unhappy housewife claims that the probability that her marriage will end in a divorce is 0.9, this means that the evidence she has at hand (no romantic dinners, etc.) entails the hypothesis that the marriage will end in a divorce to a certain degree, which can be represented by the number 0.9. Coin tossing can be analyzed along the same lines. The evidence one has about the shape of the coin and past outcomes entails the hypothesis that it will land heads up to a certain degree, and this degree is identical with the probability of the hypothesis being true.

Carnap's analysis of the logical interpretation is quite sophisticated, and cannot be easily summarized here. However, a general difficulty with logical interpretations is that they run a risk of being too dependent on evidence. Sometimes we wish to use probabilities for expressing mere guesses that have no correlation whatsoever to any evidence. For instance, I think the probability that it will be sunny in Rio de Janeiro tomorrow is 0.4. This guess is not based on any meteorological evidence. I am just guessing – the set of premises leading up to the hypothesis that it will be sunny is empty; hence, there is no genuine "entailment" going on here. So how can the hypothesis that it will be sunny in Rio de Janeiro tomorrow be entailed to degree 0.4, or any other degree?

It could be replied that pure guesses are irrational, and that it is therefore not a serious problem if the logical interpretation cannot handle this example. However, it is not evident that this is a convincing reply. People do use probabilities for expressing pure guesses, and the probability calculus can easily be applied for checking whether a set of such guesses are coherent or not. If one thinks that the probability for sun is 0.4 it would for instance be correct to conclude that the probability that it will not be sunny is 0.6. This is no doubt a legitimate way of applying the probability calculus. But if we accept the logical interpretation we cannot explain why this is so, since this interpretation defines probability as a *relation* between a (non-empty) set of evidential propositions and a hypothesis.

Let us now take closer look at the *subjective* interpretation. The main idea is that probability is a kind of mental phenomenon. Probabilities are not part of the external world; they are entities that human beings somehow create in their minds. If you claim that the probability for sun tomorrow is, say, 0.8 this merely means that your subjective degree of belief that it will be sunny tomorrow is strong and that the strength of this belief can be represented by the number 0.8. Of course, whether it *actually* will rain tomorrow depends on objective events in the external world, rather than on your beliefs. So it is *probable* that it will rain tomorrow just in case you believe that it will rain to a certain degree, irrespective of what the weather is actually like tomorrow. However, this should not be taken to mean that any subjective degree of belief is a probability. Advocates of the subjective approach stress that for a partial belief to qualify as a probability, one's degrees of belief must be compatible with the axioms of the probability calculus.

Subjective probabilities can vary across people. Mary's degree of belief that it will rain tomorrow might be strong, at the same time as your degree of belief is much lower. This just means that your mental dispositions are different. When two decision makers hold different subjective probabilities, they just happen to believe something to different degrees. It does not follow that at least one person has to be wrong. Furthermore, if there were no humans around at all, i.e., if all believing entities were to be extinct, it would simply be false that some events happen with a certain probability, including quantum-mechanical events. According to the pioneering subjectivist Bruno de Finetti, "Probability does not exist."

Subjective views have been around for almost a century. de Finetti's pioneering work was published in 1931. Ramsey presented a similar subjective theory in a paper written in 1926 and published posthumously in 1931. However, most modern accounts of subjective probability start off from Savage's theory, presented in 1954, which is more precise from a technical point of view. The key idea in all three accounts is to introduce an ingenious way in which subjective probabilities can be measured. The measurement process is based on the insight that the degree to which a decision maker believes something is closely linked to his or her behavior. Imagine, for instance, that we wish to measure Mary's subjective probability that the coin she is holding in her hand will land heads up the next time it is tossed. First, we ask her which of the following very generous options she would prefer.

(a) "If the coin lands heads up you win a trip to Bahamas; otherwise you win nothing"
(b) "If the coin *does not* land heads up you win a trip to Bahamas; otherwise you win nothing"

Suppose Mary prefers A to B. We can then safely conclude that she thinks it is *more probable* that the coin will land heads up rather than not. This follows from the assumption that Mary prefers to win a trip to Bahamas rather than nothing, and that her preference between uncertain prospects is entirely determined by her beliefs and desires with respect to her prospects of winning the trip to Bahamas. If she on the other hand prefers B to A, she thinks it is *more probable* that the coin will not land heads

up, for the same reason. Furthermore, if Mary is indifferent between A and B, her subjective probability that the coin will land heads up is exactly 1/2. This is because no other probability would make both options come out as equally attractive, irrespective of how strongly she desires a trip to Bahamas, and irrespective of which decision rule she uses for aggregating her desires and beliefs into preferences.

Next, we need to generalize the measurement procedure outlined above such that it allows us to always represent Mary's degree of belief with precise numerical probabilities. To do this, we need to ask Mary to state preferences over a *much larger* set of options and then *reason backwards*. Here is a rough sketch of the main idea: Suppose that Mary wishes to measure her subjective probability that her etching by Picasso worth $20,000 will be stolen within one year. If she considers $1,000 to be a fair price for insuring her Picasso, that is, if that amount is the highest price she is prepared to pay for a gamble in which she gets $20,000 if the event *S*: "The Picasso is stolen within a year" takes place, and nothing otherwise, then Mary's subjective probability for *S* is $\frac{1,000}{20,000} = 0.05$, given that she forms her preferences in accordance with the principle of maximizing expected monetary value. If Mary is prepared to pay up to $2,000 for insuring her Picasso, her subjective probability is $\frac{2,000}{20,000} = 0.1$, given that she forms her preferences in accordance with the principle of maximizing expected monetary value.

Now, it seems that we have a general method for measuring Mary's subjective probability: We just ask her how much she is prepared to pay for "buying a contract" that will give her a fixed income if the event we wish to assign a subjective probability to takes place. The highest price she is prepared to pay is, by assumption, so high that she is indifferent between paying the price and not buying the contract. (This assumption is required for representing probabilities with precise numbers; if buying and selling prices are allowed to differ we can sometimes use intervals for representing probabilities. See e.g., Borel and Baudain 1962 and Walley 1991.)

The problem with this method is that very few people form their preferences in accordance with the principle of maximizing expected monetary value. Most people have a decreasing marginal utility for money. However, since we do not know anything about Mary's utility function for money we cannot replace the monetary outcomes in the examples with the corresponding utility numbers. Furthermore, it also makes little sense to *presuppose* that Mary uses a specific decision rule, such as the expected utility principle, for forming preferences over uncertain prospects. Typically, we do not know anything about how people form their preferences.

Fortunately, there is a clever solution to all these problems. The main idea is to impose a number of structural conditions on preferences over uncertain options. The structural conditions, or axioms, merely restrict what *combinations* of preferences it is legitimate to have. For example, if Mary strictly prefers option A to option B in the Bahamas example, then she must not strictly prefer B to A. Then, the subjective probability function is established by reasoning backwards while taking the structural axioms into account: Since the decision maker preferred some uncertain options to others, and her preferences over uncertain options satisfy a number of structural axioms, the decision maker behaves *as if* she were forming her preferences over uncertain options by first assigning subjective probabilities and utilities to each option, and thereafter maximizing expected utility. A peculiar feature of this approach is, thus, that probabilities (and utilities) are derived from "within" the theory. The decision maker does not prefer an uncertain option to another *because* she judges the subjective probabilities (and utilities) of the outcomes to be more favorable than those of another. Instead, the well-organized structure of the decision maker's preferences over uncertain options logically imply that they can be described *as if* her choices were governed by a subjective probability function and a utility function, constructed such that a preferred option always has a higher expected utility than a non-preferred option. These probability and utility functions need not coincide with the ones outlined above in the Bahamas example; all we can be certain of is that there exist *some* functions that have the desired technical properties.

Cross References

►Axioms of Probability
►Bayes' Theorem
►Bayesian Statistics
►Foundations of Probability
►Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Philosophical Foundations of Statistics
►Probability Theory: An Outline
►Probability, History of

References and Further Reading

Borel E, Baudain M (1962) Probabilities and life. Dover Publishers, New York

de Finetti B (1931/89) Probabilism: a critical essay on the theory of probability and on the value of science. Erkenntnis 31:169–223 (trans: de Finetti B (1931) Probabilismo. Logos 14:163–219)

Humphreys P (1985) Why propensities cannot be probabilities. Philos Rev 94:557–570

Jeffrey R (1983) The logic of deciosion, 2nd edn. (significant improvements from 1st edn). University of Chicago Press, Chicago
Kreps DM (1988) Notes on the theory of choice, Westview Press, Boulder
Laplace PS (1814) A philosophical essay on probabilities (English transl 1951). Dover Publications Inc, New York
Mellor DH (1971) The Matter of Chance. Cambridge University Press, Cambridge
Popper K (1957) The propensity interpretation of the calculus of probability and the quantum theory. In: Körner S (ed) The colston papers, vol 9, pp 65–70
Ramsey FP (1926) Truth and probability in Ramsey, 1931. In: Braithwaite RB (ed) The foundations of mathematics and other logical essays, Ch. VII, Kegan, Paul, Trench, Trubner & Co, London, Harcourt, Brace and Company, New York, pp 156–198
Savage LJ (1954) The foundations of statistics. Wiley, New York. 2nd edn. 1972, Dover
Walley P (1991) Statistical reasoning with imprecise probabilities, monographs on statistics and applied probability. Chapman & Hall, London

Point Processes

David Vere-Jones
Professor Emeritus
Victoria University of Wellington, Wellington, New Zealand

"Point Processes" are locally finite (i.e., no finite accumulation points) families of events, typically occurring in time, but often with additional dimensions (marks) to describe their locations, sizes and other characteristics.

The subject originated in attempts to develop life-tables. The first such studies included one by Newton and another by his younger contemporary Halley.

Newton's study was provoked by his life-long religious concerns. In his book, "Chronology of Ancient Kingdoms Amended" he set out to estimate the dates of various biblical events by counting the number of kings or other rulers between two such biblical events and allowing each ruler a characteristic length of reign. The value he used seems to have been arrived at by putting together all the observations from history that he could find (he included rulers from British, classical, European and biblical histories and even included Methuselah's quoted age among his data points) and taking some average of their lengths of rules, but he acknowledged himself that his methods were informal and personal.

Halley, by contrast, was involved in a scientific exercise, namely the development of actuarial tables for use in calculating pensions and annuities. Indeed, he was requested by the newly established Royal Society to prepare such a table from records in Breslau, a city selected because it had been less severely affected by the plague than most of the larger cities in Europe, so that its mortality data were felt more likely to be typical of those of cities and periods in normal times.

Another important early stimulus for point process studies was the development of telephone engineering. This application prompted Erlang's early studies of the Poisson process (see ▶Poisson Processes), which laid down many of the concepts and procedures, such as ▶renewal processes, and forward and backward recurrence times, subsequently entering into point process theory. It was also the context of Palm's (1943) deep studies of telephone-traffic issues. Palm was the first to use the term "point process" (Punkt-Prozesse) itself.

A point process can be treated in many different ways: for example, as a sequence of delta functions, as a sequence of time intervals between event occurrences; as an integer-valued random measure; or as the sum of a regular increasing component and a jump-type martingale.

Treating each point as a delta-function in time yields a time series which has generalized functions as realizations, but in other respects has many similarities with a continuous time series. In particular, stationarity, ergodicity, and a "point process spectrum" can be developed through this approach.

If attention is focused on the process of intervals between successive points, the paradigm example is the renewal process, where the intervals between events are independent, and, save possibly for the first interval, identically distributed.

Counting the number of events, say $N(A)$ falling into a pre-specified set A (an interval, or more general Borel set), leads to treating the point process as an integer-valued random measure.

An important underlying theorem, first enunciated and analyzed by Slivnyak (1962), asserts the equivalence of the counting and interval based approaches.

Random measures form a generalization of point processes which are of considerable importance in their own right. Their first and second order moment measures

$$M_1(A) = E[N(A)]$$
$$M_2(A \times B) = E[N(A)N(B)]$$

and the associated signed measure, the covariance measure

$$C_2(A \times B) = M_2(A \times B) - M_1(A)M_1(B)$$

form non-random measures which have been extensively studied.

A third point of view originated more recently from martingale ideas applied to point processes, and has proved a rich source of both new models and new methods of analysis. The key here is to define the point process in terms of its conditional intensity (or conditional hazard) function, representing the conditional rate of occurrence of events, given the history (record of previously occurring events and any other relevant information) up to the present.

These ideas are closely linked to the martingale representation of point processes. This takes the form of an increasing step function, with unit steps at the time points when the events occur, less a continuous increasing part given by the integral of the conditional intensity.

The Hawkes' processes [introduced by Hawkes (1971a, b)] form an important class of point processes introduced and defined by their conditional intensities, which take the general form

$$\lambda(t) = \mu + \sum_{i:t_i - t} g(t - t_i)$$

where μ is a non-negative constant ("the arrival rate") and g is a non-negative integrable function ("the infectivity function").

The archetypal point process is the simple Poisson process, where the intervals between successive events in time are independent, and, (with the possible exception of the initial interval) are identically and exponentially distributed with a common mean, say m. For this process, the conditional intensity is constant, equal to the rate of occurrence (intensity) of the Poisson process, here $1/m$. Processes with continuous conditional intensities are sometimes referred to as "processes of Poisson type" since they behave locally like Poisson processes over intervals small enough for the conditional intensity to be considered approximately constant.

For the Poisson process itself, and also Poisson cluster processes, where cluster centers follow a simple Poisson process, and the clusters are independent subprocesses, identically distributed relative to their cluster centers, it is possible to write down simple expressions for the characteristic functional

$$\Phi[h] = E\left[e^{i\int h(t)dN(t)}\right]$$

where the carrying function h is integrable against M_1, or the essentially equivalent probability generating functional

$$G[\xi] = E[\Pi\xi(t_i)]$$

where ξ plays the role of h in the characteristic functional.

For example, the Poisson process with continuous intensity m has probability generating functional

$$G[h] = \exp\left\{-m\int [1 - h(u)]du\right\}.$$

Such functionals provide a comprehensive summary of the process and its attributes, especially the moment structure.

However, the usefulness of characteristic or generating functionals in practice is restricted by the relatively few examples for which they can be obtained in convenient closed form. Nevertheless, where available, their form is essentially independent of the space (phase space) in which the points are located. By contrast, finding extensions of the conditional intensity to spaces of more than one dimension has proved extremely difficult, on account of the absence of a clear linear ordering, a problem which equally affects other attempts to extend martingale ideas to higher dimensions.

About the Author

Dr. David Vere-Jones is Emeritus Professor, Victoria University of Wellington (since 2000). He is Past President, New Zealand Statistical Association (1981–1983), Past President of Interim Executive of International Association for Statistical Education (1991–1992), Founding President of the New Zealand Mathematical Society (1974). He received the Rutherford Medal, New Zealand's top science award, in 1999 for "outstanding and fundamental contributions to research and education in probability, statistics and the mathematical sciences, and for services to the statistical and mathematical communities, both within New Zealand and internationally." Professor Vere-Jones was also awarded the NZ Science and Technology Gold Medal (2000), and the NZSA Campbell Award for 2009, in recognition for his contributions to the statistical sciences. He has (co-)authored over 100 refereed publications, and three books. In 2001 (April 19–21), a Symposium in Honor of David Vere-Jones on the Occasion of His 65th Birthday was held in Wellington.

"David Vere-Jones is New Zealand's leading resident mathematical statistician. He has made major contributions to the theory of statistics, its applications, and to the teaching of statistics in New Zealand. He is highly regarded internationally and is involved in numerous international activities. One of his major research areas has been concerned with Point Processes... . A substantial body of the existing theory owes its origins to him, either directly or via his students. Of particular importance and relevance to New Zealand is his pioneering work on the applications of point process theory to seismology" (NZMS Newsletter 24, August 1982).

Cross References

►Khmaladze Transformation
►Martingales
►Non-Uniform Random Variate Generations
►Poisson Processes
►Renewal Processes
►Spatial Point Pattern
►Spatial Statistics
►Statistics of Extremes
►Stochastic Processes

References and Further Reading

Cox DR, Isham V (1980) Point Processes. Chapman and Hall, London
Cox DR, Lewis PAW (1966) The statistical analysis of series of events. Methuen, London
Daley DJ, Vere-Jones D (1988) An introduction to the theory of point processes, 1st edn. Springer, New York; 2nd. edn. (2002) vols 1 and 2, Springer, New York
Hawkes AG (1971a) Spectra of some self-exciting and mutually exciting point processes. Biometrika 58:83–90
Hawkes AG (1971b) Point spectra of some mutually exciting point processes. J R Stat Soc 33:438–443
Palm C (1943) Intensitasschwankungen im fernsprechverkehr. Ericsson Technics 44:1–189
Slivnyak IM (1962) Some properties of stationary flows of homogeneous random events. Teor. Veroyantnostei I Primen 7:347–352, Translation in Theory of Probability and Applications, 7:36–341
Stoyan D, Stoyan H (1994) Fractals, random shapes and point fields. Wiley, Chichester
Stoyan D, Kendall WS, Mecke J (1995) Stochastic geometry. 2nd edn. Wiley, Chichester; 1st edn. Academie, Berlin, 1987

Poisson Distribution and Its Application in Statistics

Lelys Bravo de Guenni
Professor
Universidad Simón Bolívar, Caracas, Venezuela

The Poisson distribution was first introduced by the French Mathematician Siméon-Denis Poisson (1781–1840) to describe the probability of a number of events occurring in a given time or space interval, with the probability of occurrence of these events being very small. However, since the number of trials is very large, these events do actually occur.

It was first published in 1837 in his work *Recherches sur la probabilité des jugements en matières criminelles et matière civile* (Research on the Probability of Judgments in Criminal and Civil Matters). In this work, the behavior of certain random variables X that count the number of occurrences (or *arrivals*) of such events in a given interval in time or space was described. Some examples of these events are infant mortality in a city, the number of misprints in a book, the number of bacteria on a plate, the number of activations of a geiger counter, and so on.

Assuming that λ is the expected value of such arrivals in a time interval of fixed length, the probability of observing exactly k events is given by the probability mass function

$$f(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

for $k = 0, 1, 2, \ldots$. The parameter distribution λ is a positive real number, which represents the average number of events occurring during a fixed time interval. For example, if the event occurs on average three times per second, in 10 s the event will occur on average 30 times and $\lambda = 30$. When the number of trials n is large and the probability of occurrence of the event λ/n approaches to zero, the ►binomial distribution with parameters n and $p = \lambda/n$ can be approximated to a Poisson distribution with parameter λ. The binomial distribution gives the probability of x successes in n trials.

If X is a random variable with a Poisson distribution, the expected value of X and the variance of X are both equal to λ. To estimate λ by maximum likelihood, given a sample $k_1, k_2, \ldots, k_n$, the log-likelihood function

$$L(\lambda) = \log \prod_{i=1}^{n} \frac{\lambda_i^k e^{-\lambda}}{k_i!}$$

is maximized with respect to λ and the resulting estimate is $\hat{\lambda} = \frac{\sum_{i=1}^{n} k_i}{n}$. This is an unbiased estimator since the expected value of each k_i is equal to λ; and it is also an efficient estimator since its estimator variance achieves the Cramer–Rao lower bound. From the Bayesian inference perspective, a conjugate prior distribution for the parameter λ is the Gamma distribution. Suppose that λ follows a Gamma prior distribution with parameters α and β, such that

$$p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad \lambda > 0$$

If a sample $k_1, k_2, \ldots, k_n$ of size n is observed, the posterior probability distribution for λ is given by

$$p(\lambda|k_1, k_2, \ldots, k_n) \sim Gamma\left(\alpha + \sum_{i=1}^{n} k_i, \beta + n\right).$$

When $\alpha \to$ and $\beta \to 0$, we have a diffuse prior distribution, and the posterior expected value of λ ($E[\lambda|k_1, k_2, \ldots, k_n]$) approximates to the maximum likelihood estimator $\hat{\lambda}$.

A use for this distribution was not found until 1898, when an individual named Bortkiewicz (O'Connor and

Robertson 1950) was asked by the Prussian Army to investigate accidental deaths of soldiers attributed to being kicked by horses. In 1898, he published *The Law of Small Numbers*. In this work he was the first to note that events with low frequency in a large population followed a Poisson distribution even when the probabilities of the events varied. Bortkiewicz studied the distribution of 122 men kicked to death by horses among 14 Prussian army corps over 20 years. This famous data set has been used in many statistical textbooks as a classical example on the use of the Poisson distribution (see Yule and Kendall [1950] or Fisher [1954]). He found that in about half of every army corps–year combination, there were no deaths for horse kicking. For other combinations of corps–years, the number of deaths were from 1 to 4. Although the probability of horse kick deaths might vary from corps and years, the overall observed frequencies were very close to the expected frequencies estimated by using a Poisson distribution.

In epidemiology, the Poisson distribution has been used as a model for deaths. In one of the oldest textbooks of statistics published by Bowley in 1901 (cited by Hill [2002]), he fitted a Poisson distribution to deaths from splenic fever in the years 1875–1894 and showed a reasonable agreement with the theory. At that time, splenic fever was a synonym for present-day anthrax.

A more extensive use of the Poisson distribution can be found within the Poisson generalized linear models, usually called the *Poisson regression models* (see ►Poisson Regression). These models are used to model count data and contingency tables. For contingency tables, the Poisson regression model is best known as the *log-linear model*. In this case, the response variable Y has a Poisson distribution and usually, the logarithm of its expected value ($E[Y]$) is expressed as a linear predictor $X\beta$ where X is a $n \times p$ matrix of explanatory variables and β is a parameter vector of size p. In this case, the *link* function $g(.)$, which relates the expected value of the response variable Y with the linear predictor is the logarithm function, in such a way that the mean of the response variable $\mu = g^{-1}(X\beta)$.

Poisson regression can also be used to model what is called the *relative risk*. This is the ratio between the counts and an *exposure* factor E. For example, to model the relative risk of disease in a region, we make $\eta = Y/E$, where Y is the observed number of cases and E is the expected number of cases, which depends on the number of persons at risk. The usual model for Y is

$$Y|\eta \sim \text{Poisson}(E\eta)$$

where η is the true relative risk of disease (Banerjee et al. (2004)). When applying the Poisson model to data, the main assumption is that the variance is equal to the mean. In many cases, this may not be assumed since the variance of counts are usually greater than the mean. In this case, we have *overdispersion*. One way to deal with this problem is to use the negative binomial distribution, which is a two-parameter family that allows the mean and variance to be fitted separately. In this case, the mean of the Poisson distribution λ is assumed a random variable as drawn from a Gamma distribution.

Another common problem with Poisson regression is excess zeros: if there are two processes at work, one determining whether there are zero events or any events, and a Poisson process (see ►Poisson Processes) determining how many events there are, there will be more zeros than a Poisson regression would predict. An example would be the distribution of cigarettes smoked in an hour by members of a group where some individuals are nonsmokers. These data sets can be modeled as *zero inflated Poisson models*, where p is the probability of observing zero counts, and $1 - p$ is the probability of observing a count variable modeled as a Poisson(λ).

About the Author

For biography *see* the entry ►Normal Scores.

Cross References

►Contagious Distributions
►Dispersion Models
►Expected Value
►Exponential Family Models
►Fisher Exact Test
►Geometric and Negative Binomial Distributions
►Hypergeometric Distribution and Its Application in Statistics
►Modeling Count Data
►Multivariate Statistical Distributions
►Poisson Processes
►Poisson Regression
►Relationships Among Univariate Statistical Distributions
►Spatial Point Pattern
►Statistics, History of
►Univariate Discrete Distributions: An Overview

References and Further Reading

Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis of spatial data. Chapman and Hall/CRC, Boca Raton

Fisher RA (1954) Statistical methods for research workers. Oliver and Boyd, Edinburgh

Hill G (2002) Horse kicks, antrax and the poisson model for deaths. Chronic Dis Can 23(2):77

O'Connor JJ, Robertson EF (1950) http://www-groups.dcs.st-and.ac.uk/history/Mathematicians/Bortkiewicz.html

Yule GU, Kendall MG (1950) An introduction to the theory of statistics. Charles Griffin, London

Poisson Processes

MR Leadbetter
Professor
University of North Carolina, Chapel Hill, NC, USA

Introduction

Poisson Processes are surely ubiquitous in the modeling of point events in widely varied settings, and anything resembling a brief exhaustive account is impossible. Rather we aim to survey several "Poisson Habitats" and properties, with glimpses of underlying mathematical framework for these processes and close relatives. We refer to three (of many) authoritative works (Cox and Lewis 1966; Daley and Vere-Jones 1988; Kallenberg 1986) for convenient detailed accounts of Poisson Process and general related theory, tailored to varied mathematical tastes.

In this entry we first consider Poisson Processes in their classical setting as series of random events (▶point processes) on the real line (e.g., in time), their importance in one dimension for stochastic modeling being rivaled only by the Wiener Process - both being basic (for different purposes) in their own right, and as building blocks for more complex models. Classical applications of the Poisson process abound - exemplified by births, radioactive disintegrations, customer arrival times in queues, instants of new cases of disease in an epidemic, and (a favorite of ours), the crossings of a high level by a stationary process. The classical format for Poisson Processes will be described in section "▶Poisson Processes on the Real Line", and important variants in section "▶Important Variants".

Unlike many situations in which generalizations to more than one dimension seem forced, the Poisson process has important and natural extensions to two and higher dimensions, as indicated in section "▶Poisson Processes in Higher Dimensions". Further, an even more attractive feature (at least to those with theoretical interests) is the fact that Poisson Processes can be defined on spaces with very little structure, as indicated in section "▶Poisson Processes on Abstract Spaces".

Poisson Processes on the Real Line

A *Point Process* on the real line is simply a sequence of events occurring in time (or some other 1-dimensional (e.g., distance) parameter) according to a probabilistic mechanism. One way to describe the probability structure is to define a sequence $0 \le \tau_1 < \tau_2 < \tau_3 \ldots < \infty$ where the τ_i are the "times" of occurrences of interest ("events" of the point process). They are assumed to be random variables (written here as distinct, i.e., strictly increasing, when the point process is termed "simple" but successive τ_i can be taken to be equal if "multiple events" are to be considered.) It is assumed that the points τ_i tend to infinity as $i \to \infty$ so that there are no "accumulation points". One may define a point process as a *random set* $\{\tau_i : i = 1, 2, \ldots\}$ of such points (cf Ryll-Nardzewski 1961). Alternatively the occurrence times τ_i are a family of random variables for $i = 1, 2, \ldots$ and may be discussed within the framework of *random sequences* (discrete parameter stochastic processes). However often the important quantities for a point process on the real line are the random variables $N(B)$ which are the (random) numbers of events (τ_i) in sets B of interest (usually Borel sets). When $B = (0, t]$ we write N_t for $N((0, t])$, i.e., the number of events occurring from time zero up to and including time t. $\{N_t\}$ is thus a family of random variables for positive t – or a non-negative integer-valued continuous parameter stochastic process on the positive real line. Likewise $\{N(B)\}$ defines a non-negative integer-valued stochastic process indexed by the (Borel) sets B (finite for bounded B).

Here (as is customary) we focus on the "counting" r.v.s $\{N_t\}$ or $\{N(B)\}$ rather than the consideration of more geometric properties of the sets $\{\tau_i\}$. Note that these two families are essentially equivalent since knowledge of $N(B)$ for each Borel set determines that of the sub-family $\{N_t\}(B = (0, t])$ and the converse is also true since $N(B)$ is a measure defined on the Borel sets and is determined by its values on the intervals $(0, t]$. Further their distributions are determined by those of the occurrence times τ_k and conversely, in view of equality of the events $(\tau_k > t)$, $(N_t < k)$.

We (finally!) come to the subject of this article – the Poisson Process. In its simplest context this is defined as a family $N(B)$, (or N_t) as above on the positive real line by the requirement that each N_t be Poisson, $P\{N_t = r\} = e^{-\lambda t}(\lambda t)^r/r!$, $r = 0, 1, 2, \ldots$ and that "increments" $(N_{t_2} - N_{t_1})$, $(N_{t_4} - N_{t_3})$, are independent for $0 < t_1 < t_2 \le t_3 < t_4$. Equivalently $N(B)$ is Poisson with mean $\lambda m(B)$ where m denotes Lebesgue measure, and $N(B_1)$, $N(B_2)$ are independent for disjoint B_1, B_2. $\lambda = \mathcal{E}N_1$ is the expected number of events per unit time, or the *intensity* of the Poisson process.

That the Poisson – rather than some other – distribution plays a central role is due in part to the long history of its use to describe rare events – such as the classical number of deaths by horse kicks in the Prussian army. But more significantly the very simplest modeling of a point process would surely require independence of increments which holds as noted above for the Poisson Process. Further this process has the stationarity property that the distribution of the "increment" $(N_{t+h} - N_t)$ depends only on the length

h of the interval, not on its starting point t. Moreover the probability of an event in a small interval of length h is approximately λh and the probability of more than one in that interval is of smaller order, i.e.,

$$P\{(N_{t+h} - N_t) = 1\} = \lambda h + o(h),$$
$$P\{(N_{t+h} - N_t) \geq 2\} = o(h) \text{ as } h \to o.$$

It turns out that the Poisson Process as defined is the *only* point process exhibiting stationarity and these two latter properties [see e.g., Durrett (2005) for proof] which demonstrates what Kingman aptly describes as the "inevitability" of the Poisson distribution, in his volume (Kingman 1993).

The Poisson Process has extensive properties which are well described in many works [e.g., Cox and Lewis (1966) and Daley and Vere-Jones (1988)] For example the equivalence of the events $(\tau_k > t)$, $(N_t < k)$ noted above readily shows that the inter-arrival times $(\tau_k - \tau_{k-1})$ are i.i.d. exponential random variables with mean $\lambda^{-1}(\tau_0 = 0)$, and τ_k itself is the sum of the first k of these, thus distributed as $(2\lambda)^{-1}\chi^2_{2k}$. On the other hand we have the famous apparent paradox that the random interval which contains a fixed point t_0 is distributed as the sum of two independent such exponential variables – the time from the preceding event plus that to the following event. This and a host of other useful properties may be conveniently found in Cox and Lewis (1966) and Daley and Vere-Jones (1988). Finally, the above discussion has focused on Poisson Processes on the positive real line. It is a simple matter to add an independent Poisson Process on the negative real line to obtain one on the entire real line $(-\infty, \infty)$.

Important Variants

The stationarity requirement of a constant intensity λ may be generalized to include a time varying intensity function $\lambda(t)$ for which the number of events $N(B)$ in a (Borel) set B is still Poisson but with mean $\Lambda(B) = \int_B \lambda(t)dt$, keeping independence of $N(B_1)$, $N(B_2)$ for disjoint B_1, B_2. Then N_t is Poisson with mean $\int_0^t \lambda(t)dt$. More generally one may take Λ to be a "measure" on the Borel sets but not necessarily of this integral (absolutely continuous) form which unlike the simple Poisson Process above, does not necessarily prohibit the occurrence of more than one event at the same instant, (multiple events) and may allow positive probability of an event occurring at a given fixed time point. Further one may consider random versions of the intensity e.g., with $\lambda(t)$ being itself a stochastic process ("*stochastic intensity*") as for example the blood pressure of an individual (varying randomly in time) leading to (conditionally) Poisson chest pain incidents. The resulting point processes are termed *doubly stochastic Poisson* or *Cox Processes*, and are widely used in medical trials e.g., of new treatments. For other widely used variants of Poisson Processes (e.g., "Mixed" and "Compound" Poisson processes) as well as extensive theory of point process properties, we recommend the very readable volume (Daley and Vere-Jones (1988)).

Poisson Processes in Higher Dimensions

Point processes (especially Poisson) have also been traditionally very useful in modeling point events in space and space-time dimensions. The locations of ore deposits in two or three spatial dimensions and the occurrences of earthquakes in two dimensions and perhaps time ("spatio-temporal") are important examples. The mathematical framework extends naturally from one dimension, $N(B)$ being the number of point events in the two- or 3-dimensional (Borel) set B, and notational extensions such as $N_t(B)$ for the number of events in a spatial set B up to time t.

Not infrequently a time parameter is considered simply as equivalent to the addition of just one more spatial dimension, but the obvious differences in the questions to be asked for space and time suggest that the notation reflect the different character of the parameters. Further natural dependence structure (correlation assumptions, mixing conditions, long range dependence) may differ for spatial and time parameters. Further "*coordinatewise mixing*" (introduced in Leadbetter and Rootzen (1998)) seems promising in current investigation to facilitate point process theory in higher dimensions, where the parameters have different roles. A reader interested in the theory and applications in higher dimensions should consult (Daley and Vere-Jones 1988) and the wealth of references therein.

Poisson Processes on Abstract Spaces

There is substantial development of point process (and more general "random measure") theory in more abstract spaces, usually with an intricate topological structure (see Kallenberg (1986)). However for discussion of existence and useful basic modeling properties, the topological assumptions are typically solely used for definition of a natural simple measure-theoretic structure without any necessary underlying topology - though useful for deeper considerations such as weak convergence, beyond pure modeling. Further a charming property of Poisson processes in particular is that they may be defined simply on spaces with very little structure as we now indicate.

Specifically let S be a space, and $\mathcal{S}$ a σ-ring (here a σ-field for simplicity) of subsets of S. For a givenprobability

space $(\Omega, \mathcal{F}, P)$, a *random measure* is defined to be any family of non-negative- (possibly infinite) valued random variables $N_\omega(B)$ for each $B \in \mathcal{S}$ which is a measure (countably additive) on $\mathcal{S}$ for each $\omega \in \Omega$. A point process is a random measure for which each $N_\omega(B)$ is integer-valued (or $+\infty$). In this very general context one may construct a Poisson Process (see ►Poisson Processes) by simple steps (cf Kallenberg (1986) and Kingman (1993)) which we indicate. Define i.i.d. random elements $\tau_1, \tau_2, \ldots, \tau_n$ on S for each positive integer n with common distribution $\nu = P\tau_j^{-1}$ yielding a point process on S consisting of a finite number (n) of points. By regarding n as random having a Poisson distribution with mean $a > 0$ (or mixing the (joint) distributions of this point process with Poisson weights) one obtains a finite valued Poisson process $\{N(B)\}$ with the finite intensity measure $\mathcal{E}N(B) = a\nu(B)$. Finally if λ is a σ-finite measure on $\mathcal{S}$ we may write $\mathcal{S} = \bigcup_1^\infty S_i$ where $\lambda(S_i) < \infty$. Let $\{N_i(B), i = 1, 2, \ldots\}$ be point processes with the finite intensity measures $\mathcal{E}N_i(B) = \lambda(B \cap S_i)$. The superposition of these point processes gives a Poisson Process with intensity λ.

Relatives of the Poisson Process such as those above (Mixed, Compound, Doubly Stochastic...) may be constructed in a similar way to the one-dimensional framework. These sometimes require small or modest additional assumptions about the space S such as measurability of singleton sets, and the separation of two of its points by measurable sets. One may also obtain many general results analogous results to those of one dimension by assuming the existence of a *countable semiring* which covers S, and plays the role of bounded sets on which the point process is finite-valued, in this general non-topological context. Finally, as noted, the reference Kingman (1993) gives an account of Poisson processes primarily in this general framework, along with the basic early paper (Kendall 1974). In a topological setting the monograph (Kallenberg 1986) gives a comprehensive development of random measures, motivating our own non-topological approach.

About the Author

M.R. Leadbetter obtained his B.Sc. (1953), M.Sc. (1954), University of New Zealand, and Ph.D. (1963), UNC-Chapel Hill. He is a Fellow of American Statistical Association, Institute of Mathematical Statistics and member of the International Statistical Institute. He has published a book with Harald Cramér, *Stationary and Related Stochastic Processes* (Wiley, 1967). Professor Leadbetter was awarded an Honorary Doctorate, from Lund University. His name is associated with the Theorem of Cramér and Leadbetter.

Cross References

►Erlang's Formulas
►Lévy Processes
►Markov Processes
►Non-Uniform Random Variate Generations
►Point Processes
►Poisson Distribution and Its Application in Statistics
►Probability on Compact Lie Groups
►Queueing Theory
►Radon–Nikodým Theorem
►Renewal Processes
►Spatial Point Pattern
►Statistical Modelling in Market Research
►Stochastic Models of Transport Processes
►Stochastic Processes
►Stochastic Processes: Classification
►Univariate Discrete Distributions: An Overview

References and Further Reading

Cox DR, Lewis PAW (1966) The statistical analysis of series of events. Methuen Monograph, London
Daley D, Vere-Jones D (1988) An introduction to the theory of point processes. Springer, New York
Durrett R (2005) Probability: theory and examples, 3rd edn. Duxbury, Belmont, CA
Kallenberg O (1986) Random measures, 4th edn. Academic, New York
Kendall DG (1974) Foundations of a theory of random sets. In: Kendall DG, Harding EF (eds) Stochastic geometry. Wiley, London
Kingman JFC (1993) Poisson processes. Clarendon, Oxford
Leadbetter MR, Rootzen H (1998) On extreme values in stationary random fields. In: Karatzas et al. (eds) Stochastic processes and related topics, volume in memory of Stamatis Cambanis, Birkhäuser
Ryll-Nardzewski C (1961) Remarks on processes of calls. Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability 2:455–465

Poisson Regression

Gerhard Tutz
Professor
Ludwig-Maximilians-Universität München, Germany

Introduction

The Poisson regression model is a standard model for count data where the response variable is given in the form of event counts such as the number of insurance claims within a given period of timeor the number of cases with a

specific disease in epidemiology. Let $(Y_i, \boldsymbol{x}_i)$ denote n independent observations, where $\boldsymbol{x}_i$ is a vector of explanatory variables and Y_i is the response variable. It is assumed that the response given $\boldsymbol{x}_i$ follows a Poisson distribution which has probability function

$$P(Y_i = r) = \begin{cases} \frac{\lambda_i^r}{r!} e^{-\lambda_i} & \text{for } r \in \{0, 1, 2, \dots\} \\ 0 & \text{otherwise.} \end{cases}$$

Mean and variance of the Poisson distribution are given by $E(Y_i) = var(Y_i) = \lambda_i$. Equality of the mean and variances is often referred to as the *equidispersion property* of the Poisson distribution. Thus, in contrast to the normal distribution, for which mean and variance are unlinked, the Poisson distribution implicitly models stronger variability for larger means, a property which is often found in real life data. The support of the Poisson distribution is $0, 1, 2, \dots$, which makes it an appropriate distribution model for count data.

A Poisson regression model assumes that the conditional mean $\mu_i = E(Y_i|\boldsymbol{x}_i)$ is determined by

$$\mu_i = h\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) \quad \text{or equivalently} \quad g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where g is a known link function and $h = g^{-1}$ denotes the response function. Since the Poisson distribution is from the simple exponential family the model is a *generalized linear model* (GLM, see ▶Generalized Linear Models). The most widely used model uses the canonical link function by specifying

$$\mu_i = \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) \quad \text{or} \quad \log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Since the logarithm of the conditional mean is linear in the parameters the model is called a *log-linear* model. The log-linear version of the model is particularly attractive because interpretation of parameters is very easy. The model implies that the conditional mean given $\boldsymbol{x}^T = (x_1, \dots, x_p)$ has a multiplicative form given by

$$\mu(\boldsymbol{x}) = \exp(\boldsymbol{x}^T \boldsymbol{\beta}) = e^{x_1 \beta_1} \dots e^{x_p \beta_p}.$$

Thus e^{β_j} represents the multiplicative effect on $\mu(\boldsymbol{x})$ if variable x_j changes by one unit to $x_j + 1$.

Inference

Since the model is a generalized linear model inference may be based on the methods that are available for that class of models (see for example McCullagh and Nelder 1989). One obtains for the derivative of the log-likelihood, which is the so-called score function

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \frac{h'\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)}{h(\boldsymbol{x}_i \boldsymbol{\beta})} \left(y_i - h\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right),$$

and the Fisher matrix $F(\boldsymbol{\beta}) = E(-\partial h / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \frac{h'(\boldsymbol{x}_i^T \boldsymbol{\beta})^2}{h(\boldsymbol{x}_i^T \boldsymbol{\beta})}$. Under regularity conditions, $\hat{\boldsymbol{\beta}}$ defined by $s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ is consistent and asymptotically normal distributed,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, F(\boldsymbol{\beta})^{-1}),$$

where $F(\boldsymbol{\beta})$ may be replaced by $F(\hat{\boldsymbol{\beta}})$ to obtain standard errors.

Goodness-of fit and tests on the significance of parameters based on deviance are provided within the framework of GLMs.

Extensions

In many applications count data are overdispersed, with conditional variance exceeding conditional mean. Several extensions of the basic model that account for overdispersion are available, in particular *quasi-likelihood methods* and more general distribution models like the Gamma–Poisson or *negative binomial model*. Quasi-likelihood uses the same estimation equations as maximum likelihood estimates, which are computed by solving

$$\sum_{i=1}^{n} \boldsymbol{x}_i \frac{\partial \mu_i}{\partial \eta} \frac{y_i - \mu_i}{v(\mu_i)} = \mathbf{0},$$

where $\mu_i = h(\eta_i)$ and $v(\mu_i)$ is the variance function. But instead of assuming the variance function of the Poisson model $v(\mu_i) = \mu_i$ one uses a more general form. For example, Poisson with overdispersion uses $v(\mu_i) = \phi \mu_i$ for some unknown constant ϕ. The case $\phi > 1$ represents *overdispersion* of the Poisson model, the case $\phi < 1$, which is rarely found in applications, is called *underdispersion*. Alternative variance functions usually continue to model the variance as a function of the mean. The variance function $v(\mu_i) = \mu_i + \gamma \mu_i^2$ with additional parameter γ corresponds to the variance of the negative binomial distribution.

It may be shown that the asymptotic properties of quasi-likelihood estimates are similar to that for GLMs. In particular one obtains asymptotically a normal distribution with the covariance given in the form of a pseudo-Fisher matrix, see McCullagh (1983), and McCullagh and Nelder (1989).

A source book for the modeling of count data which includes many applications is Cameron and Trivedi (1998). An econometric view on count data is outlined in Winkelmann (1997) and Kleiber and Zeileis (2008).

About the Author

Prof. Dr. Gerhard Tutz works at the Department of Statistics, Ludwig-Maximilians University Munich. He served as Head of the department for several years. He coauthored

the book *Multivariate Statistical Modeling Based on Generalized Linear Models* (with Ludwig Fahrmeir, Springer, 2001).

Cross References

►Dispersion Models
►Generalized Linear Models
►Geometric and Negative Binomial Distributions
►Modeling Count Data
►Poisson Distribution and Its Application in Statistics
►Robust Regression Estimation in Generalized Linear Models
►Statistical Methods in Epidemiology

References and Further Reading

Cameron AC, Trivedi PK (1998) Regression analysis of count data. econometric society monographs no. 30. Cambridge University Press, Cambridge
Kleiber C, Zeileis A (2008) Applied Econometrics with R. Springer, New York
McCullagh P (1983) Quasi-likelihood functions. Ann Stat 11:59–67
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, New York
Winkelmann R (1997) Count data models: econometric theory and application to labor mobility, 2nd edn. Springer, Berlin

Population Projections

Janez Malačič
Professor, Faculty of Economics
University of Ljubljana, Ljubljana, Slovenia

Population projections are a basic tool that demographers use to forecast a future population. They can be produced in the form of a *prognosis* or as *prospects*. The first is the most likely future development according to the expectations of the projections' author(s) and is produced in a single variant. The second type is based more on an "if-then" approach and is calculated in more variants. Usually, there are three or four variants, namely, low, medium, high, and constant variants. In practice, the medium variant is the most widely used and is taken as the most likely or accurate variant, that is, as a prognosis.

Population projections can be produced by *mathematical* or *analytical methods*. The *mathematical methods* use extrapolation(s) of various mathematical functions. For example, a census population can be extrapolated for a certain period into the future based on a linear, geometric, exponential, or some other functional form. The functional form is chosen on the basis of (1) past population development(s), (2) the developments of a neighboring and other similar populations, as well as (3) on the basis of general and particular demographic knowledge. In the great majority of cases, the mathematical methods are used for short- and midterm periods in the future. For population projections of small settlements and regions, only mathematical methods can be used in any reasonable way. Exceptionally, for very long-term periods (several centuries) a logistic curve can be used for a projection of the total population.

Analytical methods for population projections are much more complex. The population development in the projection period is decomposed at the level of basic components. These components are mortality, fertility, and migration. For each component, a special hypothesis of future development is produced. Very rich and complex statistical data are needed for analytical population projections. They are considered suitable for a period of 10–25 years into the future. They cannot be used to project populations in small areas because such life tables (see ►Life Table) and some other data will not be available. Analytical population projections offer very detailed information on particular population structures at present and in the future as well as data on the development of basic population components (e.g., mortality, fertility, and migration). They are also the basis for several other derived projections like those of households, of the active, rural, urban, and pensioned population.

Suppose that we have the census population divided by gender and age (in five-year age groups, say) as our starting point: ${}_{x+5}V^{t}_{m,x}$ and ${}_{x+5}V^{t}_{f,x}$, where V stands for population size, x for age, t for time, and m and f for male and female. To make a projection, we need three hypotheses. *The mortality hypothesis* is constructed on the basis of life table indicators. We take survival ratios ${}_{x+5}P_{m,x}$ and ${}_{x+5}P_{f,x}$ for all five-year age groups $(0-4, 5-9, \ldots, 80-84, 85+)$. Our hypothesis can be that mortality is constant, declining, or increasing, or we can have a combination of all three for each gender and each age group. Then we apply a population projection model aging procedure in the following form (spelled out for males):

$$ {}_{5}V^{t}_{m,0} * {}_{5}P_{m,0} = {}_{5}V^{t+5}_{m,5} \rightarrow {}_{5}V^{t+5}_{m,5} * {}_{5}P_{m,0} = {}_{5}V^{t+10}_{m,10}, \text{ etc.} $$

Evidently, in this case constant mortality hypothesis was used. The aging procedure would be applied for both genders and for all five-year age groups.

In the next step, we would construct a *fertility hypothesis*. This one can also be that fertility is constant, declining, increasing, or a combination of all three. It provides the newborn population for each year in the projection period. A set of different fertility indicators are available. The most convenient indicators are age-specific fertility rates, ${}_{x+5}f_{x}$,

where x is equal to 15, 20, 25, 30, 35, 40, and 45. In principle, we calculate the future number of births (N) for seven five-year age-groups with the formula:

$$ {}_5N_x^{t-(t+5)} = 5\left(\left({}_5V_{f,x}^{t} + {}_5V_{f,x}^{t+5}\right)/2 * {}_5f_x^{t}\right). $$

The number of births, $N^{t-(t+5)}$, should be subgrouped by gender. We can apply the demographic "constant" alternative and suppose that 485 girls are born per 1000 births. Then we calculate

$$ {}_5V_{m,0}^{t+5} = {}_5P_{m,r} * N_m^{t-(t+5)} \text{ and } {}_5V_{f,0}^{t+5} = {}_5P_{f,r} * N_f^{t-(t+5)} $$

${}_5P_m$, and ${}_5P_{f,r}$ are survival ratios for newborn boys and girls. In the case of a closed population or a population with zero migration, our projection is finished.

However, real populations have in- and out-migration. To cover this case, we should construct a *migration hypothesis*. The procedure is similar to the mortality and fertility hypotheses. Suppose we use net migration rates for five-year age groups, separately by gender. In principle, we calculate age-specific net migration factors (for males), nm is net migration:

$$ {}_5F_{m,x} = 1 + ({}_5nm_{m,x}/1,000). $$

The population aging procedure changes slightly:

$$ {}_5V_{m,x+5}^{t+5} = {}_5V_{m,x}^{t} * {}_5P_{m,x} * {}_5F_{m,x}. $$

The procedure for the female population is parallel to the procedure for the male population. The most serious problem in practice is that age-specific migration data may be unavailable or of poor quality.

Such simple analytical population projection procedures have been improved considerably in the literature during recent decades. Probably the most important improvement is the construction of *probabilistic population projections*, for which considerable progress has been made in recent decades (Lutz et al. 1998). Many analytical population projections for countries and regions are now supplemented by several national probabilistic population forecasts.

About the Author

Dr. Janez Malačić is a Professor of demography and labor economics, Faculty of Economics, Ljubljana University, Slovenia. He is a Former President of the Slovenian Statistical Society (1985–1987), a Former President of the Society of Yugoslav Statistical Societies (1986–1988), and a member of the IUSSP (from 1986). He has authored two books and more than 150 papers. His papers have been published in eight languages.

Cross References

►Actuarial Methods
►African Population Censuses
►Census
►Demographic Analysis: A Stochastic Approach
►Demography
►Life Table
►Survival Data

References and Further Reading

Eurostat, European Commission (2007) Work session on demographic projections. Bucharest, 10–12 October 2007. Methodologies and working papers. 370 p

Lutz W, Goldstein JR (guest eds) (2004) How to deal with uncertainty in population forecasting? Int Stat Rev 72(1–2):1–106, 157–208

Lutz W, Vaupel JW, Ahlburg DA (eds) (1999) Frontiers of population Forecasting. A Supplement to vol 24, 1998, population and Development Review. The Population Council, New York

Portfolio Theory

Harry M. Markowitz
Professor, Winner of the Nobel Memorial Prize in Economic Sciences in 1990
University of California, San Diego, CA, USA

Portfolio Theory considers the trade-off between some measure of risk and some measure of return on the portfolio-as-a-whole. The measures used most frequently in practice are expected (or mean) return and variance or, equivalently, standard deviation. This article discusses the justification for the use of mean and variance, sources of data needed in a mean-variance analysis, how mean-variance tradeoff curves are computed, and semi-variance as an alternative to variance.

Mean-Variance Analysis and its Justification

While the idea of trade-off curves goes back at least to Pareto, the notion of a trade-off curve between risk and return (later dubbed the efficient frontier) was introduced in Markowitz (1952). Markowitz proposed expected return and variance as both a hypothesis about how investors act and as a rule for guiding action in fact. By Markowitz (1959) he had given up the notion of mean and variance as a hypothesis but continued to propose them as criteria for action.

Tobin (1958) said that the use of mean and variance as criteria assumed either a quadratic utility function or a

Gaussian probability distribution. This view is sometimes ascribed to Markowitz, but he never justified the use of mean and variance in this way. His views evolved considerably from Markowitz (1952) to Markowitz (1959). Concerning these matters Markowitz (1952) should be ignored. Markowitz (1959) accepts the views of Von Neumann and Morgenstern (1944) when probability distributions are known, and L.J. Savage (1954) when probabilities are not known. The former asserts that one should maximize expected utility; the latter asserts that when probabilities are not known one should maximize expected utility using probability beliefs when objective probabilities are not known.

Markowitz (1959) conjectures that a suitably chosen point from the efficient frontier will approximately maximize expected utility for the kinds of utility functions that are commonly proposed for cautious investors, and for the kinds of probability distributions that are found in practice. Levy and Markowitz (1979) expand on this notion considerably. Specifically, Levy and Markowitz show that for such probability distributions and utility functions there is typically a correlation between the actual expected utility and the mean-variance approximation between 0.95 and of 0.99. They also show that the Pratt (1964) and Arrow (1971) objection to quadratic utility does not apply to the kind of approximations used by Levy and Markowitz, or in Markowitz (1959).

Models of Covariance

If covariances are computed from historical returns with more securities than there are observations, e.g., 5,000 securities and 60 months of observations, then the covariance matrix will be singular. A preferable alternative is to use a model of covariance where the return on the *i*th security is assumed to obey the following relationship

$$r_i = \alpha_i + \sum \beta_{ik} f_k + u_i$$

where the u_i are independent of each other and the f_k. The f_k may be either factors or scenarios or some of each. These ideas are carried out in, for example, Sharpe (1963), Rosenberg (1974) and Markowitz and Perold (1981a, 1981b).

Estimation of Parameters

Covariance matrices are sometimes estimated from historical returns and sometimes from factor or scenario models such as the one-factor model of Sharpe, the many-factor model of Rosenberg, or the scenario models of Markowitz and Perold cited above.

Expected returns are estimated in a great variety of ways. I do not believe that anyone suggests that, in practice, historical average returns should be used as the expected returns of individual stocks. The Ibbotson and Sinquefield (2007) series are frequently used to estimate expected returns for asset classes. Black and Litterman (1991, 1992) propose a very interesting Bayesian approach to the estimation of expected returns. Richard Michaud (1989) proposes to use estimates for asset classes based on what he refers to as a "resampled frontier". Additional methods for estimating expected return are based on statistical methods for "disentangling" various anomalies, see Jacobs and Levy (1988), or estimates based on factors that Graham and Dodd (1940) might use: see Lakonishok et al. (1994), Ohlson (1979), and Bloch et al. (1993). The last mentioned paper is based on results obtained by back-testing many alternate hypotheses concerning how to achieve excess returns. When many estimation methods are tested, the expected future return for the best of the lot should not be estimated as if this were the only procedure tested. Estimates should be corrected for "data mining." See Markowitz and Xu (1994).

Computation of M-V Efficient Sets

The set of mean-variance efficient portfolios is piecewise linear. The critical line algorithm (CLA) traces out this set, one linear piece at a time, without having to search for optima. CLA is described in Appendix A of Markowitz (1959) and, less compactly, in Markowitz and Todd (2000).

Downside Risk

"Semi-variance" or downside risk is like variance, but only considers deviations below the mean or below some target return. It is proposed by Markowitz (1959) Chap. 9 and championed by Sortino and Satchell (2001). It is used less frequently in practice than variance.

About the Author

Professor Markowitz has applied computer and mathematical techniques to various practical decision making areas. In finance: in an article in 1952 and a book in 1959 he presented what is now referred to as MPT, "modern portfolio theory." This has become a standard topic in college courses and texts on investments, and is widely used by institutional investors for asset allocation, risk control and attribution analysis. In other areas: Dr. Markowitz developed "sparse matrix" techniques for solving very large mathematical optimization problems. These techniques are now standard in production software for optimization programs. Dr. Markowitz also designed and supervised the development of the SIMSCRIPT programming language. SIMSCRIPT has been widely used for programming computer simulations of systems like factories, transportation systems and communication networks.

In 1989 Dr. Markowitz received The John von Neumann Award from the Operations Research Society of America for his work in portfolio theory, sparse matrix techniques and SIMSCRIPT. In 1990 he shared The Nobel Prize in Economics for his work on portfolio theory.

Cross References

►Actuarial Methods
►Business Statistics
►Copulas in Finance
►Heteroscedastic Time Series
►Optimal Statistical Inference in Financial Engineering
►Semi-Variance in Finance
►Standard Deviation
►Statistical Modeling of Financial Markets
►Variance

References and Further Reading

Arrow K (1971) Aspects of the theory of risk bearing. Markham Publishing Company, Chicago
Black F, Litterman R (1991) Asset allocation: combining investor views with market equilibrium. J Fixed Income 1(2):7–18
Black F, Litterman R (1992) Global portfolio optimization. Financ Anal J 48(5):28–43
Bloch M, Guerard J, Markowitz H, Todd P, Xu G (1993) A comparison of some aspects of the US and Japanese equity markets. Jpn World Econ 5:3–26
Graham B, Dodd DL (1940) Security analysis, 2nd edn. McGraw-Hill, New York
Ibbotson RG, Sinquefield RA (2007) Stocks, bonds, bills and inflation yearbook. Morningstar, New York
Jacobs BI, Levy KN (1988) Disentangling equity return regularities: new insights and investment opportunities. Financ Anal J 44(3):18–44
Lakonishok J, Shleifer A, Vishny RW (1994) Contrarian investment, extrapolation and risk. J Financ 49(5):1541–1578
Levy H, Markowitz HM (1979) Approximating expected utility by a function of mean and variance. Am Econ Rev 69(3):308–317
Markowitz HM (1952) Portfolio selection. J Financ 7(1):77–91
Markowitz HM (1959) Portfolio selection: efficient diversification of investments. Wiley, New York (Yale University Press, 1970, 2nd edn, Basil Blackwell, 1991)
Markowitz HM, Perold AF (1981a) Portfolio analysis with factors and scenarios. J Financ 36(14):871–877
Markowitz HM, Perold AF (1981b) Sparsity and piecewise linearity in large portfolio optimization problems. In: Duff IS (ed) Sparse matrices and their uses. Academic Press, New York, pp 89–108
Markowitz HM, Todd P (2000) Mean-variance analysis in portfolio choice and capital markets. Frank J. Fabozzi Associates, New Hope [revised reissue of Markowitz (1987) (with chapter by Peter Todd)]
Markowitz HM, Xu GL (1994) Data mining corrections. J Portfolio Manage 21:60–69
Michaud RO (1989) The Markowitz optimization enigma: is optimized optimal? Financ Anal J 45(1):31–42
Ohlson JA (1979) Risk return, security-valuation and the stochastic behavior of accounting numbers. J Financ Quant Anal 14(2):317–336
Pratt JW (1964) Risk aversion in the small and in the large. Econometrica 32:122–136
Rosenberg B (1974) Extra-market components of covariance in security returns. J Financ Quant Anal 9(2):263–273
Savage LJ (1954) The foundations of statistics. Wiley, New York (Second revised edn. Dover, New York)
Sharpe WF (1963) A simplified model for portfolio analysis. Manage Sci 9(2):277–293
Sortino F, Satchell S (2001) Managing downside risk in financial markets: theory, practice and implementation. Butterworth-Heinemann, Burlington
Tobin J (1958) Liquidity preference as behavior towards risk. Rev Econ Stud 25(1):65–86
Von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. 3rd edn. (1953), Princeton University Press, Princeton

Posterior Consistency in Bayesian Nonparametrics

Jayanta K. Ghosh[1], R. V. Ramamoorthi[2]
[1]Professor of Statistics
Purdue University, West Lafayette, IN, USA
[2]Professor
Michigan State University, East Lansing, MI, USA

Bayesian Nonparametrics (see ►Bayesian Nonparametric Statistics) took off with two papers of Ferguson (Ferguson 1974, 1983) and followed by Antoniak (Antoniak 1974). However consistency or asymptotics were not major issue in those papers, which were more concerned with taking the first steps towards a usable, easy to interpret prior with easy to choose hyperparameters and a rich support. Unfortunately, the fact that the Dirichlet sits on discrete distributions diminished the early enthusiasm.

The idea of consistency came from Laplace and informally may be defined as : Let $\mathcal{P}$ be a set of probability measures on a sample space $\mathcal{X}$, Π be a prior on $\mathcal{P}$. The posterior is said to be consistent at a true value P_0 if the following holds: For sample sequences with P_0 probability 1, the posterior probability of any neighborhood U of P_0 converges to 1.

The choice of neighborhoods U determines the strength of consistency. One choice, when the sample space is separable metric, is weak neighborhoods U of P_0. When elements of $\mathcal{P}$ have densities, L_1 neighborhoods of P_0 is often the relevant choice. If the family $\mathcal{P}$ is parametrized by θ, then these notions easily translate to θ, via continuity requirements of the map $\theta \mapsto P_\theta$.

The early papers on consistency were by Freedman (Freedman 1963, 1965) on multinomials with (countably) infinite classes. They were very interesting but provided a

somewhat negative picture, namely, that in a topological sense, for most priors (i.e., outside a class of first category) consistency fails to hold. This lead Freedman to consider tail-free priors including the Dirichlet prior for the countably many probabilities of the countably infinite multinomial. Around the same time, in her posthumous paper (Schwartz 1965) of 1965, arising from her thesis at Berkeley, written under Le Cam, Schwartz showed among other things that, if the prior assigned positive probability to all Kullback-Leibler neighborhoods of the true density, then consistency holds in the sense of weak convergence. This is an important result which showed that this is the right notion of support in these problems, not the more usual one adopted by Freedman.

These early papers were followed by Ferguson's Dirichlet process on the set of all probability measures along with Antoniak's study of mixtures. Even though the set of discrete measures had full measure under the Dirichlet process, it still enjoyed the property of consistency at all distributions. The paper by Diaconis and Freedman (Diaconis and Freedman 1986), along with the discussions revived interest in consistency issues. Diaconis and Freedman showed that with Dirichlet process prior consistency can go awry in the presence of a location parameter. Barron, in his discussion of the paper provided insight as to why it would be unreasonable to expect consistency in this example. Ghosh and Ramamoorthi (Ghosh and Ramamoorthi 2003) discuss several different explanation for lack of consistency if one uses the Dirichlet Process in a semiparametric problem with a location parameter.

Other major contributions have been made by Barron (Barron 1988), Barron, Schervish and Wasserman (Barron et al. 1999), Walker (Walker 2004) and Coram and Lalley (Coram and Lalley 2006). A thorough review up to 2008 is available in Choi and Ramamoorthi (Choi and Ramamoorthi 2008). Contributions to rates of convergence have been made by Ghosal, Ghosh and van der Vaart, (Ghosal et al. 2000), Ghosal and van der Vaart (Ghosal and van der Vaart 2001), Shen and Waseerman (2001), Kleijn and van der Vaart (Kleijn and van der Vaart 2006) and (van der Vaart and van Zanten 2009), (van der Vaart and van Zanten 2008). see also the book by Ghosh and Ramamoorthi (Ghosh and Ramamoorthi 2003) for many basic early results on consistency and other aspects of, like choice of priors and consistency for density estimation, semiparametric problems and survival analysis.

Ghosh and Ramamoorthi (Ghosh and Ramamoorthi 2003) deal only with nonparametric estimation problems. Work on nonparametric testing and consistency problems there have begun only recently. A survey is available in Tokdar, Chakravarti and Ghosh (Tokdar et al. 2010).

About the Authors

For biography of Professor Ghosh *see* the entry ►Bayesian P-Values.

Professor R.V. Ramamoorthi obtained his doctoral degree from the Indian Statistical Institute. His early work was on sufficiency and decision theory, a notable contribution there being the joint result with Blackwell showing that Bayes sufficiency is not equivalent to Fisherian sufficiency. Over the last few years his research has been on Bayesian nonparamterics where J.K. Ghosh and he authored one of the first books on the topic, *Bayesian Nonparametrics* (Springer 2003).

"This is the first book to present an exhaustive and comprehensive treatment of Bayesian nonparametrics. Ghosh and Ramamoorthi present the theoretical underpinnings of nonparametric priors in a rigorous yet extremely lucid style…It is indispensable to any serious Bayesian. It is bound to become a classic in Bayesian nonparametrics." (Jayaram Sethuraman, Review Of Bayesian Nonparametrics, *Sankhya*, 2004, 66, 208–209).

Cross References

►Bayesian Nonparametric Statistics
►Bayesian Statistics
►Nonparametric Statistical Inference

References and Further Reading

Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann Stat 2:1152–1174
Barron A (1988) The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, Department Statistics, University of Illinois, Champaign
Barron A, Schervish MJ, Wasserman L (1999) The consistency of posterior distributions in nonparametric problems. Ann Stat 27:536–561
Choi T, Ramamoorthi RV (2008) Remarks on consistency of posterior distributions. In: Pushing the limits of contemporary statistics: contributions in honor of Ghosh JK, vol 3 of Inst Math Stat Collect. Institute of Mathematical Statistics, Beachwood, pp 170–186
Coram M, Lalley SP (2006) Consistency of Bayes estimators of a binary regression function. Ann Stat 34:1233–1269
Diaconis P, Freedman D (1986) On the consistency of Bayes estimates. Ann Stat 14:1–67. With a discussion and a rejoinder by the authors
Ferguson TS (1974) Prior distributions on spaces of probability measures. Ann Stat 2:615–629
Ferguson TS (1983) Bayesian density estimation by mixtures of normal distributions. In: Recent advances in statistics. Academic, New York, pp 287–302
Freedman DA (1963) On the asymptotic behavior of Bayes' estimates in the discrete case. Ann Math Stat 34:1386–1403
Freedman DA (1965) On the asymptotic behavior of Bayes estimates in the discrete case II. Ann Math Stat 36:454–456
Ghosal S, Ghosh JK, van der Vaart AW (2000) Convergence rates of posterior distributions. Ann Stat 28:500–531

Ghosal S, van der Vaart AW (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann Stat 29:1233–1263
Ghosh JK, Ramamoorthi RV (2003) Bayesian nonparametrics. Springer Series in Statistics. Springer, New York
Kleijn BJK, van der Vaart AW (2006) Misspecification in infinite-dimensional Bayesian statistics. Ann Stat 34:837–877
Schwartz L (1965) On Bayes procedures. Z Wahrscheinlichkeitstheorie und Verw Gebiete 4:10–26
Shen X, Wasserman L (2001) Rates of convergence of posterior distributions. Ann Stat 29(3):687–714
Tokdar S, Chakrabarti A, Ghosh J (2010) Bayesian nonparametric goodness of fit tests. In: M-H Chen, DK Dey, P Mueller, D Sun, K Ye (eds) Frontiers of statistical decision making and Bayesian analysis. Inst Math Stat Collect. Institute of Mathematical Statistics, Beachwood
van der Vaart AW, van Zanten JH (2008) Reproducing kernel Hilbert spaces of Gaussian priors. In: Pushing the limits of contemporary statistics: contributions in honor of Ghosh JK, vol 3 of Inst Math Stat Collect. Institute of Mathematical Statistics, Beachwood, pp 200–222
van der Vaart AW, van Zanten JH (2009) Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. Ann Stat 37:2655–2675
Walker S (2004) New approaches to Bayesian consistency. Ann Stat 32:2028–2043

Power Analysis

Kevin R. Murphy
Professor
Pennsylvania State University, University Park, PA, USA

One of the most common applications of statistics in the social and behavioral science is in testing null hypotheses. For example, a researcher wanting to compare the outcomes of two treatments will usually do so by testing the hypothesis that in the population there is no difference in the outcomes of the two treatments. The power of a statistical test is defined as the likelihood that a researcher will be able to reject a specific null hypothesis when it is in fact false.

Cohen (1988), Lipsey (1990), and Kraemer and Thiemann (1987) provided excellent overviews of the methods, assumptions, and applications of power analysis. Murphy and Myors (2003) extended traditional methods of power analysis to tests of hypotheses about the size of treatment effects, not merely tests of whether or not such treatment effects exist.

The power of a null hypothesis test is a function of sample size (n), effect size (ES), and the standard used to define statistical significance (α), and the equations that define this relation can be easily rearranged to solve for any of four quantities (i.e., power, n, ES, and α), given the other three. The two most common applications of statistical power analysis are in: (1) determining the power of a study, given n, ES, and α, and (2) determining how many observations will be needed (i.e., n required), given a desired level of power, an ES estimate, and the α value. Both of these methods are widely used in designing studies; one widely-accepted convention is that studies should be designed so that they achieve power levels of 0.80 or greater (i.e., so that they have at least an 80% chance of rejecting a false null hypothesis; Cohen 1988; Murphy and Myors 2003).

There are two other applications of power analysis that are less common, but no less informative. First, power analysis may be used to evaluate the sensitivity of studies. That is, power analysis can indicate what sorts of effect sizes might be reliably detected in a study. If one expects the effect of a treatment to be small, it is important to know whether the study will detect that effect, or whether the study as planned only has sufficient sensitivity to detect larger effects. Second, one may use power analysis to make rational decisions about the criteria used to define "statistical significance."

Power analyses are included as part of several statistical analysis packages (e.g., SPSS provides Sample Power, a flexible and powerful program) and it is possible to use numerous websites to perform simple power analyses. Two notable software packages designed for power analysis are:

- *G*Power* (Faul et al. 2007; http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/ is distributed as a freeware program that is available for both Macintosh and Windows environments. It is simple, fast, and flexible.
- *Power and Precision*, distributed by Biostat, was developed by leading researchers in the field (e.g., J. Cohen). This program is very flexible, covers a large array of statistical tests, and provides power analyses and confidence intervals for most tests.

About the Author

Kevin Murphy is a Professor of Psychology and Information Sciences and Technology at the Pennsylvania State University. He has served as President of the Society for Industrial and Organizational Psychology and Editor of *Journal of Applied Psychology*. He is the author if eleven books, including *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests* (with Brett Myors, Erlbaum 2009), and over 150 articles and chapters.

Cross References

►Effect Size
►Presentation of Statistical Testimony
►Psychology, Statistics in
►Sample Size Determination
►Significance Testing: An Overview
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Significance

References and Further Reading

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Erlbaum, Hillsdale

Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Meth 39:175–191

Kraemer HC, Thiemann S (1987) How many subjects? Sage, Newbury Park

Lipsey MW (1990) Design sensitivity. Sage, Newbury Park

Murphy K, Myors B (2009) Statistical power analysis: a simple and general model for traditional and modern hypothesis tests, 3rd edn. Erlbaum, Mahwah

Preprocessing in Data Mining

Edgar Acuña
Professor
University of Puerto Rico at Mayaguez, Mayaguez, Puerto Rico

Introduction

►Data mining is the process of extracting hidden patterns in a large dataset. Azzopardi (2002) breaks the data mining process into five stages:

(a) *Selecting the domain* – data mining should be assessed to determine whether there is a viable solution to the problem at hand and a set of objectives should be defined to characterize these problems.
(b) *Selecting the target data* – this entails the selection of data that is to be used in the specified domain; for example, selection of subsets of features or data samples from larger databases.
(c) *Preprocessing the data* – this phase is primarily aimed at preparing the data in a suitable and useable format, so that a knowledge extraction process can be applied.
(d) *Extracting the knowledge/information* – during this stage, the types of data mining operations (association rules, regression, supervised classification, clustering, etc.), the data mining techniques, and data mining algorithms are chosen and the data is then mined.
(e) *Interpretation and evaluation* – this stage of the data mining process is the interpretation and evaluation of the discoveries made. It includes filtering information that is to be presented, visualizing graphically, or locating the useful patterns and translating the patterns discovered into an understandable form.

In the process of data mining, many patterns are found in the data. Patterns that are interesting for the miner are those that are easily understood, valid, potentially useful, and novel (Fayyad et al. 1996). These patterns should validate the hypothesis that the user seeks to confirm. The quality of patterns obtained depends on the quality of the data analyzed. It is common practice to prepare data before applying traditional data mining techniques such as regression, association rules, clustering, and supervised classification.

Section "►Reasons for Applying Data Preprocessing" of this article provides a more precise justification for the use of data preprocessing techniques. This is followed by a description in section "►Techniques for Data Preprocessing" of some of the data preprocessing techniques currently in use.

Reasons for Applying Data Preprocessing

Pyle (1999) suggests that about 60% of the total time required to complete a data mining project should be spent on data preparation since it is one of the most important contributors to the success of the project. Transforming the data at hand into a format appropriate for knowledge extraction has a significant influence on the final models generated, as well as on the amount and quality of the knowledge discovered during the process. At the same time, the effect caused by changes made to a dataset during data preprocessing can either facilitate or complicate even further the knowledge discovery process; thus changes made must be selected with care.

Today's real-world datasets are highly susceptible to noise, missing and inconsistent data due to human errors, mechanical failures, and to their typically large size. Data affected in this manner is known as "dirty." During the past decades, a number of techniques have been developed to preprocess data gathered from real-world applications before the data is further processed for other purposes.

Cases where data mining techniques are applied directly to raw data without any kind of data preprocessing are still frequent; yet, data preprocessing has been recommended as an obligatory step. Data preprocessing techniques should never be applied blindly to a dataset, however. Prior to any data preprocessing effort, the dataset should be explored and characterized. Two methods for

exploring the data prior to preprocessing are *data characterization* and *data visualization*.

Data Characterization

Data characterization describes data in ways that are useful to the miner and begins the process of understanding what is in the data. Engels and Theusinger (1998) describe the following characteristics as standard for a given dataset: the number of classes, the number of observations, the percentage of missing values in each attribute, the number of attributes, the number of features with numeric data type, and the number of features with symbolic data type. These characteristics can provide a first indication of the complexity of the problem being studied.

In addition to the above-mentioned characteristics, parameters of location and dispersion can be calculated as single-dimensional measurements that describe the dataset. Location parameters are measurements such as minimum, maximum, arithmetic mean, median, and empirical quartiles. On the other hand, dispersion parameters such as range, standard deviation, and quartile deviation provide measurements that indicate the dispersion of values of the feature.

Location and dispersion parameters can be divided in two classes: those that can deal with extreme values and those that are sensitive to them. A parameter that can deal well with extreme values is called robust. Some statistical software packages provide the computation of robust parameters in addition to the traditional non-robust parameters. Comparing robust and non-robust parameter values can provide insight to the existence of ►outliers during the data characterization phase.

Data Visualization

Visualization techniques can also be of assistance during this exploration and characterization phase. Visualizing the data before preprocessing it can improve the understanding of the data, thereby increasing the likelihood that new and useful information will be gained from the data. Visualization techniques can be used to identify the existence of missing values, and outliers, as well as to identify relationships among attributes. These techniques can, in effect, assist in ranking the "impurity" of the data and in selecting the most appropriate data preprocessing technique to apply.

Techniques for Data Preprocessing

Applying the correct data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Lu et al. (1996), Pyle (1999), and Azzopardi (2002) present descriptions of common techniques for preparing data for analysis. The techniques described by both authors can be summarized as follows:

(a) *Data cleaning* – filling in missing values, smoothing noisy data, removing outliers, and resolving inconsistencies.
(b) *Data reduction* – reducing the volume of data (but preserving the patterns) by removing repeated observations and applying *instance selection* as well as *feature selection* techniques. Discretization of continuous attributes is also a way of data reduction.
(c) *Data transformation* – converting text and graphical data to a format that can be processed, normalizing or scaling the data, aggregation, and generalization.
(d) *Data integration* – correcting differences in coding schemes due to the combining of several sources of data.

Data Cleaning

Data cleaning provides methods to deal with dirty data. Since dirty datasets can cause problems for data exploration and analysis, data cleaning techniques have been developed to clean data by filling in missing values (value imputation), smoothing noisy data, identifying and/or removing outliers, and resolving inconsistencies. Noise is a random error or variability in a measured feature, and several methods can be applied to remove it. Data can also be smoothed by using regression to find a mathematical equation to fit the data. Smoothing methods that involve *discretization* are also methods of data reduction since they reduce the number of distinct values per attribute. Clustering methods can also be used to remove noise by detecting outliers.

Data Integration

Some studies require the integration of multiple databases, or files. This process is known as *data integration*. Since attributes representing a given concept may have different names in different databases, care must be taken to avoid causing inconsistencies and redundancies in the data. Inconsistencies are observations that have the same values for each of the attributes but that are assigned to different classes. Redundant observations are observations that contain the same information.

Attributes that have been derived or inferred from others may create redundancy problems. Again, having a large amount of redundant and inconsistent data may slow down the knowledge discovery process for a given dataset.

Data Transformation

Many data mining algorithms provide better results if the data has been normalized or scaled to a specific range before these algorithms are applied. The use of normalization techniques is crucial when distance-based algorithms are applied, because the distance measurements taken on by attributes that assume many values will generally outweigh distance measurements taken by attributes that assume fewer values. Other methods of *data transformation* include data aggregation and generalization techniques. These methods create new attributes from existing information by applying summary operations to data or by replacing raw data by higher-level concepts. For example, monthly sales data may be aggregated to compute annual sales.

Data Reduction

The increased size of current real-world datasets has led to the development of techniques that can reduce the size of the dataset without jeopardizing the data mining results. The process known as *data reduction* obtains a reduced representation of the dataset that is much smaller in volume, yet maintains the integrity of the original data. This means that data mining on the reduced dataset should be more efficient yet produce similar analytical results. Han and Kamber (2006) mention the following strategies for data reduction:

(a) *Dimension reduction*, where algorithms are applied to remove irrelevant, weakly relevant, or redundant attributes.
(b) *Data compression*, where encoding mechanisms are used to obtain a reduced or compressed representation of the original data. Two common types of data compression are wavelet transforms and ►principal component analysis.
(c) *Numerosity reduction*, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which store only the model parameters instead of the actual data), or nonparametric methods such as clustering and the use of histograms.
(d) *Discretization and concept hierarchy generation*, where raw data values for attributes are replaced by ranges or higher conceptual levels. For example, concept hierarchies can be used to replace a low-level concept such as age, with a higher-level concept such as young, middle-aged, or senior. Some detail may be lost by such data generalizations.
(e) *Instance selection*, where a subset of best instances of the whole dataset is selected. Some of the instances are more relevant than others to perform a data mining, and working only with an optimal subset of instances, it will be more cost-and time-efficient. Variants of the classical sampling techniques can be used.

Final Remarks

Acuna (2009) has developed Drep, an R package for data preprocessing and visualization. Drep performs most of the data preprocessing techniques mentioned in this article. Currently, research is being done in order to apply preprocessing methods to data streams, see Aggarwal (2007) for more details.

About the Author

Dr. Edgar Acuña, is a Professor, Department of Mathematical Sciences, University of Puerto Rico at Mayaguez. He is also the leader of the group in Computational and statistical Learning from databases at the University of Puerto Rico. He has authored and co-authored more than 20 papers mainly on data preprocessing. He is the author of book (in Spanish) *Analisis Estadistico De Datos usando* Minitab (John Wiley & Sons, 2002). In 2003, he was honored with the Power Hitter Award in Business and Technology. In 2008, he was selected as a Fullbright visiting Scholar. Currently, he is an Associate editor for the *Revista Colombiana de Estadistica*.

Cross References

►Box–Cox Transformation
►Data Mining
►Multi-Party Inference and Uncongeniality
►Outliers

References and Further Reading

Acuna E (2009) Dprep: data preprocessing and visualization functions for classification. URL http://cran.r-hproject.org/package=dprep. R package version 2.1
Aggarwal CC (ed) (2007) Data streams: models and algorithms. Springer, New York
Azzopardi L (2002) "Am I Right?" asked the classifier: preprocessing data in the classification process. Comput Inform Syst 9:37–44
Engels R, Theusinger C (1998) Using a data metric for preprocessing advice for data mining applications. In: Proceedings of 13th European conference on artificial intelligence, pp 430–434
Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery: an overview. In: Advances in knowledge discovery and data mining, Chapter 1, AAAI Press/MIT Press, pp 1–34
Han J, Kamber M (2006) Data mining: concepts and techniques, 2nd edn. Morgan Kaufman Publishers
Lu H, Sun S, Lu Y (1996) On preprocessing data for effective classification. ACM SIGMOD'96 workshop on research issues on data mining and knowledge discovery, Montreal, QC
Pyle D (1999) Data preparation for data mining. Morgan Kaufmann, San Francisco

Presentation of Statistical Testimony

Joseph L. Gastwirth
Professor of Statistics and Economics
George Washington University, Washington, DC, USA

Introduction

Unlike scientific research, where we have the luxury of carrying out new studies to replicate and determine the domain of validity of prior investigations, the law also considers other social goals. For example, in many nations one spouse cannot be forced to testify against the other. A main purpose of the law is to resolve a dispute, so a decision needs to be made within a reasonable amount of time after the charge is filed. Science is primarily concerned with determining the true mechanism underlying a phenomenon. Typically no limits are placed on the nature of the experiment or approach an investigator may take to a problem. In most uses of statistics in legal proceedings, the relevant events happened several years ago; rarely will you be able to collect additional data. (One exception occurs in cases concerned with violations of laws protecting intellectual property, such as the Lanham Act in the USA, which prohibits a firm from making a product that infringes on an established one. Surveys of potential consumers are conducted to estimate the percentage who might be confused as to the source of the product in question.) Often, the data base will be one developed for administrative purposes, e.g., payroll or attendance records, which you will need to rely on.

Civil cases differ from criminal cases in that the penalty for violating a civil statue is not time in prison but rather compensation for the harm done. Consequently, the burden of proof a plaintiff has in discrimination or tort case is to show that the defendant caused the alleged harm by "the preponderance of the evidence" rather than the stricter standard of "beyond a reasonable doubt" used in criminal cases. Thus, many statistical studies are more useful in civil cases.

A major difference between presenting testimony in court or a regulatory hearing and giving a talk at a major scientific conference is that expert witnesses, like all others, are only allowed to answer questions put to them by the lawyers. Thus, particular findings or analyses that you believe are very important may not be submitted as evidence if the lawyer who hired you does not ask you about them when you are testifying. Unless the judge, or in some jurisdictions a juror, asks you a question related to that topic you are not allowed to discuss it.

Courts have also adopted criteria to assess the reliability of scientific evidence as well as some traditional ways of presenting and analyzing some types of data. (The leading case is *Daubert v. Merrell-Dow Pharmaceuticals Inc.*, 509 U.S. 579 (1993). The impact of this case and two subsequent ones on scientific evidence is described by Berger (2000) and Rosenblum (2000). Some of the criteria courts consider are: can the methodology was subject to peer review, can it be replicated and tested and whether the potential error rates are known and considered in the expert's report and testimony.) This may limit the range of analyses you can use in the case at hand; although subsequently it can stimulate interesting statistical research. A potentially more serious threat to the credibility of your research and subsequent testimony case is due to the fact that the lawyers provide you with the data and background information. Thus, you may not even be informed that other information or data sets exist.

A related complication can arise when the lawyer hires both a consulting expert who has complete access to all the data as he or she is protected by the "work product" rule and then hires a "testifying expert". This second expert may only be asked to analyze the data favorable to the defendant and not told that any other data exists. Sometimes, the analytic approach, e.g., regression analysis, may be suggested to this expert because the lawyer already knows the outcome. If one believes an alternative statistical technique would be preferable or at least deserves exploration, the expert may be constrained as to the choice of methodology.

This entry describes examples of actual uses of statistical evidence, along with suggestions to aid courts in understanding the implications of the data. Section "►Presenting the Data or Summary Tables That will be Analyzed" discusses the presentation of data and the results of statistical tests. One dataset illustrates the difficulty of getting lawyers and judges to appreciate the statistical concept of "power", the ability of a test to detect a real or important difference. As a consequence an analysis that used a test with *no* power was accepted by a court. In section "►A More Informative Summary of Promotion Data: Hogan v. Pierce (31 F.E.P. 115 (D.D.C. 1983))" will show how in a subsequent case I was able to present more detailed data, which helped make the data clearer to the court. The last section offers some suggestions for improving the quality of statistical analyses and their presentation.

Presenting the Data or Summary Tables That will be Analyzed

In the classic *Castenada v. Partida* (430 U.S. 482, 97 S. Ct. 1272 (1997)) case concerning whether Mexican-Americans

were discriminated against in the jury selection process, the Court summarized the data by years, i.e., the data for *all* juries during the year were aggregated and the minority fraction compared to their fraction of the total population as well as the subgroup eligible for jury service. (The data is reported in footnote 7 of the opinion as well as in the texts: Finkelstein and Levin (2000) and Gastwirth (1988).) The data showed a highly significant difference between the Mexican-American fraction of jurors (39%) and both their fraction of the total population (79.1%) and of adults with some schooling (65%). From a statistical view the case is important as it established that formal statistical hypothesis testing would be used rather than intuitive judgments about whether the difference between the percentage of jurors who were from the minority group differed sufficiently from the percentage of minorities eligible for jury service. When the lower courts followed the methodology laid out in *Castenada*, they also adopted the tradition of presenting yearly summaries of the data in discrimination cases.

Unlike jury discrimination cases, which are typically brought by a defendant in a criminal case rather than the minority juror who was dismissed, in equal employment cases the plaintiff is the individual who suffered the alleged discriminatory act. In the United States the plaintiff has 180 days from the time of the alleged act, e.g., not being hired or promoted or of being laid off, to file a formal complaint with the Equal Employment Opportunity Commission (EEOC). Quite often after receiving notice of the complaint, the employer will modify their system to mitigate the effect of the employment practice under scrutiny on the minority group in question. The impact of this "change" in policy on statistical analysis has often been overlooked by courts. In particular, if a change occurs during the year the charge occurs and employer may change their policy and include the post-charge minority hires or promotions in their analysis.

Let me use data from a case, *Capaci v. Katz & Besthoff* (525 F. Supp. 317 (E.D. La. 1981), *aff'd in part, rev'd in part*, 711 F.2d 647 (5th Cir. 1983)), in which I was an expert for the plaintiffs to illustrate this. On January, 11, 1973 the plaintiff filed a charge of discrimination against women in promotions in the pharmacy department. One way such discriminatory practices may be carried out is to require female employees to work longer at the firm before promotion than males. Therefore, a study comparing the length of time males and female Pharmacists served before they were promoted to Chief Pharmacist was carried out. The time frame considered started in July 1, 1965 the effective date of the Civil Rights Act until the date of the charge. The times each Pharmacist who was promoted had served until they received their promotion are reported in Table 1. Only their initials of the employees are given.

Presentation of Statistical Testimony. Table 1 Months of service for male and female pharmacists employed at K&B during the period July 1, 1965 thru January 11, 1973 before receiving a promotion to chief pharmacist

Females: 229; 453.
Males: 5; 7; 12; 14; 14; 14; 18; 21; 22, 23; 24; 25; 25; 34; 34; 37; 47; 49; 64; 67; 69; 125; 192; 483.

Applying the Wilcoxon test (see ▶Wilcoxon–Mann–Whitney Test), incorporating ties yielded a statistically significant difference (p-value = 0.02). The average number of months the two females worked until they were promoted was 341, while the average male worked for 59 months before their promotion. The corresponding medians were 341 and 25 months, respectively. The defendant's expert presented the data, given in Table 1, broken out into two time periods ending at the end of a year. The first was from 1965 until the end of 1973 and the second was from 1974 until 1978. The defendant's data includes three more females and eight more males in the defendant's data because their expert included essentially the first year's data *subsequent* to the charge. Furthermore, the three females fell into the seniority categories of 20–29, 30–39 and 70–79 months, i.e., they had much less seniority than the two females who were promoted *prior* to the complaint

The defendant's expert did not utilize the Wilcoxon test; rather he analyzed all the data sets with the *median* test and found no significant difference differences. In contrast with the Wilcoxon analysis of the data in Table 1, the median test did not find the difference in time to promotion data in the pre-charge period to be statistically significant.

Because only two females were promoted from July 1, 1965 until the charge was filed in January 1973, the median test has *zero* power of detecting a difference between the two samples. Thus, I suggested that the plaintiffs' lawyer ask the following series of questions to the defendant's expert on cross exam:

1. What is the difference between the average time to promotion of the male and female pharmacists in Table 1.

   Expected Answer: about 20 years. The actual difference was 23 years as the mean female took 341 months while the mean male took 59 months to be promoted.
2. Suppose the difference in the two means or averages was *50* years, would the median test have found a

statistically significant difference between the times that females had to work before being promoted than males?

Expected Answer: No.

3. Suppose the difference in the two means or averages was *100* years, would the median test have found a statistically significant difference between the times that females had to work before being promoted than males?

   Expected Answer: No.

4. Suppose the difference in the two means or averages was *1,000* years, would the median test have found a statistically significant difference between the times that females had to work before being promoted than males?

   Expected Answer: No.

5. Suppose the difference in the two means or averages was *one million* years, would the median test have found a statistically significant difference between the times that females had to work before being promoted than males?

   Expected Answer: No.

   My thought was that the above sequence of questions would have shown the judge the practical implication of finding a non-significant result with a test that did not have any power, in the statistical sense. Unfortunately, after the lawyer asked the first question, she jumped to the last one. By doing so, the issue was not made clear to the judge. When I asked why, I was told that she felt that the other expert realized the point. Of course, the questions were designed to explain the practical meaning of statistical "power" to the trial judge, not the expert. A while later while describing the trial to another, more experienced lawyer, he told me that after receiving the No answers to the five questions he would have turned to the expert and asked him:

6. As your statistical test could not detect a difference of a million years between the times to promotion of male and female employees, just how long would my client and other females have to work without receiving a promotion before your test would find a statistically significant difference?

This experience motivated me to look further into the power properties of nonparametric tests, especially in the unbalanced sample size setting (Gastwirth and Wang 1987; Freidlin and Gastwirth 2000a). The data from the *Capaci* case is discussed by Finkelstein and Levin (2000, p. 344) and Gastwirth (1988, p. 312) and the need to be cautious when a test with low power accepts the null hypothesis is emphasized by Zeisel and Kaye (1997, p. 88).

To further illustrate the change in practices the employer made one could examine the data for 1974–1978. It turns out the mean (median) time to promotion for males was 65.725 (35) and for females was 11.66 (15). Thus, *after* the charge males had to work at least a year *more* than females before they were promoted to Chief Pharmacist. This is an example of a phenomenon I refer to as "A Funny Thing Happens on the Way to the Courtroom." From both a "common sense" standpoint as well as legal one, the employment actions in the period leading up to the complaint have the most relevance for determining what happened when the plaintiff was being considered for promotion. (Similar issues of timing occur in contract law, where the meaning and conditions of a contract at the time it was signed are used to determine whether it has been properly carried out by both parties. In product liability law, a manufacturer is not held liable for risks that were not known when the product was sold to the plaintiff but were discovered subsequently.) Indeed, quite often a plaintiff applies for promotion on several occasions and only after being denied it on all of them, files a formal charge. (For example, in *Watson v. Fort Worth Bank & Trust*, 487 U.S. 977 (1988) the plaintiff had applied for promotion four times. The opinion indicates that the Justices felt that she was unfairly denied promotion on her fourth attempt.)

Comment 1: Baldus and Cole (1987, p. 190) refer to the dispute concerning the Wilcoxon and Median tests in the *Capaci* case in a section concerning the difference between practical and statistical significance. Let me quote them:

> ► Exclusive concern for the ►statistical significance of a disparity encourages highly technical skirmishes between plaintiff's and defendants' experts who may find competing methods of computing statistical significance advantageous in arguing their respective positions (citing the *Capaci* case). The importance of such skirmishes maybe minimized by limiting the role of a test of significance to that of aiding in the interpretation of a measure of impact whose practical significance may be evaluated on nontechnical grounds.

Thus, even the authors of perhaps a commonly cited text on statistical proof of discrimination at the time did not appreciate the importance of the theory of hypothesis testing and the role of statistical power in choosing between tests. More importantly, a difference in the median time to promotion of 341 – 25 = 316 or about 15 years (or the difference in the average times of 23.5 years) would appear to me to be of practical significance. Thus, well respected authors as well as the judiciary allowed the defendant's expert to use

the median test, which had no power to detect a difference in time to promotion, to obfuscate a practically meaningful difference.

Comment 2: The expert for the defendant was a social scientist rather than a statistician. Other statisticians who have faced experts of this type have mentioned to me that often non-statisticians haven't had sufficient training in our subject to know how one should properly choose a procedure. They may select the first procedure that comes to mind or choose a method that helps their client even though it is quite inappropriate and their ignorance of statistical theory makes it difficult for the lawyer you are working for to get them to admit that their method is not as powerful (or accurate or reliable) as yours. An example of this occurred in a case concerning sex discrimination in pay when an "expert" compared the wages of small groups of roughly similar males and females with the t-test. It is well known that typically income and wage data are quite skewed and that the distribution of the two-sample t-test in small samples depends on the assumption of normality. I provided this information to the lawyer who then asked the other "expert" whether he had ever done any research using income or wage data (Ans. No) and whether he had ever carried out any research or read literature on the *t*-test and its properties (Ans. No). Thus, it was difficult for the lawyer to get this "expert" to admit that using the t-test in such a situation is questionable and the significance levels might not be reliable. On a more positive note, the *Daubert* (509 U.S. 579 (1993)) opinion listed several criteria for courts to evaluate expert scientific testimony, one of which is that the method used has a known error rate. Today one might be able to fit a skewed distribution to the data and then show by simulation that a nominal 0.05 level test has an actual level (α) of 0.10 or more. Similarly, if the one must use the t-test in such a situation one could conduct a simulation study to obtain "correct" critical values that will ensure that a nominal 0.05 level test has true level between 0.04 and 0.06. (Although I use the 0.05 level for illustration, I agree with Judge Posner (2001) that it should not be used as a talisman. Indeed, Judge P. Higginbotham's statement in *Vuyanich v. Republic National Bank*, 505 F. Supp. 224 (N.D. Texas 1980) that the p-value is a sliding-scale makes more sense than a simple yes-no dichotomy of significance or not in the legal context as the statistical evidence is only part of the story. The two-sided p-value 0.06 on the post-charge time until promotion data from the Capaci case illustrates the wisdom of the statements of these judges. Not only do the unbalanced sample sizes diminish the power of two-sample tests, the change in the promotion practices of the defendant subsequent to the charge are quite clear from the change in difference in average waiting times until promotion of males and females as well as the change from a significant difference in favor of males before the charge to a nearly significant change in favor of females after the charge.)

Another way of demonstrating that an expert does not possess relevant knowledge is for the lawyer to show them a book or article that states the point you wish to make and ask the expert to read it and then say they agree or disagree with the point. (Dr. Charles Mann told the author that he has been able to successfully use this technique.) If that expert disagrees, a follow-up question can inquire why or on what grounds does he or she disagree with it.

The tradition of reporting data by year also makes it more difficult to demonstrate that a change occurred after a charge was filed. In *Valentino v. USPS* (674 F.2d 56 (DC Circ. 1982)) the plaintiff had applied for a promotion in 1976 and filed the charge in May, 1976. The data is reported in Table 2 and has reanalyzed by Freidlin and Gastwirth (2000b) and Kadane and Woodworth (2004), suggested that females received fewer promotions than expected in the two previous years but after 1976 they received close to their expected number. Let me recall the data and analysis the yearly summaries enabled us to present.

The data for each year were analyzed by the Mantel-Haenszel test applied to the 2×2 tables for each grade grouping in Tables 2 and 3. This was done because the Civil Service Commission reported its data in this fashion. Notice that during two time periods, the number of grade advancements awarded to females for each year is significantly less than expected. After 1976, when the charge was filed the female employees start to receive their expected number of promotions.

My best recollection is that the promotion the plaintiff applied for was the 34th competitive one filled in 1976. Unfortunately, data on all of the applicants was not available even though EEOC guidelines do require employers to preserve records for at least 6 months. Of the 17 or so positions for which data on the actual applicants was available every one was given to a male candidate. Since females did receive their expected number of promotions over the entire year it would appear that the defendant changed its practices after the charge and consequently prevailed in the case. (The data discussed in the cited references considered employees in job categories classified by their level in the system. The district court accepted the defendant's criticism that since each job level contains positions in a variety of occupations, the plaintiffs should have stratified the data by occupation; see 511 F. Supp 917, 939 (D.C. DC, 1981). Later in the opinion, at 511 F. Supp. 951, the opinion accepted a regression analysis submitted by the defendant

Presentation of Statistical Testimony. Table 2 Number of employees and promotions they received: from the *Valentino v. U.S.P.S Case*

	Grade 17–19		Grade 20–22		Grade 23–25		Grade 26–28		Grade 29–31	
Time period	M	F	M	F	M	F	M	F	M	F
06/74–03/75	229	73	360	48	703	33	236	7	82	1
	67	5	74	9	132	2	28	1	8	0
03/75–01/76	205	89	373	43	716	36	277	9	85	1
	40	6	39	5	41	1	19	0	7	0
01/76–01/77	233	101	396	52	727	36	271	9	85	2
	31	10	32	4	54	5	28	2	5	0
01/77–01/78	200	86	377	52	680	35	262	8	89	3
	43	18	80	9	57	6	18	1	14	0
01/78–01/79	196	90	325	50	685	37	252	9	78	3
	29	8	45	7	42	3	14	1	6	1

Key to symbols: F=females; M=males; for any time period and grade group the number of promotions is below the number of employees. For example, in grades 17–19 during 06/74–03/75 period 67 out of 229 eligible males were promoted compared to 5 out of 73 eligible females

Presentation of Statistical Testimony. Table 3 The results obtained from Mantel-Haenszel test for equality of promotion rates applied to the stratified data for each period in Table 2

Year	Observed	Expected	p-value (two-sided)
1974–1975	17	34.1	0.0006
1975–1976	12	21.16	0.020
1976–1977	21	20.32	0.885
1977–1978	34	33.23	0.869
1978–1979	20	21.66	0.674

that used grade level as a predictor of salary noting that 'level" is a good proxy for occupation. While upholding the ultimate finding that U.S.P.S did not discriminate against women, the appellate opinion, fn. 15 at 71, did not accept the district court's criticism that a regression submitted by plaintiffs should have included the level of a position. The reason is that in a promotion case, it is advancement in one's job level that is the issue. Thus, courts do accept regressions that include the major other job-related characteristics such as experience, education, special training and objective measures of productivity.) There was one unusual aspect of the case; the Postal Service had changed the system of reviewing candidates for promotions as of January 1, 1976. As no other charges of discrimination in promotion had been filed in either 1974 or 1975, the analysis of data for those years is considered background information unless the plaintiff can demonstrate that the same process carried over into the time period when the promotion in question was made. (In *Evans v. United Airlines*, the Supreme Court stated that since charges of employment discrimination need to be filed within 180 days of the charge, earlier data is useful background information but is not sufficient by itself to establish a *prima facie* case of discrimination. If there is a continuing violation, however, then the earlier data can be used in conjunction with more recent data by the plaintiffs. The complex legal issues involved in determining when past practices have continued into the time period relevant to a particular case are beyond the scope of the present paper.)

When data is reported in yearly periods, invariably some post-charge data will be included in the data for the year in which the charge was file and statisticians should examine it to see if there is evidence of a change in employment practice subsequent to the charge. In the *Capaci* case, the defendant included eleven months of post-charge data in their first time period, 1965–1973. The inclusion of three additional females promoted in that period lessens the impact of the fact that only two female Pharmacists were promoted during the previous seven and a half years. Similarly, reporting the data by year in *Valentino* enabled the

defendant to monitor the new promotion system begun at the start of 1976 subsequent to the complaint, so females did receive their expected number of promotions for the 1976, the year of the charge.

A More Informative Summary of Promotion Data: Hogan v. Pierce (31 F.E.P. 115 (D.D.C. 1983))

The plaintiff in the case alleged that he had been denied a promotion to a GS-14 level position in the computer division of a government agency and filed a formal complaint in 1977. As the files on the actual applicants were unavailable, we considered all individuals employed in GS-13 level jobs whose records indicated that they met the Civil Service criteria for a promotion to the next job-level to be the pool of "qualified applicants." (These qualifications were that they were employed in an appropriate computer-related position and had at lest one year of experience at the previous level (GS-13) or its equivalent.) There were about ten opportunities for promotion to a GS-14 post during the several years prior to the complaint. About three of the successful GS-14 job applicants were outside candidates who had a Ph.D. in computer science; all of whom were white. Since they had a higher level of education than the internal candidates, they are excluded from Table 4, which gives the number of employees who were eligible and the number promoted, by race, for the ten job announcements.

Although the data is longitudinal in nature and technically one might want to apply a survival test such as the log-rank procedure, it was easier to analyze the data by the Mantel-Haenszel (MH) test that combines the observed minus expected numbers from the individual the individual 2×2 tables (this is, of course, the way the log-rank test is also computed and the resulting statistics are the same). Although there were only 18 promotions awarded to internal candidates during the period under consideration *none* of them went to a black. Moreover, the exact p-value of the MH test was 0.007, a clearly significant result. The analysis can be interpreted as a test for the effect of race controlling for eligibility by Feinberg (1989, p. 100) and has been discussed by Agresti (1996) in the STATXACT manual (2003, p. 597) where it is shown that the *lower* end of a 95% confidence interval of the odds a white employee receives a promotion relative to a minority employee is about two. Thus, we can conclude that the odds of a black employee had of being promoted were half those of a white, which is clearly of practical as well as statistical significance.

In order to demonstrate that the most plausible potential explanation of the promotion data, the white employees had greater seniority, data was submitted that showed that by 1977 the average black employee had worked over a year more at the GS-13 level than the average white one. Thus, if seniority were a major factor that could explain why whites received the earlier promotions, it should have worked in favor of the black employees in the later part of the period. The defendant did not suggest an alternative analysis of this data but concentrated on post-charge data but Judge A. Robinson observed at the trial that their analysis should have considered a time frame around the time of the complaint.

Presentation of Statistical Testimony. Table 4
Promotion data for GS-14 positions obtained by internal job candidates, by Race, from *Hogan v. Pierce* (From Plaintiff's exhibit on file with D.C. District Court. Reproduced in Gastwirth (1988, p. 266) and in the STATXACT manual (Version 6, p. 597))

	Whites		Blacks	
Date of Promotion	Eligible	Promoted	Eligible	Promoted
July 1974	20	4	7	0
August 1974	17	4	7	0
Sept. 1974	15	2	8	0
April 1975	18	1	8	0
May 1975	18	1	8	0
Oct. 1975	30	1	10	0
Nov. 1975	31	2	10	0
Feb. 1976	31	1	10	0
March 1976	31	1	10	0
Nov. 1977	34	1	13	0

A Few Suggestions to Statisticians Desiring to Increase the Validity of and Weight Given to Statistical and Scientific Evidence

Mann (2000), cited and endorsed by Mallows (2003), noted that if your analysis does not produce results the lawyer who hired you desired; you will likely be replaced. Indeed, he notes that many attorneys act as though they will be able to find a statistician who will give them the results they want and that "regrettably they may often be correct." This state of affairs unfortunately perpetuates the statement that there are "lies, damn lies and statistics" attributed to both B. Disraeli and M. Twain. Statisticians

have suggested questions that judges may ask experts (Feinberg et al. 1995) and discussed related ethical issues arising in giving testimony (Meier 1986; Kadane 2005). Here we mention a few more suggestions for improving statistical testimony and close with a brief discussion questioning the wisdom of a recent editorial that appeared in *Science*.

1. Mann (2000) is correct in advising statisticians who are asked to testify inquire as to the existence of other data or the analysis of any other previous expert the lawyer consulted. I would add that these are especially important considerations if you are asked to examine data very shortly before a trial as you will have a limited time to understand the data and how it was generated so you will need to totally rely on the data and background information the lawyer provides. In civil cases, it is far preferable for an expert to be involved during the period when discovery is being carried out. Both sides are asking the other for relevant data and information and it is easier for you to tell the lawyer the type of data that you feel would help answer the relevant questions. If the lawyers ignore your requests or don't give you a sensible justification (let me give an example of a business justification. You are asked to work for a defendant in an equal employment case concerning the fairness of a layoff carried out in a plant making brakes for cars and SUVs. Data on absenteeism would appear to be quite useful. The employer knows that the rate of absenteeism in the plant was "higher" than normal for the industry and might be concerned that if this information was put into the public record, they might become involved in many suits arising from accidents involving cars with brakes made there. Since the corporation is a profit-making organization not a scientific one, they will carry out a "cost-benefit" analysis to decide whether you will be given the data), you should become concerned.

   The author's experience in *Valentino* illustrates this point. One reason the original regression equation we developed for the plaintiffs was considered incomplete was because it did not contain relevant information concerning any special degrees the employees had. This field was omitted from the original file we were given and we only learned about it a week or so before trial. After finding that employees with business, engineering or law degrees received about the same salary increase, we then included a single indicator for having a degree in one of the three areas. To assist the court, it would have been preferable to use separate indicators for each degree. (Both opinions, see 674 F.2d 56, 70 fn. 21, downplayed this factor because the defendant's expert testified that it was unreliable as the information depended on whether applicants listed any specialty on their form. More importantly, use of the single indicator variable for three different degrees may not have made it clear that they all had a similar effect on salary and that the indicator was restricted to employees in these three specialties.) Given the very short time available to incorporate the new, but important information that did reduce the estimated coefficient on sex by a meaningful amount, although it remained significant, one did not have time to explore how best to present the analysis.
2. When you are approached by counsel, you should ask for permission to write an article about any interesting statistical problems or data arising in the case. (Of course, data and exhibits submitted formally into evidence are generally considered to be in the public domain so one can use them in scholarly writings. Sometimes, however, the data in the exhibits are summaries and either the individual data or summaries of smaller sub-groups are required to illustrate the methodology.) While the lawyers and client might request that some details of the case not be reported so that the particular case or parties involved are not identifiable, you should be given permission to show the profession how you used or adapted existing methodology. If you perceive that the client or lawyer want to be very restrictive about the use and dissemination of the data and your analysis, you should think seriously about becoming involved. I realize that it is much easier for an academic statistician to say "no" in this situation, than statisticians who do consulting for a living; especially if they have a long-term business relationship with a law firm.
3. Statisticians are now being asked by judges to advise them in "*Daubert*" hearings, which arise when one party in the case challenges the scientific validity or reliability of the expert testimony the other side desires to use. This is a useful service that all scientific professions can provide the legal system. Before assessing a proposed expert's testimony and credentials, it is wise for you to look over the criteria the court suggested in the *Daubert* opinion as well as some examples of decisions in such matters. Because courts take a skeptical view of "new techniques" that have not been subject to peer review but appear to have been developed specifically for the case at hand, one should not fault an expert who does not use the most recently developed method but uses a previously existing method that has been generally regarded in the field as appropriate for

the particular problem. The issue is not whether you would have performed the same analysis the expert conducted but whether the analysis is appropriate and statistically sound. Indeed, Kadane (1990) conducted a Bayesian analysis of data arising in an equal employment and confirmed it with a frequentist method used in ►biostatistics that had been suggested for the problem by Gastwirth and Greenhouse (1987). He noted that it is reassuring when two approaches lead to the same inference. In my experience, this often occurs when data sets of moderate to large sample size are examined and one uses some common sense in interpreting any difference. (By common sense I mean that if one statistician analyzes a data set with test A and obtains a p-value of 0.04, i.e., a statistically significant one, but the other statistician uses another appropriate test B and obtains a p-value of 0.06, a so-called non-significant one, the results are really not meaningfully different.)

4. A "*Daubert*" hearing can be used to provide the judge with questions to ask an expert about the statistical methodology utilized. (This also gives the court's expert the opportunity to find the relevant portions of books and articles that can be shown the expert, thereby implementing the approach described in Comment 2 of section "►A More Informative Summary of Promotion Data: Hogan v. Pierce (31 F.E.P. 115 (D.D.C. 1983))".) This may be effective in demonstrating to a court that a non-statistician really does not have a reasonable understanding of some basic concepts such as power or the potential effect of violations of the assumptions underlying the methods used.
5. One task experts are asked to perform is to criticize or raise questions about the findings and methodology of the testimony given by the other party's expert. This is one place where is easy for one to become overly partisan and lose their scientific objectivity. This important problem is discussed by (Meier 1986). Before raising a criticism, one should ask whether it is important. Could it make a difference in either the conclusion or the weight given to it? In addition to checking that the assumptions underlying the analysis they are presenting are basically satisfied, one step an expert can carry out to protect their own statistical analyses from being unfairly criticized is to carry out a ►sensitivity analysis. Many methods for assessing whether an omitted variable could change an inference have been developed (Rosenbaum 2002) and van Belle (2002) has discussed the relative importance of violations of the assumptions underlying commonly used techniques.
6. While this entry focused on statistical testimony in civil cases, similar issues arise in the use of statistical evidence in criminal cases. The reader is referred to Aitken and Taroni (2004), Balding (2005) for the commonly used statistical methods used in this setting. Articles by a number of experts in both civil and criminal cases appear in Gastwirth (2000) provide additional perspectives on issues arising in the use of statistical evidence in the legal setting.

Acknowledgments

This entry was adapted from Gastwirth (2005). The author wishes to thank Prof. M. Lovic for his helpful suggestions.

About the Author

Professor Joseph L. Gastwirth is a Fellow of ASA (1970), IMS (1972), and AAS (1971), and an Elected member of ISI (1980). He was President of the Washington Statistical Society (1982–1983). He has served three times on the American Statistical Association's Law and Justice Statistics Committee and currently is its Chair. He has written over 150 articles concerning both the theory and application of statistics; especially in the legal setting. In 1985, he received a Guggenheim Fellowship, which led to his two-volume book *Statistical Reasoning in Law and Public Policy* (1988), and the Shiskin Award (1998) for Research in Economic Statistics for his development of statistical methods for measuring economic inequality and discrimination. He edited the book *Statistical Science in the Courtroom* (2000) and in 2002, his article with Dr. B. Freidlin using change-point methods to analyze data arising in equal employment cases shared the "Applications Paper of the Year" award of the American Statistical Association. He has served on the editorial boards of several major statistical journals and recently, he was appointed Editor of the journal, *Law, Probability and Risk*. On August 1, 2009, a workshop celebrating 45 years of statistical activity of Professor Gastwirth was organized at the George Washington University to "recognize his outstanding contributions to the development of nonparametric and robust methods and their use in genetic epidemiology, his pioneering research in statistical methods and measurement for the analysis of data arising in the areas of economic inequality, health disparities and equal employment, and in other legal applications".

Cross References

►Frequentist Hypothesis Testing: A Defense
►Null-Hypothesis Significance Testing: Misconceptions
►Power Analysis
►P-Values

►Significance Testing: An Overview
►Significance Tests: A Critique
►Statistical Evidence
►Statistics and the Law
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Agresti A (1996) An introduction to categorical data analysis. Wiley, New York
Aitken C, Taroni F (2004) Statistics and the evaluation of evidence for forensic scientists, 2nd edn. Wiley, Chichester, UK
Balding DJ (2005) Weight of the evidence for forensic DNA profiles. Wiley, Chichester, UK
Baldus DC, Cole JWL (1987) Statistical proof of discrimination: cumulative supplement. Shepard's/McGraw-Hill, Colorado Springs
Berger MA (2000) The supreme court's trilogy on the admissibility of expert testimony. In: Reference manual on scientific evidence, 2nd edn. Federal Judicial Center, Washington, DC, pp 9–38
Feinberg SE (1989) The evolving role of statistical assessments as evidence in courts. Springer, New York
Feinberg SE, Krislov SH, Straf M (1995) Understanding and evaluating statistical evidence in litigation. Jurimetrics J 36:1–32
Finkelstein MO, Levin B (2000) Statistics for lawyers, 2nd edn. Springer, New York
Freidlin B, Gastwirth JL (2000a) Should the median test be retired from general use? Am Stat 54:161–164
Freidlin B, Gastwirth JL (2000b) Change point tests designed for the analysis of hiring data arising in equal employment cases. J Bus Econ Stat 18:315–322
Gastwirth JL, Greenhouse SW (1987) Estimating a common relative risk: application in equal employment. J Am Stat Assoc 82:38–45
Gastwirth JL, Wang JL (1987) Nonparametric tests in small unbalanced samples: application in employment discrimination cases. Can J Stat 15:339–348
Gastwirth JL (1988) Statistical reasoning in law and public policy vol. 1 statistical concepts and issues of fairness. Academic, Orlando, FL
Gastwirth JL (ed) (2000) Statistical science in the courtroom. Springer, NY
Gastwirth JL (2005) Some issues arising in the presentation of statistical testimony. Law, Probability and Risk 4:5–20
Kadane JB (1990) A statistical analysis of adverse impact of employer decisions. J Am Stat 85:925–933
Kadane JB (2005) Ethical issues in being an expert witness. Law, Probability and Risk 4:21–23
Kadane JB, Woodworth GG (2004) Hierarchical models for employment decisions. J Bus Econ Stat 22:182–xx
Mallows C (2003) Parity: implementing the telecommunications act of 1996. Stat Sci 17:256–270
Mann CR (2000) Statistical consulting in the legal environment. In: Gastwirth JL (ed) Statistical science in the courtroom. Springer, New York, pp 245–262
Meier P (1986) Damned liars and expert witnesses. J Am Stat Assoc 81:269–276
Rosenbaum PR (2002) Observational studies, 2nd edn. Springer, New York
Rosenblum M (2000) On the evolution of analytic proof, statistics and the use of experts in EEO litigation. In: Gastwirth JL (ed) Statistical science in the courtroom. Springer, New York, pp 161–194
van Belle G (2002) Statistical rules of thumb. Wiley, New York
Zeisel H, Kaye DH (1997) Prove it with figures: empirical methods in law and litigation. Springer, New York

Principal Component Analysis

Ian Jolliffe
Professor
University of Exeter, Exeter, UK

Introduction

Large or massive data sets are increasingly common and often include measurements on many variables. It is frequently possible to reduce the number of variables considerably while still retaining much of the information in the original data set. Principal component analysis (PCA) is probably the best known and most widely used dimension-reducing technique for doing this. Suppose we have n measurements on a vector $\mathbf{x}$ of p random variables, and we wish to reduce the dimension from p to q, where q is typically much smaller than p. PCA does this by finding linear combinations, $\mathbf{a}_1'\mathbf{x}, \mathbf{a}_2'\mathbf{x}, \ldots, \mathbf{a}_q'\mathbf{x}$, called *principal components*, that successively have maximum variance for the data, subject to being uncorrelated with previous $\mathbf{a}_k'\mathbf{x}$s. Solving this maximization problem, we find that the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q$ are the eigenvectors of the covariance matrix, $\mathbf{S}$, of the data, corresponding to the q largest eigenvalues (see ►Eigenvalue, Eigenvector and Eigenspace). The eigenvalues give the variances of their respective principal components, and the ratio of the sum of the first q eigenvalues to the sum of the variances of all p original variables represents the proportion of the total variance in the original data set accounted for by the first q principal components. The familiar algebraic form of PCA was first presented by Hotelling (1933), though Pearson (1901) had earlier given a geometric derivation. The apparently simple idea actually has a number of subtleties, and a surprisingly large number of uses, and has a vast literature, including at least two comprehensive textbooks (Jackson 1991; Jolliffe 2002).

An Example

As an illustration we use an example that has been widely reported in the literature, and which is originally due to

Principal Component Analysis. Table 1 Principal Component Analysis Vectors of coefficients for the first two principal components for data from Yule et al. (1969)

Variable	a_1	a_2
x_1	0.34	0.39
x_2	0.34	0.37
x_3	0.35	0.10
x_4	0.30	0.24
x_5	0.34	0.32
x_6	0.27	−0.24
x_7	0.32	−0.27
x_8	0.30	−0.51
x_9	0.23	−0.22
x_{10}	0.36	−0.33

Yule et al. (1969). The data consist of scores, between 0 and 20, for 150 children aged $4^1/_2 - 6$ years from the Isle of Wight, on ten subtests of the Wechsler Pre-School and Primary Scale of Intelligence. Five of the tests were "verbal" tests and five were 'performance' tests. Table 1 gives the vectors $\mathbf{a}_1, \mathbf{a}_2$ that define the first two principal components for these data.

The first component is a linear combination of the ten scores with roughly equal weight (maximum 0.36, minimum 0.23) given to each score. It can be interpreted as a measure of the overall ability of a child to do well on the full battery of ten tests, and represents the major (linear) source of variability in the data. On its own it accounts for 48% of the original variability. The second component contrasts the first five scores (verbal tests) with the five scores on the performance tests. It accounts for a further 11% of the total variability. The form of this second component tells us that once we have accounted for overall ability, the next most important (linear) source of variability in the tests scores is between those children who do well on the verbal tests *relative to* the performance tests and those children whose test score profile has the opposite pattern.

Covariance or Correlation

Principal components successively maximize variance, and can be found from the eigenvalues/eigenvectors of a covariance matrix. Often a modification is adopted, in order to avoid two problems. If the p variables are measured in a mixture of units, then it is difficult to interpret the principal components. What is meant by a linear combination of weight, height and temperature, for example? Furthermore, if we measure temperature and weight in °F and pounds respectively, we may get completely *different principal components* from those obtained from the *same data* but using °C and kilograms. To avoid this arbitrariness, we standardize each variable to have zero mean and unit variance. Finding linear combinations of these standardized variables that successively maximize variance, subject to being uncorrelated with previous linear combinations, leads to principal components defined by the eigenvalues and eigenvectors of the correlation matrix, rather than the covariance matrix, of the original variables. When all variables are measured in the same units, covariance-based PCA may be appropriate, but even here they can be uninformative when a few variables have much larger variances than the remainder. In such cases the first few components are dominated by the high-variance variables and tell us little that could not have been deduced by inspection of the original variances. Circumstances exist where covariance-based PCA is of interest but most PCAs encountered in practice are correlation-based. Our example is a case where either approach could be used. The results given above are based on the correlation matrix but, because the variances of all 10 tests are similar, results from a covariance-based analysis would be little different.

How Many Components?

We have talked about q principal components accounting for most of the variation in the p variables? What is meant by "most" and, more generally, how do we decide how many components to keep? There is a large literature on this topic – see, for example, Jolliffe (2002), Chap. 6. Perhaps the simplest procedure is to set a threshold, say 80%, and stop when the first q components account for a percentage of total variation greater than this threshold. In our example the first two components accounted for only 59% of the variation. The threshold is often set higher than this – 70% to 90% are the usual sort of values, but it depends on the context of the data set and can be higher or lower. Other techniques are based on the values of the eigenvalues or on the differences between consecutive eigenvalues. Some of these simple ideas, as well as more sophisticated ones (Jolliffe 2002, Chap. 6) have been borrowed from factor analysis (see ▶Factor Analysis and Latent Variable Modelling). This is unfortunate because the different objectives of PCA and factor analysis (see below for more on this) mean that typically fewer dimensions should be retained in

factor analysis than in PCA, so the factor analysis rules are often inappropriate. It should also be noted that although it is usual to discard low-variance principal components, they can sometimes be useful in their own right, for example in finding ►outliers (Jolliffe 2002, Chap. 10) and in quality control (Jackson 1991).

Confusion with Factor Analysis

There is much confusion between principal component analysis and factor analysis, partly because some widely used software packages treat PCA as a special case of factor analysis, which it is not. There are several technical differences between PCA and factor analysis, but the most fundamental difference is that factor analysis explicitly specifies a model relating the observed variables to a smaller set of underlying unobservable factors. Although some authors express PCA in the framework of a model, its main application is as a descriptive, exploratory technique, with no thought of an underlying model. This descriptive nature means that distributional assumptions are unnecessary to apply PCA in its usual form. It can be used, although caution may be needed in interpretation, on discrete and even binary data, as well as continuous variables. One notable feature of factor analysis is that it is generally a two-stage procedure; having found an initial solution, it is rotated towards simple structure. This idea can be borrowed and used in PCA; having decided to keep q principal components, we may rotate within the q-dimensional subspace defined by the components in a way that makes the axes as easy as possible to interpret. This is one of number of techniques that attempt to simplify the results of PCA by post-processing them in some way, or by replacing PCA with a modified technique (Jolliffe 2002, Chap. 11).

Uses of Principal Component Analysis

There are many variations on the basic use of PCA as a dimension reducing technique whose results are used in a descriptive/exploratory manner – see Jackson (1991), Jolliffe (2002). PCA is often used a first step, reducing dimensionality before undertaking another technique, such as multiple regression, cluster analysis (see ►Cluster Analysis: An Introduction), discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems) or independent component analysis.

Extensions to Principal Component Analysis

PCA has been extended in many ways. For example, one restriction of the technique is that it is linear. A number of non-linear versions have therefore been suggested. These include the Gifi approach to multivariate analysis. Another area in which many variations have been proposed is when the data are time series, so that there is dependence between observations as well as between variables (Jolliffe 2002, Chap. 12). There are many other extensions and modifications, and the list continues to grow.

Acknowledgments

This article is a revised and shortened version of an entry that appeared in The Encyclopedia of Statistics in Behavioral Science, published by Wiley.

About the Author

Professor Ian Jolliffe is Honorary Visiting Professor at the University of Exeter. Before his early retirement he was Professor of Statistics at the University of Aberdeen. He received the International Meetings in Statistical Climatology Achievement Award in 2004 and was elected a Fellow of the American Statistical Association in 2009. He has authored, co-authored or co-edited four books, including *Principal Component Analysis* (Springer, 2nd edition, 2002) and *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (jointly edited with D B Stephenson, Wiley, 2003, 2nd edition due 2011). He has also published over 80 papers in peer-reviewed journal and is currently Associate Editor for *Weather and Forecasting*.

Cross References

►Analysis of Multivariate Agricultural Data
►Data Analysis
►Eigenvalue, Eigenvector and Eigenspace
►Factor Analysis and Latent Variable Modelling
►Fuzzy Logic in Statistical Data Analysis
►Multivariate Data Analysis: An Overview
►Multivariate Reduced-Rank Regression
►Multivariate Statistical Analysis
►Multivariate Technique: Robustness
►Partial Least Squares Regression Versus Other Methods

References and Further Reading

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441, 498–520
Jackson JE (1991) A user's guide to principal components. Wiley, New York
Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2:559–572
Yule W, Berger M, Butler S, Newham V, Tizard J (1969) The WPPSI: an empirical evaluation with a British sample. Brit J Educ Psychol 39:1–13

Principles Underlying Econometric Estimators for Identifying Causal Effects

James J. Heckman
Winner of the Nobel Memorial Prize in Economic Sciences in 2000, Henry Schultz Distinguished Service Professor of Economics
The University of Chicago, Chicago, IL, USA
University College Dublin, Dublin, Ireland

This paper reviews the basic principles underlying the identification of conventional econometric evaluation estimators for causal effects and their recent extensions. Heckman (2008) discusses the econometric approach to causality and compares it to conventional statistical approaches. This paper considers alternative methods for identifying causal models.

The paper is in four parts. The first part presents a prototypical economic choice model that underlies econometric models of causal inference. It is a framework that is useful for analyzing and motivating the economic assumptions underlying alternative estimators. The second part discusses general identification assumptions for leading econometric estimators at an intuitive level. The third part elaborates the discussion of matching in the second part. Matching is widely used in applied work and makes strong informational assumptions about what analysts know relative to what the people they analyze know. The fourth part concludes.

A Prototypical Policy Evaluation Problem

Consider the following prototypical policy problem. Suppose a policy is proposed for adoption in a country. It has been tried in other countries and we know outcomes there. We also know outcomes in countries where it was not adopted. From the historical record, what can we conclude about the likely effectiveness of the policy in countries that have not implemented it?

To answer questions of this sort, economists build models of counterfactuals. Consider the following model. Let Y_0 be the outcome of a country (e.g., GDP) under a no-policy regime. Y_1 is the outcome if the policy is implemented. $Y_1 - Y_0$ is the "treatment effect" or causal effect of the policy. It may vary among countries. We observe characteristics X of various countries (e.g., level of democracy, level of population literacy, etc.). It is convenient to decompose Y_1 into its mean given X, $\mu_1(X)$ and deviation from mean U_1. One can make a similar decomposition for Y_0:

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1 \\ Y_0 &= \mu_0(X) + U_0. \end{aligned} \tag{1}$$

Additive separability is not needed, but it is convenient to assume it, and I initially adopt it to simplify the exposition and establish a parallel regression notation that serves to link the statistical literature on treatment effects with the economic literature. (Formally, it involves no loss of generality if we define $U_1 = Y_1 - E(Y_1 \mid X)$ and $U_0 = U_0 - E(Y_0 \mid X)$.)

It may happen that controlling for the X, $Y_1 - Y_0$ is the same for all countries. This is the case of homogeneous treatment effects given X. More likely, countries vary in their responses to the policy even after controlling for X.

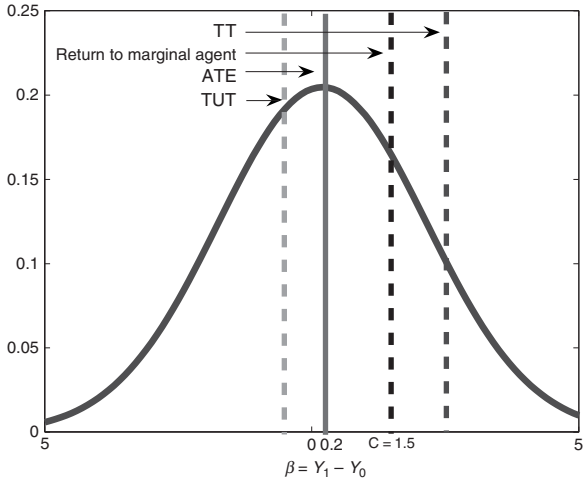
Figure 1 plots the distribution of $Y_1 - Y_0$ for a benchmark X. It also displays the various conventional treatment parameters. I use a special form of a "generalized Roy" model with constant cost C of adopting the policy (see Heckman and Vytlacil 2007a, for a discussion of this model). This is called the "extended Roy model." I use this model because it is simple and intuitive. (The precise parameterization of the extended Roy model used to generate the figure and the treatment effects is given at the base of Fig. 1.) The special case of homogeneity in $Y_1 - Y_0$ arises when the distribution collapses to its mean. It would be ideal if one could estimate the distribution of $Y_1 - Y_0$ given X and there is research that does this.

More often, economists focus on some mean of the distribution in the literature and use a regression framework to interpret the data. To turn (1) into a regression model, it is conventional to use the switching regression framework. (Statisticians sometimes attribute this representation to Rubin (1974, 1978), but it is due to Quandt (1958, 1972). It is implicit in the Roy (1951) model. See the discussion of this basic model of counterfactuals in Heckman and Vytlacil (2007a)). Define $D = 1$ if a country adopts a policy; $D = 0$ if it does not. Substituting (1) into this expression, and keeping all X implicit, one obtains

$$\begin{aligned} Y &= Y_0 + (Y_1 - Y_0)D \\ &= \mu_0 + (\mu_1 - \mu_0 + U_1 - U_0)D + U_0. \end{aligned} \tag{2}$$

This is the Roy-Quandt "switching regression" model. Using conventional regression notation,

$$Y = \alpha + \beta D + \varepsilon \tag{3}$$



TT = 2.666, TUT = –0.632
Return to Marginal Agent = $C = 1.5$
ATE = $\mu_1 - \mu_0 = \bar{\beta} = 0.2$

The Model

Outcomes	Choice Model
$Y_1 = \mu_1 + U_1 = \alpha + \bar{\beta} + U_1$ $Y_0 = \mu_0 + U_0 = \alpha + U_0$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$
General Case	
$(U_1 \neq U_0) \not\perp\!\!\!\perp D$ ATE ≠ TT ≠ TUT	

The Researcher Observes (Y, D, C)

$Y = \alpha + \beta D + U_0$ where $\beta = Y_1 - Y_0$

Parameterization

$\alpha = 0.67 \quad (U_1, U_0) \sim N(\mathbf{0}, \Sigma) \quad D^* = Y_1 - Y_0 - C$

$\bar{\beta} = 0.2 \quad \Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix} \quad C = 1.5$

Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 1 Distribution of gains, the Roy economy (Heckman et al. 2006)

where $\alpha = \mu_0$, $\beta = (Y_1 - Y_0) = \mu_1 - \mu_0 + U_1 - U_0$ and $\varepsilon = U_0$. I will also use the notation that $\upsilon = U_1 - U_0$, letting $\bar{\beta} = \mu_1 - \mu_0$ and $\beta = \bar{\beta} + \upsilon$. Throughout this paper I use treatment effect and regression notation interchangeably. The coefficient on D is the treatment effect. The case where β is the same for every country is the case conventionally assumed. More elaborate versions assume that β depends on X ($\beta(X)$) and estimates interactions of D with X. The case where β varies even after accounting for X is called the "random coefficient" or "heterogenous treatment effect" case. The case where $\upsilon = U_1 - U_0$ depends on D is the case of essential heterogeneity analyzed by Heckman et al. (2006). This case arises when treatment choices depend at least in part on the idiosyncratic return to treatment. A great deal of attention has been focused on this case in recent decades and I develop the implications of this model in this paper.

An Index Model of Choice and Treatment Effects: Definitions and Unifying Principles

I now present the model of treatment effects developed in Heckman and Vytlacil (1999, 2001, 2005, 2007a,b) and Heckman et al. (2006), which relaxes the normality, separability and exogeneity assumptions invoked in the traditional economic selection models. It is rich enough to generate all of the treatment effects in the program evaluation literature as well as many other policy parameters. It does not require separability. It is a nonparametric generalized Roy model with testable restrictions that can be used to unify the treatment effect literature, identify different treatment effects, link the literature on treatment effects to the literature in structural econometrics and interpret the implicit economic assumptions underlying ▶instrumental variables, regression discontinuity design methods, control functions and matching methods.

Y is the measured outcome variable. It is produced from the switching regression model (2). Outcomes are general nonlinear, nonseparable functions of observables and unobservables:

$$Y_1 = \mu_1(X, U_1) \tag{4}$$

$$Y_0 = \mu_0(X, U_0). \tag{5}$$

Examples of models that can be written in this form include conventional latent variable models for discrete choice that are generated by a latent variable crossing a threshold: $Y_i = \mathbf{1}\left(Y_i^* \geq 0\right)$, where $Y_i^* = \mu_i(X) + U_i, i = 0, 1$. Notice that in the general case, $\mu_i(X, U_i) - E(Y_i \mid X) \neq U_i$, $i = 0, 1$.

The individual treatment effect associated with moving an otherwise identical person from "0" to "1" is $Y_1 - Y_0 = \Delta$ and is defined as the causal effect on Y of a *ceteris paribus* move from "0" to "1". To link this framework to the literature on economic choice models, I characterize the

decision rule for program participation by an index model:

$$D^* = \mu_D(Z) - V; \qquad D = 1 \quad \text{if} \quad D^* \geq 0; \qquad D = 0 \quad \text{otherwise}, \tag{6}$$

where, from the point of view of the econometrician, (Z, X) is observed and (U_1, U_0, V) is unobserved. The random variable V may be a function of (U_1, U_0). For example, in the original Roy Model, μ_1 and μ_0 are additively separable in U_1 and U_0 respectively, and $V = -[U_1 - U_0]$. In the original formulations of the generalized Roy model, outcome equations are separable and $V = -[U_1 - U_0 - U_C]$, where U_C arises from the cost function. Without loss of generality, I define Z so that it includes all of the elements of X as well as any additional variables unique to the choice equation.

I invoke the following assumptions that are weaker than those used in the conventional literature on structural econometrics or the recent literature on semiparametric selection models and at the same time can be used both to define and to identify different treatment parameters. (A much weaker set of conditions is required to define the parameters than is required to identify them. See Heckman and Vytlacil (2007b, Appendix B).) The assumptions are:

(A-1) (U_0, U_1, V) *are independent of* Z *conditional on* X *(Independence);*

(A-2) $\mu_D(Z)$ *is a nondegenerate random variable conditional on* X *(Rank Condition);*

(A-3) *The distribution of* V *is continuous; (Absolutely continuous with respect to Lebesgue measure.)*

(A-4) *The values of* $E|Y_1|$ *and* $E|Y_0|$ *are finite (Finite Means);*

(A-5) $0 < \Pr(D = 1 \mid X) < 1$.

(A-1) assumes that V is independent of Z given X and is used below to generate counterfactuals. For the definition of treatment effects one does not need either (A-1) or (A-2). The definitions of treatment effects and their unification do not require any elements of Z that are not elements of X or independence assumptions. However, an analysis of instrumental variables requires that Z contain at least one element not in X. Assumptions (A-1) or (A-2) justify application of instrumental variables methods and nonparametric selection or control function methods. Some parameters in the recent IV literature are defined by an instrument so I make assumptions about instruments up front, noting where they are not needed. Assumption (A-4) is needed to satisfy standard integration conditions. It guarantees that the mean treatment parameters are well defined. Assumption (A-5) is the assumption in the population of both a treatment and a control group for each X. Observe that there are no exogeneity requirements for X. This is in contrast with the assumptions commonly made in the conventional structural literature and the semiparametric selection literature (see, e.g., Powell 1994).

A counterfactual "no feedback" condition facilitates interpretability so that conditioning on X does not mask the effects of D. Letting X_d denote a value of X if D is set to d, a sufficient condition that rules out feedback from D to X is:

(A-6) *Let* X_0 *denote the counterfactual value of* X *that would be observed if* D *is set to* 0. X_1 *is defined analogously. Assume* $X_d = X$ *for* $d = 0, 1$. *(The* X_D *are invariant to counterfactual manipulations.)*

Condition (A-6) is not strictly required to formulate an evaluation model, but it enables an analyst who conditions on X to capture the "total" or "full effect" of D on Y (see Pearl 2000). This assumption imposes the requirement that X is an external variable determined outside the model and is not affected by counterfactual manipulations of D. However, the assumption allows for X to be freely correlated with U_1, U_0 and V so it can be endogenous.

In this notation, $P(Z)$ is the probability of receiving treatment given Z, or the "propensity score" $P(Z) \equiv \Pr(D = 1 \mid Z) = F_{V|X}(\mu_D(Z))$, where $F_{V|X}(\cdot)$ denotes the distribution of V conditional on X. (Throughout this paper, I will refer to the cumulative distribution function of a random vector A by $F_A(\cdot)$ and to the cumulative distribution function of a random vector A conditional on random vector B by $F_{A|B}(\cdot)$. I will write the cumulative distribution function of A conditional on $B = b$ by $F_{A|B}(\cdot \mid b)$.) I denote $P(Z)$ by P, suppressing the Z argument. I also work with U_D, a uniform random variable ($U_D \sim \text{Unif}[0,1]$) defined by $U_D = F_{V|X}(V)$. (This representation is valid whether or not (A-1) is true. However, (A-1) imposes restrictions on counterfactual choices. For example, if a change in government policy changes the distribution of Z by an external manipulation, under (A-1) the model can be used to generate the choice probability from $P(z)$ evaluated at the new arguments, i.e., the model is invariant with respect to the distribution Z.) The separability between V and $\mu_D(Z)$ or $D(Z)$ and U_D is conventional. It plays a crucial role in justifying instrumental variable estimators in the general models analyzed in this paper.

Vytlacil (2002) establishes that assumptions (A-1)–(A-5) for the model of Eqs. (2)–(6) are equivalent to the assumptions used to generate the LATE model of Imbens and Angrist (1994). Thus the nonparametric selection model for treatment effects developed by Heckman

and Vytlacil is implied by the assumptions of the Imbens and Angrist instrumental variable model for treatment effects. The Heckman and Vytlacil approach is more general and links the IV literature to the literature on economic choice models. The latent variable model is a version of the standard sample selection bias model. This weaves together two strands of the literature often thought to be distinct (see e.g., Angrist and Krueger 1999). Heckman et al. (2006) develop this parallelism in detail. (The model of Eqs. (4)–(6) and assumptions (A-1)–(A-5) impose two testable restrictions on the distribution of (Y, D, Z, X). First, it imposes an index sufficiency restriction: for any set $\mathcal{A}$ and for $j = 0, 1$,

$$\Pr(Y_j \in \mathcal{A} \mid X, Z, D = j) = \Pr(Y_j \in \mathcal{A} \mid X, P(Z), D = j).$$

Z (given X) enters the model only through the propensity score $P(Z)$ (the sets of $\mathcal{A}$ are assumed to be measurable). This restriction has empirical content when Z contains two or more variables not in X. Second, the model also imposes monotonicity in p for $E(YD \mid X = x, P = p)$ and $E(Y(1-D) \mid X = x, P = p)$. Heckman and Vytlacil (2005, Appendix A) develop this condition further, and show that it is testable.

Even though this model of treatment effects is not the most general possible model, it has testable implications and hence empirical content. It unites various literatures and produces a nonparametric version of the selection model, and links the treatment literature to economic choice theory.)

Definitions of Treatment Effects in the Two Outcome Model

The difficulty of observing the same individual in both treated and untreated states leads to the use of various population level treatment effects widely used in the biostatistics literature and often applied in economics. (Heckman et al. (1999) discuss panel data cases where it is possible to observe both Y_0 and Y_1 for the same person.) The most commonly invoked treatment effect is the Average Treatment Effect (ATE): $\Delta^{\text{ATE}}(x) \equiv E(\Delta \mid X = x)$ where $\Delta = Y_1 - Y_0$. This is the effect of assigning treatment randomly to everyone of type X assuming full compliance, and ignoring general equilibrium effects. (See, e.g., Imbens (2004).) The average impact of treatment on persons who actually take the treatment is Treatment on the Treated (TT): $\Delta^{\text{TT}}(x) \equiv E(\Delta \mid X = x, D = 1)$. This parameter can also be defined conditional on $P(Z)$: $\Delta^{\text{TT}}(x, p) \equiv E(\Delta \mid X = x, P(Z) = p, D = 1)$. (These two definitions of treatment on the treated are related by integrating out the conditioning p variable: $\Delta^{\text{TT}}(x) = \int_0^1 \Delta^{\text{TT}}(x, p) dF_{P(Z)|X,D}(p|x, 1)$ where $F_{P(Z)|X,D}(\cdot|x, 1)$ is the distribution of $P(Z)$ given $X = x$ and $D = 1$.)

The mean effect of treatment on those for whom $X = x$ and $U_D = u_D$, the Marginal Treatment Effect (MTE), plays a fundamental role in the analysis of the next subsection:

$$\Delta^{\text{MTE}}(x, u_D) \equiv E(\Delta \mid X = x, U_D = u_D). \tag{7}$$

This parameter is defined independently of any instrument. I separate the definition of parameters from their identification. The MTE is the expected effect of treatment conditional on observed characteristics X and conditional on U_D, the unobservables from the first stage decision rule. For u_D evaluation points close to zero, $\Delta^{\text{MTE}}(x, u_D)$ is the expected effect of treatment on individuals with the value of unobservables that make them most likely to participate in treatment and who would participate even if the mean scale utility $\mu_D(Z)$ is small. If U_D is large, $\mu_D(Z)$ would have to be large to induce people to participate.

One can also interpret $E(\Delta \mid X = x, U_D = u_D)$ as the mean gain in terms of $Y_1 - Y_0$ for persons with observed characteristics X who would be indifferent between treatment or not if they were randomly assigned a value of Z, say z, such that $\mu_D(z) = u_D$. When Y_1 and Y_0 are value outcomes, MTE is a mean willingness-to-pay measure. MTE is a choice-theoretic building block that unites the treatment effect, selection, matching and control function literatures.

A third interpretation is that MTE conditions on X and the residual defined by subtracting the expectation of D^* from D^*: $\tilde{U}_D = D^* - E(D^* \mid Z, X)$. This is a "replacement function" interpretation in the sense of Heckman and Robb (1985a) and Matzkin (2007), or "control function" interpretation in the sense of Blundell and Powell (2003). (These three interpretations are equivalent under separability in D^*, i.e., when (6) characterizes the choice equation, but lead to three different definitions of MTE when a more general nonseparable model is developed. See Heckman and Vytlacil (2007b).) The additive separability of Eq. 6 in terms of observables and unobservables plays a crucial role in the justification of instrumental variable methods.

The LATE parameter of Imbens and Angrist (1994) is a version of MTE. I define LATE independently of any instrument after first presenting the Imbens–Angrist definition. Define $D(z)$ as a counterfactual choice variable, with $D(z) = 1$ if D would have been chosen if Z had been set to z, and $D(z) = 0$ otherwise. Let $\mathcal{Z}(x)$ denote the support of the distribution of Z conditional on $X = x$. For any $(z, z') \in \mathcal{Z}(x) \times \mathcal{Z}(x)$ such that $P(z) > P(z')$, LATE is $E(\Delta \mid X = x, D(z) = 1, D(z') = 0) = E(Y_1 - Y_0 \mid X = x, D(z) = 1, D(z') = 0)$, the mean gain to persons who would be induced to switch from $D = 0$ to $D = 1$ if Z were

manipulated externally from z' to z. In an example of the returns to education, z' could be the base level of tuition and z a reduced tuition level. Using the latent index model, Heckman and Vytlacil (1999, 2005) show that LATE can be written as

$$\begin{aligned} E(Y_1 - Y_0 \mid X = x, D(z) = 1, D(z') = 0) \\ = E\left(Y_1 - Y_0 \mid X = x, u'_D < U_D < u_D\right) \\ = \Delta^{\text{LATE}}\left(x, u_D, u'_D\right) \end{aligned}$$

for $u_D = \Pr(D(z) = 1) = P(z)$, $u'_D = \Pr(D(z') = 1) = P(z')$, where assumption (A-1) implies that $\Pr(D(z) = 1) = \Pr(D = 1 \mid Z = z)$ and $\Pr(D(z') = 1) = \Pr(D = 1 \mid Z = z')$.

Imbens and Angrist define the LATE parameter as the probability limit of an estimator. Their analysis conflates issues of definition of parameters with issues of identification. The representation of LATE given here allows analysts to separate these two conceptually distinct matters and to define the LATE parameter more generally. One can in principle evaluate the right hand side of the preceding equation at any u_D, u'_D points in the unit interval and not only at points in the support of the distribution of the propensity score $P(Z)$ conditional on $X = x$ where it is identified. From assumptions (A-1), (A-3), and (A-4), $\Delta^{\text{LATE}}\left(x, u_D, u'_D\right)$ is continuous in u_D and u'_D and $\lim_{u'_D \uparrow u_D} \Delta^{\text{LATE}}\left(x, u_D, u'_D\right) = \Delta^{\text{MTE}}(x, u_D)$. (This follows from Lebesgue's theorem for the derivative of an integral and holds almost everywhere with respect to Lebesgue measure. The ideas of the marginal treatment effect and the limit form of LATE were first introduced in the context of a parametric normal generalized Roy model by Björklund and Moffitt (1987), and were analyzed more generally in Heckman (1997). Angrist et al. (2000) also define and develop a limit form of LATE.)

Heckman and Vytlacil (1999) use assumptions (A-1)–(A-5) and the latent index structure to develop the relationship between MTE and the various treatment effect parameters shown in the first three lines of Table 1a. They present the formal derivation of the parameters and associated weights and graphically illustrates the relationship between ATE and TT. All treatment parameters may be expressed as weighted averages of the MTE:

$$\begin{aligned} &\text{Treatment Parameter } (j) \\ &= \int \Delta^{\text{MTE}}(x, u_D)\, \omega_j(x, u_D)\, du_D \end{aligned}$$

where $\omega_j(x, u_D)$ is the weighting function for the MTE and the integral is defined over the full support of u_D. Except for the OLS weights, the weights in the table all integrate to one, although in some cases the weights for IV may be negative (Heckman et al. 2006).

In Table 1a, $\Delta^{\text{TT}}(x)$ is shown as a weighted average of Δ^{MTE}:

$$\Delta^{\text{TT}}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TT}}(x, u_D)\, du_D,$$

where

$$\omega_{TT}(x, u_D) = \frac{1 - F_{P|X}(u_D \mid x)}{\int_0^1 \left(1 - F_{P|X}(t \mid x)\right) dt} = \frac{S_{P|X}(u_D \mid x)}{E(P(Z) \mid X = x)}, \tag{8}$$

and $S_{P|X}(u_D \mid x)$ is $\Pr(P(Z) > u_D \mid X = x)$ and $\omega_{\text{TT}}(x, u_D)$ is a weighted distribution. The parameter $\Delta^{\text{TT}}(x)$ oversamples $\Delta^{\text{MTE}}(x, u_D)$ for those individuals with low values of u_D that make them more likely to participate in the program being evaluated. Treatment on the untreated (TUT) is defined symmetrically with TT and oversamples those least likely to participate. The various weights are displayed in Table 1b. A central theme of the analysis of Heckman and Vytlacil is that under their assumptions all estimators and estimands can be written as weighted averages of MTE. This allows them to unify the treatment effect literature using a common functional MTE (u_D).

Observe that if $E(Y_1 - Y_0 \mid X = x, U_D = u_D) = E(Y_1 - Y_0 \mid X = x)$, so $\Delta = Y_1 - Y_0$ is mean independent of U_D given $X = x$, then $\Delta^{\text{MTE}} = \Delta^{\text{ATE}} = \Delta^{\text{TT}} = \Delta^{\text{LATE}}$. Therefore in cases where there is no heterogeneity in terms of unobservables in MTE (Δ constant conditional on $X = x$) or agents do not act on it so that U_D drops out of the conditioning set, marginal treatment effects are average treatment effects, so that all of the evaluation parameters are the same. Otherwise, they are different. Only in the case where the marginal treatment effect is the average treatment effect will the "effect" of treatment be uniquely defined.

Figure 2a plots weights for a parametric normal generalized Roy model generated from the parameters shown at the base of Fig. 2b. The model allows for costs to vary in the population and is more general than the extended Roy model used to construct Fig. 1. The weights for IV depicted in Fig. 2b are discussed in Heckman et al. (2006) and the weights for OLS are discussed in the next section. A high u_D is associated with higher cost, relative to return, and less likelihood of choosing $D = 1$. The decline of MTE in terms of higher values of u_D means that people with higher u_D have lower gross returns. TT overweights low values of u_D (i.e., it oversamples U_D that make it likely to have $D = 1$). ATE samples U_D uniformly. Treatment on the Untreated

Principles Underlying Econometric Estimators for Identifying Causal Effects. Table 1

(a) Treatment effects and estimands as weighted averages of the marginal treatment effect

$\text{ATE}(x) = E(Y_1 - Y_0 \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, du_D$

$\text{TT}(x) = E(Y_1 - Y_0 \mid X = x, D = 1) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TT}}(x, u_D)\, du_D$

$\text{TUT}(x) = E(Y_1 - Y_0 \mid X = x, D = 0) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{TUT}}(x, u_D)\, du_D$

Policy relevant treatment effect $(x) = E(Y_{a'} \mid X = x) - E(Y_a \mid X = x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{PRTE}}(x, u_D)\, du_D$ for two policies a and a' that affect the Z but not the X

$\text{IV}_J(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{IV}}^J(x, u_D)\, du_D$, given instrument J

$\text{OLS}(x) = \int_0^1 \Delta^{\text{MTE}}(x, u_D)\, \omega_{\text{OLS}}(x, u_D)\, du_D$

(b) Weights (Heckman and Vytlacil 2005)

$$\omega_{\text{ATE}}(x, u_D) = 1$$

$$\omega_{\text{TT}}(x, u_D) = \left[\int_{u_D}^1 f(p \mid X = x) dp\right] \frac{1}{E(P \mid X = x)}$$

$$\omega_{\text{TUT}}(x, u_D) = \left[\int_0^{u_D} f(p \mid X = x)\, dp\right] \frac{1}{E((1 - P) \mid X = x)}$$

$$\omega_{\text{PRTE}}(x, u_D) = \left[\frac{F_{P_{a'},X}(u_D) - F_{P_a,X}(u_D)}{\Delta \bar{P}}\right]$$

$$\omega_{\text{IV}}^J(x, u_D) = \left[\int_{u_D}^1 (J(Z) - E(J(Z) \mid X = x)) \int f_{J,P|X}(j, t \mid X = x)\, dt\, dj\right] \frac{1}{\text{Cov}(J(Z), D \mid X = x)}$$

$$\omega_{\text{OLS}}(x, u_D) = 1 + \frac{E(U_1 \mid X = x, U_D = u_D)\, \omega_1(x, u_D) - E(U_0 \mid X = x, U_D = u_D)\, \omega_0(x, u_D)}{\Delta^{\text{MTE}}(x, u_D)}$$

$$\omega_1(x, u_D) = \left[\int_{u_D}^1 f(p \mid X = x)\, dp\right] \left[\frac{1}{E(P \mid X = x)}\right]$$

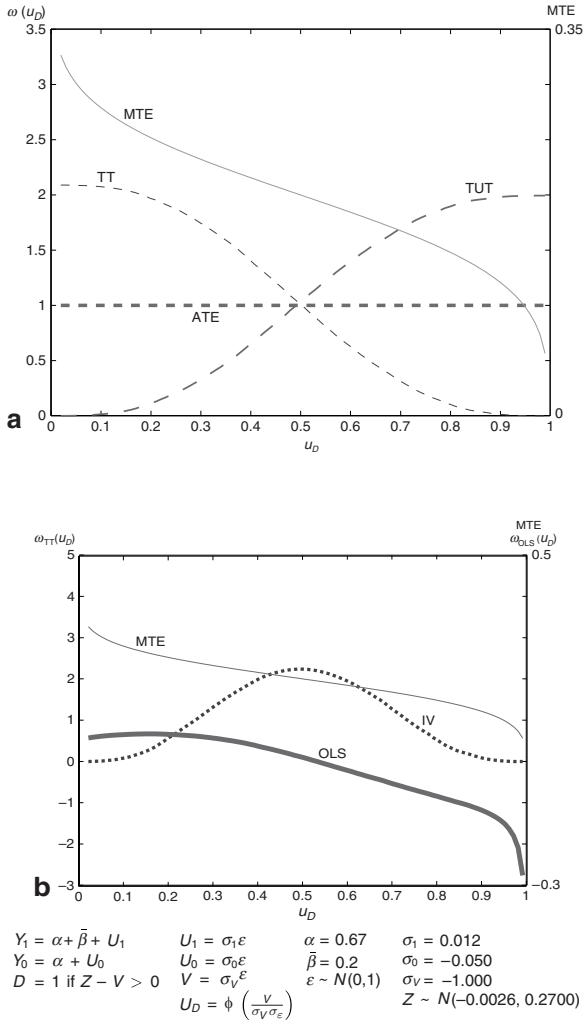
$$\omega_0(x, u_D) = \left[\int_0^{u_D} f(p \mid X = x)\, dp\right] \frac{1}{E((1 - P) \mid X = x)}$$

$(E(Y_1 - Y_0 \mid X = x, D = 0))$, or TUT, oversamples the values of U_D which make it unlikely to have $D = 1$.

Table 2 shows the treatment parameters produced from the different weighting schemes for the model used to generate the weights in Fig. 2a and 2b. Given the decline of the MTE in u_D, it is not surprising that TT>ATE>TUT. This is the generalized Roy version of the principle of diminishing returns. Those most likely to self select into the program benefit the most from it. The difference between TT and ATE is a sorting gain: $E(Y_1 - Y_0 \mid X, D = 1) - E(Y_1 - Y_0 \mid X)$, the average gain experienced by people who sort into treatment compared to what the average person would experience. Purposive selection on the basis of gains should lead to positive sorting gains of the kind found in the table. If there is negative sorting on the gains, then TUT ≥ ATE ≥ TT.

The Weights for a Generalized Roy Model

Heckman et al. (2006) show that all of the weights for treatment effects and IV estimators can be estimated over the available support. Since the MTE can be estimated by the method of Local Instrumental variables, we can form each treatment effect and each IV estimand as an integral to two estimable functions (subject to support). For the case of continuous Z, I plot the weights associated with the MTE for IV. This analysis draws on Heckman et al. (2006), who derive the weights. Figure 3 plots $E(Y \mid P(Z))$ and MTE for the extended Roy models generated by the parameters displayed at the base of the figure. In cases where $\beta \perp\!\!\!\perp D$, $\Delta^{\text{MTE}}(u_D)$ is constant in u_D. This is trivial when β is a constant. When β is random but selection into D does not depend on β, MTE is still flat. The more interesting case termed "essential heterogeneity" by



Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 2(a) Weights for the marginal treatment effect for different parameters (Heckman and Vytlacil 2005) (**b**) Marginal treatment effect vs linear instrumental variables and ordinary least squares weights (Heckman and Vytlacil 2005)

Principles Underlying Econometric Estimators for Identifying Causal Effects. Table 2 Treatment parameters and estimands in the generalized Roy example

Treatment on the treated	0.2353
Treatment on the untreated	0.1574
Average treatment effect	0.2000
Sorting gain[a]	0.0353
Policy relevant treatment effect (PRTE)	0.1549
Selection bias[b]	−0.0628
Linear instrumental variables[c]	0.2013
Ordinary least squares	0.1725

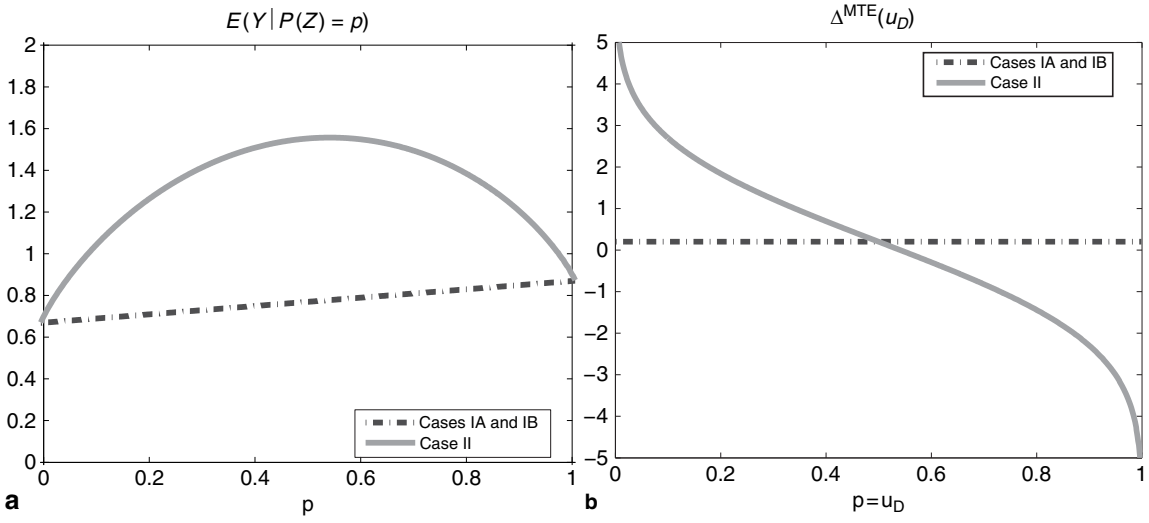
[a] $TT - ATE = E(Y_1 - Y_0 \mid D = 1) - E(Y_1 - Y_0)$
[b] $OLS - TT = E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)$
[c] Using Propensity Score $P(Z)$ as the instrument.
Note: The model used to create Table 2 is the same as those used to create Fig. 2a and b. The PRTE is computed using a policy t characterized as follows:
If $Z > 0$ then $D = 1$ if $Z(1 + t) - V > 0$.
If $Z \leq t$ then $D = 1$ if $Z - V > 0$.
For this example t is set equal to 0.2.

Heckman and Vytlacil has $\beta \not\perp\!\!\!\perp D$. The left hand side (Fig. 3a) depicts $E(Y \mid P(Z))$ in the two cases. The first case makes $E(Y \mid P(Z))$ linear in $P(Z)$. The second case is nonlinear in $P(Z)$. This arises when $\beta \not\perp\!\!\!\perp D$. The derivative of $E(Y \mid P(Z))$ is presented in the right panel (Fig. 3b). It is a constant for the first case (flat MTE) but declining in $U_D = P(Z)$ for the case with selection on the gain. A simple test for linearity in $P(Z)$ in the outcome equation reveals whether or not the analyst is in cases I and II ($\beta \perp\!\!\!\perp D$) or case III ($\beta \not\perp\!\!\!\perp D$). (Recall that we keep the conditioning on X implicit.) These cases are the extended Roy counterparts to $E(Y \mid P(Z) = p)$ and MTE shown for the generalized Roy model in Figs. 4a and 4b.

MTE gives the mean marginal return for persons who have utility $P(Z) = u_D$. Thus, $P(Z) = u_D$ is the margin of indifference. Those with low u_D values have high returns. Those with high u_D values have low returns. Figure 3 highlights that in the general case MTE (and LATE) identify average returns for persons at the margin of indifference at different levels of the mean utility function ($P(Z)$).

Figure 5 plots MTE and LATE for different intervals of u_D using the model generating Fig. 3. LATE is the chord of $E(Y \mid P(Z))$ evaluated at different points. The relationship between LATE and MTE is depicted in the right panel (b) of Fig. 5. LATE is the integral under the MTE curve divided by the difference between the upper and lower limits.

Treatment parameters associated with the second case are plotted in Fig. 6. The MTE is the same as that presented in Fig. 3. ATE has the same value for all p. The effect of treatment on the treated for $P(Z) = p$, $\Delta^{TT}(p) = E(Y_1 - Y_0 \mid D = 1, P(Z) = p)$ declines in p (equivalently it declines in u_D). Treatment on the untreated given p, $\text{TUT}(p) = \Delta^{TUT}(p) = E(Y_1 - Y_0 \mid D = 0, P(Z) = p)$ also declines in p.



Outcomes

$$Y_1 = \alpha + \bar{\beta} + U_1$$
$$Y_0 = \alpha + U_0$$

Choice model

$$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$$

Case IA	Case IB	Case II
$U_1 = U_0$	$U_1 - U_0 \perp\!\!\!\perp D$	$U_1 - U_0 \not\perp\!\!\!\perp D$
$\bar{\beta}$ =ATE=TT=TUT=IV	$\bar{\beta}$ =ATE=TT=TUT=IV	$\bar{\beta}$ =ATE≠TT≠TUT≠IV

Parameterization

Cases IA, IB, and II	Cases IB and II	Case II
$\alpha = 0.67$	$(U_1, U_0) \sim N(\mathbf{0}, \Sigma)$	$D^* = Y_1 - Y_0 - \gamma Z$
$\bar{\beta} = 0.2$	with $\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$	$Z \sim N(\mu_Z, \Sigma_Z)$
		$\mu_Z = (2, -2)$ and $\Sigma_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$
		$\gamma = (0.5, 0.5)$

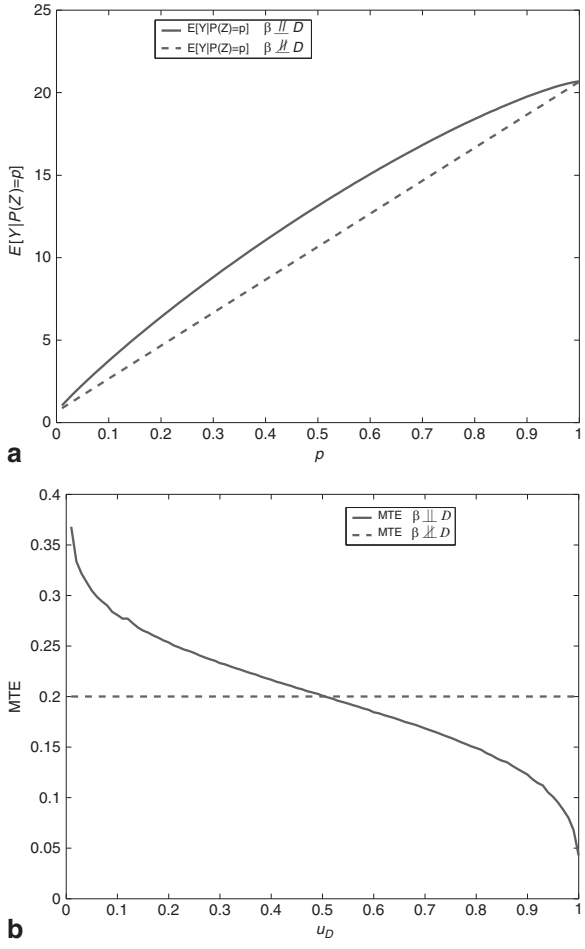
Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 3 Conditional expectation of Y on $P(Z)$ and the marginal treatment effect (MTE) the extended Roy economy (Heckman et al. 2006)

$$LATE(p, p') = \frac{\Delta^{TT}(p')p' - \Delta^{TT}(p)p}{p' - p}, \qquad p' \neq p$$

$$MTE = \frac{\partial[\Delta^{TT}(p)p]}{\partial p}.$$

One can generate all of the treatment parameters from $\Delta^{TT}(p)$.

Matching on $P = p$ (which is equivalent to nonparametric regression given $P = p$) produces a biased estimator of $TT(p)$. Matching assumes a flat MTE (average return



Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 4 (**a**) Plot of the $E(Y|P(Z) = p)$, (**b**) Plot of the identified marginal treatment effect from Fig. 2a (the derivative). Note: Parameters for the general heterogeneous case are the same as those used in Fig. 2a and 2b. For the homogeneous case we impose $U_1 = U_0$ ($\sigma_1 = \sigma_0 = 0.012$). (Heckman and Vytlacil 2005)

equals marginal return) as we develop below. Therefore it is systematically biased for $\Delta^{TT}(p)$ in a model with essential heterogeneity, where $\beta \not\perp\!\!\!\perp D$. Making observables alike makes the unobservables dissimilar. Holding p constant across treatment and control groups understates $TT(p)$ for low values of p and overstates it for high values of p. I develop this point further in section "►Matching", where I discuss the method of matching. First I present a unified approach that integrates all evaluation estimators in a common framework.

The Basic Principles Underlying the Identification of the Leading Econometric Evaluation Estimators

This section reviews the main principles underlying the evaluation estimators commonly used in the econometric literature. I assume two potential outcomes (Y_0, Y_1). $D = 1$ if Y_1 is observed, and $D = 0$ corresponds to Y_0 being observed. The observed outcome is

$$Y = DY_1 + (1 - D)Y_0. \tag{9}$$

The *evaluation problem* arises because for each person we observe either Y_0 or Y_1 but not both. Thus in general it is not possible to identify the individual level treatment effect $Y_1 - Y_0$ for any person. The typical solution to this problem is to reformulate the problem at the population level rather than at the individual level and to identify certain mean outcomes or quantile outcomes or various distributions of outcomes as described in Heckman and Vytlacil (2007a). For example, a commonly used approach focuses attention on average treatment effects, such as ATE = $E(Y_1 - Y_0)$.

If treatment is assigned or chosen on the basis of potential outcomes, so

$$(Y_0, Y_1) \not\perp\!\!\!\perp D,$$

where $\not\perp\!\!\!\perp$ denotes "is not independent" and "$\perp\!\!\!\perp$" denotes independent, we encounter the problem of selection bias. Suppose that we observe people in each treatment state $D = 0$ and $D = 1$. If $Y_j \not\perp\!\!\!\perp D$, then the observed Y_j will be selectively different from randomly assigned Y_j, $j = 0, 1$. Thus $E(Y_0 \mid D = 0) \neq E(Y_0)$ and $E(Y_1 \mid D = 1) \neq E(Y_1)$. Using unadjusted data to construct $E(Y_1 - Y_0)$ will produce one source of evaluation bias:

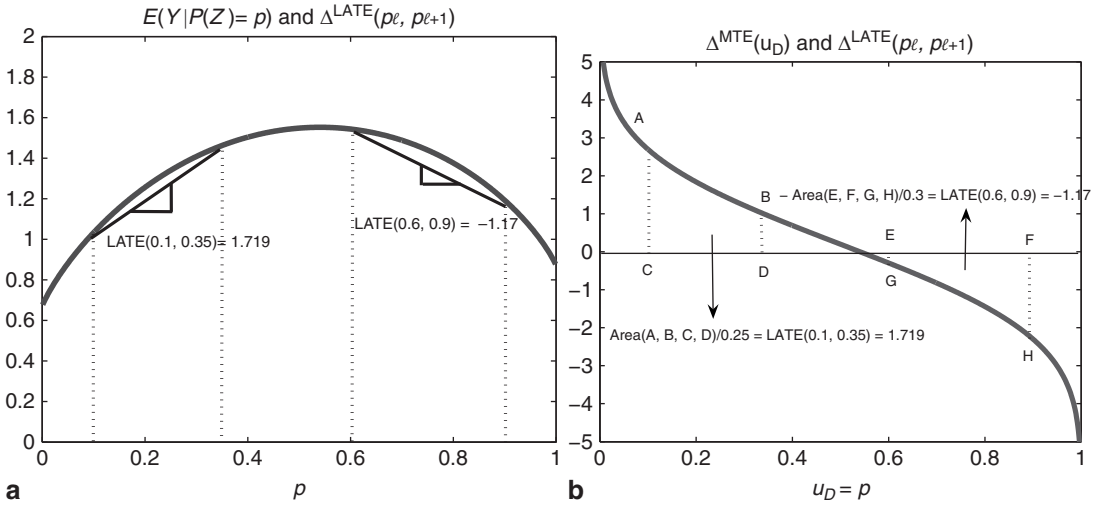
$$E(Y_1 \mid D = 1) - E(Y_0 \mid D = 0) \neq E(Y_1 - Y_0).$$

The selection problem underlies the evaluation problem. Many methods have been proposed to solve both problems.

The method with the greatest intuitive appeal, which is sometimes called the "gold standard" in evaluation analysis, is the method of random assignment. Nonexperimental methods can be organized by how they attempt to approximate what can be obtained by an ideal random assignment. If treatment is chosen at random with respect to (Y_0, Y_1), or if treatments are randomly assigned and there is full compliance with the treatment assignment,

(R-1) $(Y_0, Y_1) \perp\!\!\!\perp D$.

It is useful to distinguish several cases where (R-1) will be satisfied. The first is that agents (decision makers whose choices are being analyzed) pick outcomes that are random with respect to (Y_0, Y_1). Thus agents may not know



$$\Delta^{\text{LATE}}(p_\ell, p_{\ell+1}) = \frac{E(Y|P(Z) = p_{\ell+1}) - E(Y|P(Z) = p_\ell)}{p_{\ell+1} - p_\ell} = \frac{\int_{p_\ell}^{p_{\ell+1}} \Delta^{\text{MTE}}(u_D)du_D}{p_{\ell+1} - p_\ell}$$

$$\Delta^{\text{LATE}}(0.6, 0.9) = -1.17$$
$$\Delta^{\text{LATE}}(0.1, 0.35) = 1.719$$

Outcomes	Choice model
$Y_1 = \alpha + \bar{\beta} + U_1$	$D = \begin{cases} 1 \text{ if } D^* > 0 \\ 0 \text{ if } D^* \leq 0 \end{cases}$
$Y_0 = \alpha + U_0$	with $D^* = Y_1 - Y_0 - \gamma Z$

Parameterization

$$(U_1, U_0) \sim N(\mathbf{0}, \Sigma) \text{ and } Z \sim N(\mu_Z, \Sigma_Z)$$
$$\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \ \mu_Z = (2, -2) \text{ and } \Sigma_Z = \begin{bmatrix} 9 & -2 \\ -2 & 9 \end{bmatrix}$$
$$\alpha = 0.67, \bar{\beta} = 0.2, \gamma = (0.5, 0.5)$$

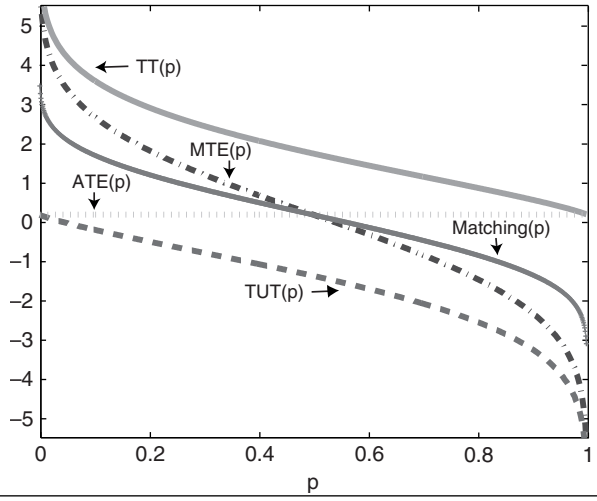
Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 5 The local average treatment effect the extended Roy economy (Heckman et al. 2006)

(Y_0, Y_1) at the time they make their choices to participate in treatment or at least do not act on (Y_0, Y_1), so that $\Pr(D = 1 \mid X, Y_0, Y_1) = \Pr(D = 1 \mid X)$ for all X. Matching assumes a version of (R-1) conditional on matching variables X: $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$.

A second case arises when individuals are randomly assigned to treatment status even if they would choose to self select into no-treatment status, and they comply with the randomization protocols. Let ξ be randomized assignment status. With full compliance, $\xi = 1$ implies that Y_1 is observed and $\xi = 0$ implies that Y_0 is observed. Then, under randomized assignment,

(R-2) $(Y_0, Y_1) \perp\!\!\!\perp \xi$,

even if in a regime of self-selection, $(Y_0, Y_1) \not\perp\!\!\!\perp D$. If ►randomization is performed conditional on X, we obtain $(Y_0, Y_1) \perp\!\!\!\perp \xi \mid X$.



Parameter	Definition	Under assumptions[a]
Marginal treatment effect	$E\left[Y_1 - Y_0 \mid D^* = 0, P(Z) = p\right]$	$\bar{\beta} + \sigma_{U_1 - U_0}\Phi^{-1}(1-p)$
Average treatment effect	$E\left[Y_1 - Y_0 \mid P(Z) = p\right]$	$\bar{\beta}$
Treatment on the treated	$E\left[Y_1 - Y_0 \mid D^* > 0, P(Z) = p\right]$	$\bar{\beta} + \sigma_{U_1 - U_0}\frac{\phi\left(\Phi^{-1}(1-p)\right)}{p}$
Treatment on the untreated	$E\left[Y_1 - Y_0 \mid D^* \leq 0, P(Z) = p\right]$	$\bar{\beta} - \sigma_{U_1 - U_0}\frac{\phi\left(\Phi^{-1}(1-p)\right)}{1-p}$
OLS/matching on $P(Z)$	$E\left[Y_1 \mid D^* > 0, P(Z) = p\right]$	$\bar{\beta} + \left(\frac{\sigma^2_{U_1} - \sigma_{U_1 U_0}}{\sqrt{\sigma_{U_1 - U_0}}}\right)\left(\frac{1-2p}{p(1-p)}\right)\phi\left(\Phi^{-1}(1-p)\right)$
	$-E\left[Y_0 \mid D^* \leq 0, P(Z) = p\right]$	

Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 6 Treatment parameters and OLS/matching as a function of $P(Z) = p$. Note: $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cdf and pdf of a standard normal distribution, respectively. $\Phi^{-1}(\cdot)$ represents the inverse of $\Phi(\cdot)$. [a] The model in this case is the same as the one presented below Fig. 5. (Heckman et al. 2006)

Let A denote actual treatment status. If the randomization has full compliance among participants, $\xi = 1 \Rightarrow A = 1$; $\xi = 0 \Rightarrow A = 0$. This is entirely consistent with a regime in which a person would choose $D = 1$ in the absence of randomization, but would have no treatment ($A = 0$) if suitably randomized, even though the agent might desire treatment.

If treatment status is chosen by self-selection, $D = 1 \Rightarrow A = 1$ and $D = 0 \Rightarrow A = 0$. If there is imperfect compliance with randomization, $\xi = 1 \nRightarrow A = 1$ because of agent choices. In general, $A = \xi D$ so that $A = 1$ only if $\xi = 1$ and $D = 1$. If treatment status is randomly assigned, either through randomization or randomized self-selection,

(R-3) $(Y_0, Y_1) \perp\!\!\!\perp A$.

This version of randomization can also be defined conditional on X. Under (R-1), (R-2), or (R-3), the average treatment effect (ATE) is the same as the marginal treatment effect of Björklund and Moffitt (1987) and Heckman and Vytlacil (1999, 2005, 2007a), and the parameters treatment on the treated (TT) ($E(Y_1 - Y_0 \mid D = 1)$) and treatment on the untreated (TUT) ($E(Y_1 - Y_0 \mid D = 0)$). (The marginal treatment effect is formally defined in the next section.) These parameters can be identified from population means:

$$\text{TT} = \text{MTE} = \text{TUT} = \text{ATE} = E(Y_1 - Y_0) = E(Y_1) - E(Y_0).$$

Forming averages over populations of persons who are treated ($A = 1$) or untreated ($A = 0$) suffices to identify this parameter. If there are conditioning variables X, we can define the mean treatment parameters for all X where (R-1) or (R-2) or (R-3) hold.

Observe that even with random assignment of treatment status and full compliance, one cannot, in general, identify the distribution of the treatment effects

$(Y_1 - Y_0)$, although one can identify the marginal distributions $F_1(Y_1 \mid A = 1, X = x) = F_1(Y_1 \mid X = x)$ and $F_0(Y_0 \mid A = 0, X = x) = F_0(Y_0 \mid X = x)$. One special assumption, common in the conventional econometrics literature, is that $Y_1 - Y_0 = \Delta(x)$, a constant given x. Since $\Delta(x)$ can be identified from $E(Y_1 \mid A = 1, X = x) - E(Y_0 \mid A = 0, X = x)$ because A is randomly allocated, in this special case the analyst can identify the joint distribution of (Y_0, Y_1). (Heckman (1992); Heckman et al. (1997).) This approach assumes that (Y_0, Y_1) have the same distribution up to a parameter Δ (Y_0 and Y_1 are perfectly dependent). One can make other assumptions about the dependence across ranks from perfect positive or negative ranking to independence. (Heckman et al. (1997).) The joint distribution of (Y_0, Y_1) or of $(Y_1 - Y_0)$ is not identified unless the analyst can pin down the dependence across (Y_0, Y_1). Thus, even with data from a randomized trial one cannot, without further assumptions, identify the proportion of people who benefit from treatment in the sense of gross gain ($\Pr(Y_1 \geq Y_0)$). This problem plagues all evaluation methods. Abbring and Heckman (2007) discuss methods for identifying joint distributions of outcomes. (See also Aakvik et al. (2005); Carneiro et al. (2001, 2003); and Cunha et al. (2005).)

Assumption (R-1) is very strong. In many cases, it is thought that there is *selection bias* with respect to Y_0, Y_1, so persons who select into status 1 or 0 are selectively different from randomly sampled persons in the population. The assumption most commonly made to circumvent problems with (R-1) is that even though D is not random with respect to potential outcomes, the analyst has access to control variables X that effectively produce a randomization of D with respect to (Y_0, Y_1) given X. This is the method of matching, which is based on the following conditional independence assumption:

(M-1) $\quad (Y_0, Y_1) \perp\!\!\!\perp D \mid X.$

Conditioning on X randomizes D with respect to (Y_0, Y_1). (M-1) assumes that any selective sampling of (Y_0, Y_1) can be adjusted by conditioning on observed variables. (R-1) and (M-1) are different assumptions and neither implies the other. In a linear equations model, assumption (M-1) that D is independent from (Y_0, Y_1) given X justifies application of ▶least squares on D to eliminate selection bias in mean outcome parameters. For means, matching is just nonparametric regression. (Barnow et al. (1980) present one application of matching in a regression setting.) In order to be able to compare X-comparable people in the treatment regime one must assume

(M-2) $\quad 0 < \Pr(D = 1 \mid X = x) < 1.$

Assumptions (M-1) and (M-2) justify matching. Assumption (M-2) is required for *any* evaluation estimator that compares treated and untreated persons. It is produced by random assignment if the randomization is conducted for all $X = x$ and there is full compliance.

Observe that from (M-1) and (M-2), it is possible to identify $F_1(Y_1 \mid X = x)$ from the observed data $F_1(Y_1 \mid D = 1, X = x)$ since we observe the left hand side of

$$F_1(Y_1 \mid D = 1, X = x) = F_1(Y_1 \mid X = x) = F_1(Y_1 \mid D = 0, X = x).$$

The first equality is a consequence of conditional independence assumption (M-1). The second equality comes from (M-1) and (M-2). By a similar argument, we observe the left hand side of

$$F_0(Y_0 \mid D = 0, X = x) = F_0(Y_0 \mid X = x) = F_0(Y_0 \mid D = 1, X = x),$$

and the equalities are a consequence of (M-1) and (M-2). Since the pair of outcomes (Y_0, Y_1) is not identified for anyone, as in the case of data from randomized trials, the joint distributions of (Y_0, Y_1) given X or of $Y_1 - Y_0$ given X are not identified without further information. This is a problem that plagues all selection estimators.

From the data on Y_1 given X and $D = 1$ and the data on Y_0 given X and $D = 0$, since $E(Y_1 \mid D = 1, X = x) = E(Y_1 \mid X = x) = E(Y_1 \mid D = 0, X = x)$ and $E(Y_0 \mid D = 0, X = x) = E(Y_0 \mid X = x) = E(Y_0 \mid D = 1, X = x)$ we obtain

$$E(Y_1 - Y_0 \mid X = x) = E(Y_1 - Y_0 \mid D = 1, X = x) = E(Y_1 - Y_0 \mid D = 0, X = x).$$

Effectively, we have a randomization for the subset of the support of X satisfying (M-2).

At values of X that fail to satisfy (M-2), there is no variation in D given X. One can define the residual variation in D not accounted for by X as

$$\mathcal{E}(x) = D - E(D \mid X = x) = D - \Pr(D = 1 \mid X = x).$$

If the variance of $\mathcal{E}(x)$ is zero, it is not possible to construct contrasts in outcomes by treatment status for those X values and (M-2) is violated. To see the consequences of this violation in a regression setting, use $Y = Y_0 + D(Y_1 - Y_0)$ and take conditional expectations, under (M-1), to obtain

$$E(Y \mid X, D) = E(Y_0 \mid X) + D[E(Y_1 - Y_0 \mid X)].$$

This follows because $E(Y \mid X, D) = E(Y_0 \mid X, D) + DE(Y_1 - Y_0 \mid X, D)$ but from (M-1), $E(Y_0 \mid X, D) = E(Y_0 \mid X)$ and $E(Y_1 - Y_0 \mid X, D) = E(Y_1 - Y_0 \mid X)$. If $Var(\mathcal{E}(x)) > 0$

for all x in the support of X, one can use nonparametric least squares to identify $E(Y_1 - Y_0 \mid X = x)$ =ATE(x) by regressing Y on D and X. The function identified from the coefficient on D is the average treatment effect. (Under the conditional independence assumption (M-1), it is also the effect of treatment on the treated $E(Y_1 - Y_0 \mid X, D = 1)$ and the marginal treatment effect formally defined in the next section.) If $Var(\mathcal{E}(x)) = 0$, ATE(x) is not identified at that x value because there is no variation in D that is not fully explained by X. A special case of matching is linear least squares where one can write

$$Y_0 = X\alpha + U \qquad Y_1 = X\alpha + \beta + U,$$

$U_0 = U_1 = U$ and hence under (M-1),

$$E(Y \mid X, D) = X\alpha + \beta D.$$

If D is perfectly predictable by X, one cannot identify β because of a multicollinearity problem (see ▶Multicollinearity). (M-2) rules out perfect collinearity. (Clearly (M-1) and (M-2) are sufficient but not necessary conditions. For the special case of OLS, as a consequence of the assumed linearity in the functional form of the estimating equation, we achieve identification of β if $Cov(X, U) = 0$, $Cov(D, U) = 0$ and (D, X) are not perfectly collinear. These conditions are much weaker than (M-1) and (M-2) and can be satisfied if (M-1) and (M-2) are only identified in a subset of the support of X.) Matching is a nonparametric version of least squares that does not impose functional form assumptions on outcome equations, and that imposes support condition (M-2).

Conventional econometric choice models make a distinction between variables that appear in outcome equations (X) and variables that appear in choice equations (Z). The same variables may be in (X) and (Z) but more typically, there are some variables not in common. For example, the instrumental variable estimator is based on variables that are not in X but that are in Z. Matching makes no distinction between the X and the Z. (Heckman et al. (1998) distinguish X and Z in matching. They consider a case where conditioning on X may lead to failure of (M-1) and (M-2) but conditioning on (X, Z) satisfies a suitably modified version of this condition.) It does not rely on exclusion restrictions. The conditioning variables used to achieve conditional independence can in principle be a set of variables Q distinct from the X variables (covariates for outcomes) or the Z variables (covariates for choices). I use X solely to simplify the notation. The key identifying assumption is the assumed existence of a random variable X with the properties satisfying (M-1) and (M-2).

Conditioning on a larger vector (X augmented with additional variables) or a smaller vector (X with some components removed) may or may not produce suitably modified versions of (M-1) and (M-2). Without invoking further assumptions there is no objective principle for determining what conditioning variables produce (M-1).

Assumption (M-1) is strong. Many economists do not have enough faith in their data to invoke it. Assumption (M-2) is testable and requires no act of faith. To justify (M-1), it is necessary to appeal to the quality of the data.

Using economic theory can help guide the choice of an evaluation estimator. A crucial distinction is the one between the information available to the analyst and the information available to the agent whose outcomes are being studied. Assumptions made about these information sets drive the properties of econometric estimators. Analysts using matching make strong informational assumptions in terms of the data available to them. In fact, all econometric estimators make assumptions about the presence or absence of informational asymmetries, and I exposit them in this paper.

To analyze the informational assumptions invoked in matching, and other econometric evaluation strategies, it is helpful to introduce five distinct information sets and establish some relationships among them. (See also the discussion in Barros (1987), Heckman and Navarro (2004), and Gerfin and Lechner (2002).) (1) An information set $\sigma(I_R)$ with an associated random variable that satisfies conditional independence (M-1) is defined as a *relevant* information set; (2) The minimal information set $\sigma(I_R)$ with associated random variable needed to satisfy conditional independence (M-1), the *minimal relevant* information set; (3) The information set $\sigma(I_A)$ available to the agent at the time decisions to participate are made; (4) The information available to the economist, $\sigma(I_{E^*})$; and (5) The information $\sigma(I_E)$ used by the economist in conducting an empirical analysis. I will denote the random variables generated by these sets as $I_{R^*}, I_R, I_A, I_{E^*}, I_E$, respectively. (I start with a primitive probability space (Ω, σ, P) with associated random variables I. I assume minimal σ-algebras and assume that the random variables I are measurable with respect to these σ-algebras. Obviously, strictly monotonic or affine transformations of the I preserve the information and can substitute for the I.)

Definition 1 *Define $\sigma(I_{R^*})$ as a relevant information set if the information set is generated by the random variable I_{R^*}, possibly vector valued, and satisfies condition (M-1), so*

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid I_{R^*}.$$

Definition 2 *Define $\sigma(I_R)$ as a minimal relevant information set if it is the intersection of all sets $\sigma(I_{R^*})$ and satisfies $(Y_0, Y_1) \perp\!\!\!\perp D \mid I_R$. The associated random variable I_R is a minimum amount of information that guarantees that condition (M-1) is satisfied. There may be no such set.* (Observe that the intersection of all sets $\sigma(I_{R^*})$ may be empty and hence may not be characterized by a (possibly vector valued) random variable I_R that guarantees $(Y_1, Y_2) \perp\!\!\!\perp D \mid I_R$. If the information sets that produce conditional independence are nested, then the intersection of all sets $\sigma(I_{R^*})$ producing conditional independence is well defined and has an associated random variable I_R with the required property, although it may not be unique (e.g., strictly monotonic transformations and affine transformations of I_R also preserve the property). In the more general case of non-nested information sets with the required property, it is possible that no uniquely defined minimal relevant set exists. Among collections of nested sets that possess the required property, there is a minimal set defined by intersection but there may be multiple minimal sets corresponding to each collection.)

If one defines the relevant information set as one that produces conditional independence, it may not be unique. If the set $\sigma(I_{R^*})$ satisfies the conditional independence condition, then the set $\sigma(I_{R^*}, Q)$ such that $Q \perp\!\!\!\perp (Y_0, Y_1) \mid I_{R^*}$ would also guarantee conditional independence. For this reason, I define the relevant information set to be minimal, that is, to be the intersection of all relevant sets that still produce conditional independence between (Y_0, Y_1) and D. However, no minimal set may exist.

Definition 3 *The agent's information set, $\sigma(I_A)$, is defined by the information I_A used by the agent when choosing among treatments. Accordingly, call I_A the agent's information.*

By the agent I mean the person making the treatment decision not necessarily the person whose outcomes are being studied (e.g., the agent may be the parent; the person being studied may be a child).

Definition 4 *The econometrician's full information set, $\sigma(I_{E^*})$, is defined as all of the information available to the econometrician, I_{E^*}.*

Definition 5 *The econometrician's information set, $\sigma(I_E)$, is defined by the information used by the econometrician when analyzing the agent's choice of treatment, I_E, in conducting an analysis.*

For the case where a unique minimal relevant information set exists, only three restrictions are implied by the structure of these sets: $\sigma(I_R) \subseteq \sigma(I_{R^*})$, $\sigma(I_R) \subseteq \sigma(I_A)$, and $\sigma(I_E) \subseteq \sigma(I_{E^*})$. (This formulation assumes that the agent makes the treatment decision. The extension to the case where the decision maker and the agent are distinct is straightforward. The requirement $\sigma(I_R) \subseteq \sigma(I_{R^*})$ is satisfied by nested sets.) I have already discussed the first restriction. The second restriction requires that the minimal relevant information set must be part of the information the agent uses when deciding which treatment to take or assign. It is the information in $\sigma(I_A)$ that gives rise to the selection problem which in turn gives rise to the evaluation problem.

The third restriction requires that the information used by the econometrician must be part of the information that the agent observes. Aside from these orderings, the econometrician's information set may be different from the agent's or the relevant information set. The econometrician may know something the agent doesn't know, for typically he is observing events after the decision is made. At the same time, there may be private information known to the agent but not the econometrician. Matching assumption (M-1) implies that $\sigma(I_R) \subseteq \sigma(I_E)$, so that the econometrician uses at least the minimal relevant information set, but of course he or she may use more. However, using more information is not guaranteed to produce a model with conditional independence property (M-1) satisfied for the augmented model. Thus an analyst can "overdo" it. Heckman and Navarro (2004) and Abbring and Heckman (2007) present examples of the consequences of the asymmetry in agent and analyst information sets.

The possibility of asymmetry in information between the agent making participation decisions and the observing economist creates the potential for a major identification problem that is ruled out by assumption (M-1). The methods of control functions and instrumental variables estimators (and closely related regression discontinuity design methods) address this problem but in different ways. Accounting for this possibility is a more conservative approach to the selection problem than the one taken by advocates of ▶least squares, or its nonparametric counterpart, matching. Those advocates assume that they know the X that produces a relevant information set. Heckman and Navarro (2004) show the biases that can result in matching when standard econometric model selection criteria are applied to pick the X that are used to satisfy (M-1). Conditional independence condition (M-1) cannot be tested without maintaining other assumptions. (These assumptions may or may not be testable. The required "exogeneity" conditions are discussed in Heckman and Navarro (2004). Thus randomization of assignment of treatment status might be used to test (M-1) but this requires that there be full compliance and that the

randomization be valid (no anticipation effects or general equilibrium effects).) Choice of the appropriate conditioning variables is a problem that plagues *all* econometric estimators.

The methods of control functions, replacement functions, proxy variables, and instrumental variables all recognize the possibility of asymmetry in information between the agent being studied and the econometrician and recognize that even after conditioning on X (variables in the outcome equation) and Z (variables affecting treatment choices, which may include the X), analysts may fail to satisfy conditional independence condition (M-1). (The term and concept of control function is due to Heckman and Robb (1985a,b, 1986a,b). See Blundell and Powell (2003) (who call the Heckman and Robb replacement functions control functions). A more recent nomenclature is "control variate." Matzkin (2007) provides a comprehensive discussion of identification principles for econometric estimators.) These methods postulate the existence of some unobservables θ, which may be vector valued, with the property that

(U-1) $(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, \theta,$

but allow for the possibility that

(U-2) $(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z.$

In the event (U-2) holds, these approaches model the relationships of the unobservable θ with Y_1, Y_0 and D in various ways. The content in the control function principle is to specify the exact nature of the dependence on the relationship between observables and unobservables in a nontrivial fashion that is consistent with economic theory. Heckman and Navarro present examples of models that satisfy (U-1) but not (U-2).

The early literature focused on mean outcomes conditional on covariates (Heckman and Robb 1985a, b, 1986a, b) and assumes a weaker version of (U-1) based on conditional mean independence rather than full conditional independence. More recent work analyzes distributions of outcomes (e.g., Aakvik et al. 2005; Carneiro et al. 2001, 2003). This work is reviewed in Abbring and Heckman (2007).

The normal Roy selection model makes distributional assumptions and identifies the joint distribution of outcomes. A large literature surveyed by Matzkin (2007) makes alternative assumptions to satisfy (U-1) in nonparametric settings. Replacement functions (Heckman and Robb 1985a) are methods that proxy θ. They substitute out for θ using observables. (This is the "control variate" of Blundell and Powell (2003). Heckman and Robb (1985a) and Olley and Pakes (1996) use a similar idea. Matzkin (2007) discusses replacement functions.) Aakvik et al. (1999, 2005), Carneiro et al. (2001, 2003), Cunha et al. (2005), and Cunha et al. (2006, 2010) develop methods that integrate out θ from the model assuming $\theta \perp\!\!\!\perp (X, Z)$, or invoking weaker mean independence assumptions, and assuming access to proxy measurements for θ. They also consider methods for estimating the distributions of treatment effects. These are discussed in Abbring and Heckman (2007).

The normal selection model produces partial identification of a generalized Roy model and full identification of a Roy model under separability and normality. It models the conditional expectation of U_0 and U_1 given X, Z, D. In terms of (U-1), it models the conditional mean dependence of Y_0, Y_1 on D and θ given X and Z. Powell (1994) and Matzkin (2007) survey methods for producing semiparametric versions of these models. Heckman and Vytlacil (2007a, Appendix B) or the appendix of Heckman and Navarro (2007) present a prototypical identification proof for a general selection model that implements (U-1) by estimating the distribution of θ, assuming $\theta \perp\!\!\!\perp (X, Z)$, and invoking support conditions on (X, Z).

Central to both the selection approach and the instrumental variable approach for a model with heterogenous responses is the probability of selection. This is an integral part of the Roy model previously discussed. Let Z denote variables in the choice equation. Fixing Z at different values (denoted z), I define $D(z)$ as an indicator function that is "1" when treatment is selected at the fixed value of z and that is "0" otherwise. In terms of a separable index model $U_D = \mu_D(Z) - V$, for a fixed value of z,

$$D(z) = \mathbf{1}[\mu_D(z) \geq V]$$

where $Z \perp\!\!\!\perp V \mid X$. Thus fixing $Z = z$, values of z do not affect the realizations of V for any value of X. An alternative way of representing the independence between Z and V given X due to Imbens and Angrist (1994), writes that $D(z) \perp\!\!\!\perp Z$ for all $z \in \mathcal{Z}$, where $\mathcal{Z}$ is the support of Z. The Imbens and Angrist independence condition for IV is

$$\{D(z)\}_{z \in \mathcal{Z}} \perp\!\!\!\perp Z \mid X.$$

Thus the probabilities that $D(z) = 1, z \in \mathcal{Z}$ are not affected by the occurrence of Z. Vytlacil (2002) establishes the equivalence of these two formulations under general conditions. (See Heckman and Vytlacil (2007b) for a discussion of these conditions.)

The method of instrumental variables (IV) postulates that

(IV-1) $\left(Y_0, Y_1, \{D(z)\}_{z \in \mathcal{Z}}\right) \perp\!\!\!\perp Z \mid X.$ (*Independence*)

One consequence of this assumption is that $E(D \mid Z) = P(Z)$, the propensity score, is random with respect to potential outcomes. Thus $(Y_0, Y_1) \perp\!\!\!\perp P(Z) \mid X$. So are all other functions of Z given X. The method of instrumental variables also assumes that

(IV-2) $E(D \mid X, Z) = P(X, Z)$ *is a nondegenerate function of* Z *given* X. (*Rank Condition*)

Alternatively, one can write that $Var(E(D \mid X, Z)) \neq Var(E(D \mid X))$.

Comparing (IV-1) to (M-1) in the method of instrumental variables, Z is independent of (Y_0, Y_1) given X whereas in matching D is independent of (Y_0, Y_1) given X. So in (IV-1), Z plays the role of D in matching condition (M-1). Comparing (IV-2) with (M-2), in the method of IV the choice probability $\Pr(D = 1 \mid X, Z)$ is assumed to vary with Z conditional on X, whereas in matching, D varies conditional on X. Unlike the method of control functions, no explicit model of the relationship between D and (Y_0, Y_1) is required in applying IV.

(IV-2) is a rank condition and can be empirically verified. (IV-1) is not testable as it involves assumptions about counterfactuals. In a conventional common coefficient regression model

$$Y = \alpha + \beta D + U,$$

where β is a constant and where I allow for $Cov(D, U) \neq 0$, (IV-1) and (IV-2) identify β. $\left(\beta = \frac{Cov(Z,Y)}{Cov(Z,D)}.\right)$ When β varies in the population and is correlated with D, additional assumptions must be invoked for IV to identify interpretable parameters. Heckman et al. (2006) and Heckman and Vytlacil (2007b) discuss these conditions.

Assumptions (IV-1) and (IV-2), with additional assumptions in the case where β varies in the population which I discuss in this paper, can be used to identify mean treatment parameters. Replacing Y_1 with $\mathbf{1}(Y_1 \leq t)$ and Y_0 with $\mathbf{1}(Y_0 \leq t)$, where t is a constant, the IV approach allows us to identify marginal distributions $F_1(y_1 \mid X)$ or $F_0(y_0 \mid X)$.

In matching, the variation in D that arises after conditioning on X provides the source of randomness that switches people across treatment status. Nature is assumed to provide an experimental manipulation conditional on X that replaces the randomization assumed in (R-1)–(R-3). When D is perfectly predictable by X, there is no variation in it conditional on X, and the randomization assumed to be given by nature in the matching model breaks down. Heuristically, matching assumes a residual $\mathcal{E}(X) = D - E(D \mid X)$ that is nondegenerate and is one manifestation of the randomness that causes persons to switch status. (It is heuristically illuminating, but technically incorrect to replace $\mathcal{E}(X)$ with D in (R-1) or R in (R-2) or T in (R-3). In general $\mathcal{E}(X)$ is not independent of X even if it is mean independent.)

In the IV method, it is the choice probability $E(D \mid X, Z) = P(X, Z)$ that is random with respect to (Y_0, Y_1), not components of D not predictable by (X, Z). Variation in Z for a fixed X provides the required variation in D that switches treatment status and still produces the required conditional independence:

$$(Y_0, Y_1) \perp\!\!\!\perp P(X, Z) \mid X.$$

Variation in $P(X, Z)$ produces variations in D that switch treatment status. Components of variation in D not predictable by (X, Z) do not produce the required independence. Instead, the predicted component provides the required independence. It is just the opposite in matching. Versions of the method of control functions use measurements to proxy θ in (U-1) and (U-2) and remove spurious dependence that gives rise to selection problems. These are called replacement functions (see Heckman and Robb 1985a) or control variates (see Blundell and Powell 2003).

The methods of replacement functions and proxy variables all start from characterizations (U-1) and (U-2). θ is not observed and (Y_0, Y_1) are not observed directly but Y is observed:

$$Y = DY_1 + (1 - D) Y_0.$$

Missing variables θ produce selection bias which creates a problem with using observational data to evaluate social programs. From (U-1), if one conditions on θ, condition (M-1) for matching would be satisfied, and hence one could identify the parameters and distributions that can be identified if the conditions required for matching are satisfied.

The most direct approach to controlling for θ is to assume access to a function $\tau(X, Z, Q)$ that perfectly proxies θ:

$$\theta = \tau(X, Z, Q). \tag{10}$$

This approach based on a perfect proxy is called the method of replacement functions by Heckman and Robb (1985a). In (U-1), one can substitute for θ in terms of observables (X, Z, Q). Then

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, Q.$$

It is possible to condition nonparametrically on (X, Z, Q) and without having to know the exact functional form of τ. θ can be a vector and τ can be a vector of functions. This method has been used in the economics of education for decades (see the references in Heckman and Robb 1985a). If θ is ability and τ is a test score, it is sometimes assumed

that the test score is a perfect proxy (or replacement function) for θ and that one can enter it into the regressions of earnings on schooling to escape the problem of ability bias (typically assuming a linear relationship between τ and θ). (Thus if $\tau = \alpha_0 + \alpha_1 X + \alpha_2 Q + \alpha_3 Z + \theta$, one can write

$$\theta = \tau - \alpha_0 - \alpha_1 X - \alpha_2 Q - \alpha_3 Z,$$

and use this as the proxy function. Controlling for T, X, Q, Z controls for θ. Notice that one does not need to know the coefficients $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$ to implement the method, one can condition on X, Q, Z.) Heckman and Robb (1985a) discuss the literature that uses replacement functions in this way. Olley and Pakes (1996) apply this method and consider nonparametric identification of the τ function. Matzkin (2007) provides a rigorous proof of identification for this approach in a general nonparametric setting.

The method of replacement functions assumes that (10) is a perfect proxy. In many applications, this assumption is far too strong. More often, θ is measured with error. This produces a factor model or measurement error model (Aigner et al. 1984). Matzkin (2007) surveys this method. One can represent the factor model in a general way by a system of equations:

$$Y_j = g_j(X, Z, Q, \theta, \varepsilon_j), \quad j = 1, \ldots, J. \tag{11}$$

A linear factor model separable in the unobservables writes

$$Y_j = g_j(X, Z, Q) + \alpha_j \theta + \varepsilon_j, \quad j = 1, \ldots, J, \tag{12}$$

where

$$(X, Z, Q) \perp\!\!\!\perp (\theta, \varepsilon_j), \varepsilon_j \perp\!\!\!\perp \theta, \quad j = 1, \ldots, J, \tag{13}$$

and the ε_j are mutually independent. Observe that under (11) and (12), Y_j controlling for X, Z, Q only imperfectly proxies θ because of the presence of ε_j. θ is called a factor, α_j factor loadings and the ε_j "uniquenesses" (see e.g., Aigner 1985).

A large literature, reviewed in Abbring and Heckman (2007) and Matzkin (2007) shows how to establish identification of econometric models under factor structure assumptions. Cunha et al. (2010), Schennach (2004) and Hu and Schennach (2008) establish identification in nonlinear models of the form (11). (Cunha et al. (2006, 2010) apply and extend this approach to a dynamic factor setting where the θ_t are time dependent.) The key to identification is multiple, but imperfect (because of ε_j), measurements on θ from the Y_j, $j = 1, \ldots, J$ and X, Z, Q, and possibly other measurement systems that depend on θ. Carneiro et al. (2003), Cunha et al. (2005, 2006) and Cunha and Heckman (2007, 2008) apply and develop these methods. Under assumption (13), they show how to nonparametrically identify the econometric model and the distributions of the unobservables $F_\Theta(\theta)$ and $F_{\xi_j}(\varepsilon_j)$. In the context of classical simultaneous equations models, identification is secured by using covariance restrictions across equations exploiting the low dimensionality of vector θ compared to the high dimensional vector of (imperfect) measurements on it. The recent literature (Cunha et al. 2003; Cunha et al. 2010; Hu and Schennach 2008) extends the linear model to a nonlinear setting.

The recent econometric literature applies in special cases the idea of the control function principle introduced in Heckman and Robb (1985a). This principle, versions of which can be traced back to Telser (1964), partitions θ in (U-1) into two or more components, $\theta = (\theta_1, \theta_2)$, where only one component of θ is the source of bias. Thus it is assumed that (U-1) is true, and (U-1)′ is also true:

$$\text{(U-1)}' \quad (Y_0, Y_1) \perp\!\!\!\perp D \mid X, Z, \theta_1,$$

and (U-2) holds. For example, in a normal selection model with additive separability, one can break U_1, the error term associated with Y_1, into two components:

$$U_1 = E(U_1 \mid V) + \varepsilon,$$

where V plays the role of θ_1 and is associated with the choice equation. Under normality, ε is independent of $E(U_1 \mid V)$. Further,

$$E(U_1 \mid V) = \frac{Cov(U_1, V)}{Var(V)} V, \tag{14}$$

assuming $E(U_1) = 0$ and $E(V) = 0$. Heckman and Robb (1985a) show how to construct a control function in the context of the choice model

$$D = \mathbf{1}[\mu_D(Z) > V].$$

Controlling for V controls for the component of θ_1 in (U-1)′ that gives rise to the spurious dependence. The Blundell and Powell (2003, 2004) application of the control function principle assumes functional form (14) but assumes that V can be perfectly proxied by a first stage equation. Thus they use a replacement function in their first stage. Their method does not work when one can only condition on D rather than on $D^* = \mu_D(Z) - V$ instead of directly measuring it. (Imbens and Newey (2002) extend their approach. See the discussion in Matzkin (2007).) In the sample selection model, it is not necessary to identify V. As developed in Heckman and Robb (1985a) and

Heckman and Vytlacil (2007a, b), under additive separability for the outcome equation for Y_1, one can write

$$E(Y_1 \mid X, Z, D = 1) = \mu_1(X) + \underbrace{E(U_1 \mid \mu_D(Z) > V)}_{\text{control function}},$$

so the analyst "expects out" rather than solve out the effect of the component of V on U_1 and thus control for selection bias under the maintained assumptions. In terms of the propensity score, under the conditions specified in Heckman and Vytlacil (2007a), one may write the preceding expression in terms of $P(Z)$:

$$E(Y_1 \mid X, Z, D = 1) = \mu_1(X) + K_1(P(Z)),$$

where $K_1(P(Z)) = E(U_1 \mid X, Z, D = 1)$. It is not literally necessary to know V or be able to estimate it. The Blundell and Powell (2003, 2004) application of the control function principle assumes that the analyst can condition on and estimate V.

The Blundell and Powell method and the method of Imbens and Newey (2002) build heavily on (14) and implicitly make strong distributional and functional form assumptions that are not intrinsic to the method of control functions. As just noted, their method uses a replacement function to obtain $E(U_1 \mid V)$ in the first step of their procedures. The general control function method does not require a replacement function approach. The literature has begun to distinguish between the more general control function approach and the *control variate* approach that uses a first stage replacement function.

Matzkin (2003) develops the method of unobservable instruments which is a version of the replacement function approach applied to ▶nonlinear models. Her unobservable instruments play the role of covariance restrictions used to identify classical simultaneous equations models (see Fisher, 1966). Her approach is distinct from and therefore complementary with linear factor models. Instead of assuming $(X, Z, Q) \perp\!\!\!\perp (\theta, \varepsilon_j)$, she assumes in a two equation system that $(\theta, \varepsilon_1) \perp\!\!\!\perp Y_2 \mid Y_1, X, Z$. See Matzkin (2007).

I do not discuss panel data methods in this paper. The most commonly used panel data method is difference-in-differences as discussed in Heckman and Robb (1985a), Blundell et al. (1998), Heckman et al. (1999), and Bertrand et al. (2004), to cite only a few of the key papers. Most of the estimators I have discussed can be adapted to a panel data setting. Heckman et al. (1998) develop difference-in-differences matching estimators. Abadie (2002) extends this work. (There is related work by Athey and Imbens (2006), which exposits the Heckman et al. (1998) difference-in-differences matching estimator.) Separability between errors and observables is a key feature of the panel data approach in its standard application. Altonji and Matzkin (2005) and Matzkin (2003) present analyses of nonseparable panel data methods. Regression discontinuity estimators, which are versions of IV estimators, are discussed by Heckman and Vytlacil (2007b).

Table 3 summarizes some of the main lessons of this section. I stress that the stated conditions are necessary conditions. There are many versions of the IV and control functions principle and extensions of these ideas which refine these basic postulates. See Heckman and Vytlacil (2007b). Matzkin (2007) is an additional reference on sources of identification in econometric models.

I next introduce the generalized Roy model and the concept of the marginal treatment effect which helps to link the econometric literature to the statistical literature. The Roy model also provides a framework for thinking about the difference in information between the agents and the statistician.

Matching

The method of matching is widely-used in statistics. It is based on strong assumptions which often make its application to economic data questionable. Because of its popularity, I single it out for attention. The method of matching assumes selection of treatment based on potential outcomes

$$(Y_0, Y_1) \not\perp\!\!\!\perp D,$$

so $\Pr(D = 1 \mid Y_0, Y_1)$ depends on Y_0, Y_1. It assumes access to variables Q such that conditioning on Q removes the dependence:

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid Q. \qquad \text{(Q-1)}$$

Thus,

$$\Pr(D = 1 \mid Q, Y_0, Y_1) = \Pr(D = 1 \mid Q).$$

Comparisons between treated and untreated can be made at all points in the support of Q such that

$$0 < \Pr(D = 1 \mid Q) < 1. \qquad \text{(Q-2)}$$

The method does not explicitly model choices of treatment or the subjective evaluations of participants, nor is there any distinction between the variables in the outcome equations (X) and the variables in the choice equations (Z) that is central to the IV method and the method of control functions. In principle, condition (Q-1) can be satisfied using a set of variables Q distinct from all or some of the components of X and Z. The conditioning variables do not have to be exogenous.

Principles Underlying Econometric Estimators for Identifying Causal Effects. Table 3 Identifying assumptions under commonly used methods

	Identifying assumptions	Identifies marginal distributions?	Exclusion condition needed?
Random assignment	$(Y_0, Y_1) \perp\!\!\!\perp \xi, \xi = 1 \Longrightarrow A = 1, \xi = 0 \Longrightarrow A = 0$ (full compliance) Alternatively, if self-selection is random with with respect to outcomes, $(Y_0, Y_1) \perp\!\!\!\perp D$. Assignment can be conditional on X.	Yes	No
Matching	$(Y_0, Y_1) \not\perp\!\!\!\perp D$, but $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$, $0 < \Pr(D = 1 \mid X) < 1$ for all X So D conditional on X is a nondegenerate random variable	Yes	No
Control functions and extensions	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp D \mid X, Z, \theta$. The method models dependence induced by θ or else proxies θ (replacement function) Version (i) Replacement functions (substitute out θ by observables) (Blundell and Powell, 2003; Heckman and Robb, 1985b; Olley and Pakes, 1996). Factor models Carneiro et al., (2003) allow for measurement error in the proxies. Version (ii) Integrate out θ assuming $\theta \perp\!\!\!\perp (X, Z)$ (Aakvik et al., 2005; Carneiro et al., 2003) Version (iii) For separable models for mean response expect θ conditional on X, Z, D as in standard selection models (control functions in the same sense of Heckman and Robb).	Yes	Yes (for semiparametric models)
IV	$(Y_0, Y_1) \not\perp\!\!\!\perp D \mid X, Z$, but $(Y_1, Y_0) \perp\!\!\!\perp Z \mid X$, $\Pr(D = 1 \mid Z)$ is a nondegenerate function of Z	Yes	Yes

(Y_0, Y_1) are potential outcomes that depend on X

$$D = \begin{cases} 1 \text{ if assigned (or choose) status 1} \\ 0 \text{ otherwise} \end{cases}$$

Z are determinants of D, θ is a vector of unobservables

For random assignments, A is a vector of actual treatment status. $A = 1$ if treated; $A = 0$ if not.

$\xi = 1$ if a person is randomized to treatment status; $\xi = 0$ otherwise (Heckman and Vytlacil 2007b)

From condition (Q-1) one recovers the distributions of Y_0 and Y_1 given Q – $\Pr(Y_0 \le y_0 \mid Q = q) = F_0(y_0 \mid Q = q)$ and $\Pr(Y_1 \le y_1 \mid Q = q) = F_1(y_1 \mid Q = q)$ – but not the joint distribution $F_{0,1}(y_0, y_1 \mid Q = q)$, because the analyst does not observe the same persons in the treated and untreated states. This is a standard evaluation problem common to all econometric estimators. Methods for determining which variables belong in Q rely on untested exogeneity assumptions which we discuss in this section.

OLS is a special case of matching that focuses on the identification of conditional means. In OLS linear functional forms are maintained as exact representations or valid approximations. Considering a common coefficient model, OLS writes

$$Y = \pi Q + D\alpha + U, \tag{Q-3}$$

where α is the treatment effect and

$$E(U \mid Q, D) = 0. \tag{Q-4}$$

The assumption is made that the variance-covariance matrix of (Q, D) is of full rank:

$$Var(Q, D) \text{ full rank.} \tag{Q-5}$$

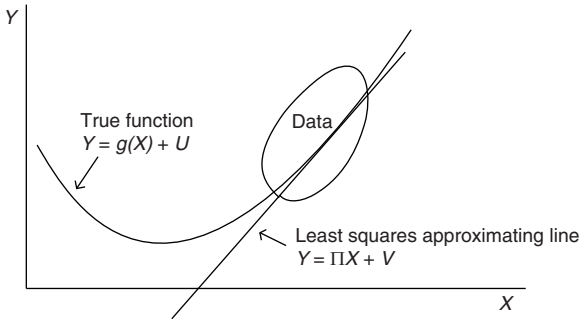
Under these conditions, one can identify α even though D and U are dependent: $D \not\perp\!\!\!\perp U$. Controlling for the observable Q eliminates any spurious mean dependence between D and U: $E(U \mid D) \neq 0$ but $E(U \mid D, Q) = 0$. (Q-3) is the linear regression counterpart to (Q-1). (Q-5) is the linear regression counterpart to (Q-2). Failure of (Q-5) would mean that using a nonparametric estimator one might perfectly predict D given Q, and that $\Pr(D = 1 \mid Q = q) = 1$ or 0. (This condition might be met only at certain values of $Q = q$. For certain parameterizations (e.g., the linear probability model), one may obtain predicted probabilities outside the unit interval.)

Matching can be implemented as a nonparametric method. When this is done, the procedure does not require specification of the functional form of the outcome equations. It enforces the requirement that (Q-2) be satisfied by estimating functions pointwise in the support of Q. Assume that $Q = (X, Z)$ and that X and Z are the same except where otherwise noted. Thus I invoke assumptions (M-1) and (M-2) presented in section "►The Basic Principles Underlying the Identification of the Leading Econometric Evaluation Estimators", even though in principle one can use a more general conditioning set.

Assumptions (M-1) and (M-2) or (Q-1) and (Q-2) rule out the possibility that after conditioning on X (or Q), agents possess more information about their choices than econometricians, and that the unobserved information helps to predict the potential outcomes. Put another way, the method allows for potential outcomes to affect choices but only through the observed variables, Q, that predict outcomes. This is the reason why Heckman and Robb (1985a, 1986b) call the method selection on observables.

Heckman and Vytlacil (2007b) establish the following points. (1) Matching assumptions (M-1) and (M-2) generically imply a flat MTE in u_D, i.e., they assume that $E(Y_1 - Y_0 \mid X = x, U_D = u_D)$ does not depend on u_D. Thus the unobservables central to the Roy model and its extensions and the unobservables central to the modern IV literature are assumed to be absent once the analyst conditions on X. (M-1) implies that all mean treatment parameters are the same. (2) Even if one weakens (M-1) and (M-2) to mean independence instead of full independence, generically the MTE is flat in u_D under the assumptions of the nonparametric generalized Roy model developed in section "►An Index Model of Choice and Treatment Effects: Definitions and Unifying Principles", so again all mean treatment parameters are the same. (3) IV and matching make distinct identifying assumptions even though they both invoke conditional independence assumptions. (4) Comparing matching with IV and control function (sample selection) methods, matching assumes that conditioning on observables eliminates the dependence between (Y_0, Y_1) and D. The control function principle models the dependence. (5) Heckman and Navarro (2004) and Heckman and Vytlacil (2007b) demonstrate that if the assumptions of the method of matching are violated, the method can produce substantially biased estimators of the parameters of interest. (6) Standard methods for selecting the conditioning variables used in matching assume exogeneity. Violations of the exogeneity assumption can produce biased estimators.

Nonparametric versions of matching embodying (M-2) avoid the problem of making inferences outside the support of the data. This problem is implicit in any application of least squares. Figure 7 shows the support problem that can arise in linear least squares when the linearity of the regression is used to extrapolate estimates determined in one empirical support to new supports. Careful attention to support problems is a virtue of any nonparametric method, including, but not unique to, nonparametric matching. Heckman, Ichimura, Smith, and Todd (1998)



Principles Underlying Econometric Estimators for Identifying Causal Effects. Fig. 7 The least squares extrapolation problem avoided by using nonparametric regression or matching (Heckman and Vytlacil 2007b)

show that the bias from neglecting the problem of limited support can be substantial. See also the discussion in Heckman, LaLonde, and Smith (1999).

Summary

This paper exposits the basic economic model of causality and compares it to models in statistics. It exposits the key identifying assumptions of commonly used econometric estimators for causal inference. The emphasis is on the economic content of these assumptions. I discuss how matching makes strong assumption about the information available to economist/statistician.

Acknowledgments

University of Chicago, Department of Economics, 1126 E. 59th Street, Chicago IL 60637, USA. This research was supported by NSF: 97-09-873, 00-99195, and SES-0241858 and NICHD: R01-HD32058-03. I thank Mohan Singh, Sergio Urzua, and Edward Vytlacil for useful comments.

About the Author

Professor Heckman shared the Nobel Memorial Prize in Economics in 2000 with Professor Daniel McFadden for his development of theory and methods for analyzing selective samples. Professor Heckman has also received numerous awards for his work, including the John Bates Clark Award of the American Economic Association in 1983, the 2005 Jacob Mincer Award for Lifetime Achievement in Labor Economics, the 2005 University College Dublin Ulysses Medal, the 2005 Aigner award from the Journal of Econometrics, and Gold Medal of the President of the Italian Republic, Awarded by the International Scientific Committee, in 2008. He holds six honorary doctorates. He is considered to be among the five most influential economists in the world, in 2010. (http://ideas.repec.org/top/top.person.all.html).

Cross References

►Causal Diagrams
►Causation and Causal Inference
►Econometrics
►Factor Analysis and Latent Variable Modelling
►Instrumental Variables
►Measurement Error Models
►Panel Data
►Random Coefficient Models
►Randomization

References and Further Reading

Aakvik A, Heckman JJ, Vytlacil EJ (1999) Training effects on employment when the training effects are heterogeneous: an application to Norwegian vocational rehabilitation programs. University of Bergen Working Paper 0599, University of Chicago

Aakvik A, Heckman JJ, Vytlacil EJ (2005) Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. J Econometrics 125:15–51

Abadie A (2002) Bootstrap tests of distributional treatment effects in instrumental variable models. J Am Stat Assoc 97:284–292

Abbring JH, Heckman JJ (2007) Econometric evaluation of social programs, part III: distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 6B. Elsevier, Amsterdam, pp 5145–5303

Aigner DJ (1985) The residential electricity time-of-use pricing experiments: what have we learned? In: Hausman JA, Wise DA (eds) Social experimentation. University of Chicago Press, Chicago, pp 11–41

Aigner DJ, Hsiao C, Kapteyn A, Wansbeek T (1984) Latent variable models in econometrics. In: Griliches Z, Intriligator MD (eds) Handbook of econometrics, vol 2, chap 23. Elsevier, Amsterdam, pp 1321–1393

Altonji JG, Matzkin RL (2005) Cross section and panel data estimators for nonseparable models with endogenous regressors. Econometrica 73:1053–1102

Angrist J, Graddy K, Imbens G (2000) The interpretation of instrumental variables estimators in simultaneous equationsmodelswith an application to the demand for fish. Rev Econ Stud 67:499–527

Angrist JD, Krueger AB (1999) Empirical strategies in labor economics. In: Ashenfelter O, Card D (eds) Handbook of labor economics, vol 3A. North-Holland, New York, pp 1277–1366

Athey S, Imbens GW (2006) Identification and inference in nonlinear difference-indifferences models. Econometrica 74:431–497

Barnow BS, Cain GG, Goldberger AS (1980) Issues in the analysis of selectivity bias. In: Stromsdorfer E, Farkas G (eds) Evaluation studies, vol 5. Sage Publications, Beverly Hills, pp 42–59

Barros RP (1987) Two essays on the nonparametric estimation of economic models with selectivity using choice-based samples. PhD thesis, University of Chicago

Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-indifferences estimates? Q J Econ 119:249–275

Björklund A, Moffitt R (1987) The estimation of wage gains and welfare gains in self-selection. Rev Econ Stat 69:42–49

Blundell R, Duncan A, Meghir C (1998) Estimating labor supply responses using tax reforms. Econometrica 66:827–861

Blundell R, Powell J (2003) Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont LPHM, Turnovsky SJ (eds) Advances in economics and econometrics: theory and applications, eighth world congress, vol 2. Cambridge University Press, Cambridge

Blundell R, Powell J (2004) Endogeneity in semiparametric binary responsemodels. Rev Econ Stud 71:655–679

Carneiro P, Hansen K, Heckman JJ (2001) Removing the veil of ignorance in assessing the distributional impacts of social policies. Swedish Econ Policy Rev 8:273–301

Carneiro P, Hansen K, Heckman JJ (2003) Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. Int Econ Rev 44:361–422

Cunha F, Heckman JJ (2007) Identifying and estimating the distributions of Ex Post and Ex Ante returns to schooling: a survey of recent developments. Labour Econ 14:870–893

Cunha F, Heckman JJ (2008) A new framework for the analysis of inequality. Macroecon Dyn 12:315–354
Cunha F, Heckman JJ, Matzkin R (2003) Nonseparable factor analysis. Unpublished manuscript, University of Chicago, Department of Economics
Cunha F, Heckman JJ, Navarro S (2005) Separating uncertainty fromheterogeneity in life cycle earnings. The 2004 Hicks Lecture. Oxford Economic Papers 57, 191–261
Cunha F, Heckman JJ, Navarro S (2006) Counterfactual analysis of inequality and social mobility. In: Morgan SL, Grusky DB, Fields GS (eds) Mobility and inequality: frontiers of research in sociology and economics, chap 4. Standford University Press, Stanford, pp 290–348
Cunha F, Heckman JJ, Schennach SM (2006) Nonlinear factor analysis. Unpublished manuscript, University of Chicago, Department of Economics, revised 2008
Cunha F, Heckman JJ, Schennach SM (2010) Estimating the technology of cognitive and noncognitive skill formation, Forthcoming. Econometrica
Fisher RA (1966) The design of experiments. Hafner Publishing, New York
Gerfin M, Lechner M (2002) Amicroeconomic evaluation of the active labormarket policy in Switzerland. Econ J 112:854–893
Heckman JJ (1992) Randomization and social policy evaluation. In: Manski C, Garfinkel I (eds) Evaluating welfare and training programs. Harvard University Press, Cambridge, pp 201–230
Heckman JJ (1997) Instrumental variables: a study of implicit behavioral assumptions used inmaking programevaluations. J Hum Resour 32:441–462; addendum published vol. 33 no. 1 (Winter 1998)
Heckman JJ (2008) Econometric causality. Int Stat Rev 76:1–27
Heckman JJ, Ichimura H, Smith J, Todd PE (1998) Characterizing selection bias using experimental data. Econometrica 66: 1017–1098
Heckman JJ, LaLonde RJ, Smith JA (1999) The economics and econometrics of active labor market programs. In: Ashenfelter O, Card D (eds) Handbook of labor economics, vol 3A, chap 31. North-Holland, New York, pp 1865–2097
Heckman JJ, Navarro S (2004) Usingmatching, instrumental variables, and control functions to estimate economic choice models. Rev Econ Stat 86:30–57
Heckman JJ, Navarro S (2007) Dynamic discrete choice and dynamic treatment effects. J Econometrics 136:341–396
Heckman JJ, Robb R (1985a) Alternative methods for evaluating the impact of interventions. In: Heckman J, Singer B (eds) Longitudinal analysis of labor market data, vol 10. Cambridge University Press, New York, pp 156–245
Heckman JJ, Robb R (1985b) Alternative methods for evaluating the impact of interventions: an overview. J Econometrics 30: 239–267
Heckman JJ, Robb R (1986a) Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In: Wainer H (ed) Drawing inferences from self-selected samples. Springer, New York, pp 63–107, reprinted in 2000, Erlbaum, Mahwah
Heckman JJ, Robb R (1986b) Postscript: a rejoinder to Tukey. In: Wainer H (ed) Drawing inferences from self-selected samples. Springer, New York, pp 111–114, reprinted in 2000, Erlbaum, Mahwah
Heckman JJ, Smith JA, Clements N (1997) Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. Rev Econ Stud 64: 487–536
Heckman JJ, Urzua S, Vytlacil EJ (2006) Understanding instrumental variables in models with essential heterogeneity. Rev Econ Stat 88:389–432
Heckman JJ, Vytlacil EJ (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proc Natl Acad Sci 96:4730–4734
Heckman JJ, Vytlacil EJ (2001) Local instrumental variables. In: Hsiao C, Morimune K, Powell JL (eds) Nonlinear statistical modeling: proceedings of the thirteenth international symposium in economic theory and econometrics: essays in honor of Takeshi Amemiya. Cambridge University Press, New York, pp 1–46
Heckman JJ, Vytlacil EJ (2005) Structural equations, treatment effects and econometric policy evaluation. Econometrica 73:669–738
Heckman JJ, Vytlacil EJ (2007a) Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 6B. Elsevier, Amsterdam, pp 4779–4874
Heckman JJ, Vytlacil EJ (2007b) Econometric evaluation of social programs, part II: using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 6B. Elsevier, Amsterdam, pp 4875–5144
Hu Y, Schennach SM (2008) Instrumental variable treatment of nonclassical measurement error models. Econometrica 76:195–216
Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 86:4–29
Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. Econometrica 62:467–475
Imbens GW, Newey WK (2002) Identification and estimation of triangular simultaneous equations models without additivity. Technical working paper 285, National Bureau of Economic Research
Matzkin RL (2003) Nonparametric estimation of nonadditive randomfunctions. Econometrica 71:1339–1375
Matzkin RL (2007) Nonparametric identification. In: Heckman J, Leamer E (eds) Handbook of econometrics, vol 6B. Elsevier, Amsterdam
Olley GS, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. Econometrica 64:1263–1297
Pearl J (2000) Causality. Cambridge University Press, Cambridge
Powell JL (1994) Estimation of semiparametric models. In: Engle R, McFadden D (eds) Handbook of econometrics, vol 4. Elsevier, Amsterdam, pp 2443–2521
Quandt RE (1958) The estimation of the parameters of a linear regression systemobeying two separate regimes. J Am Stat Assoc 53:873–880
Quandt RE (1972) A new approach to estimating switching regressions. J Am Stat Assoc 67:306–310
Roy A (1951) Some thoughts on the distribution of earnings. Oxford Econ Pap 3:135–146
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701
Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6:34–58

Schennach SM (2004) Estimation of nonlinear models with measurement error. Econometrica 72:33–75
Telser LG (1964) Iterative estimation of a set of linear regression equations. J Am Stat Assoc 59:845–862
Vytlacil EJ (2002) Independence, monotonicity, and latent index models: an equivalence result. Econometrica 70:331–341

Prior Bayes: Rubin's View of Statistics

Herman Rubin
Purdue University, West Lafayette, IN, USA

Introduction

What is statistics? The common way of looking at it is a collection of methods, somehow or other produced, and one should use one of those methods for a given set of data. The typical user who has little more than this comes to a statistician asking for THE answer, as if the data are sufficient to get this without knowing the problem.

No; the first thing is to formulate the problem. One cannot do better than to assume that there is an unknown state of nature, that there is a probability distribution of observations given a state of nature, a set of possible actions, and that each action in each state of nature has (possibly random) consequences.

Following the von Neumann–Morgenstern (1944) axioms for utility, in 1947 (see Rubin 1987a) I was able to show that if one has a self-consistent evaluation of actions in each state of nature, the utility function for an unknown state of nature has to be an integral of the utilities for the states of nature. Another way of looking at this in the discrete case is that one assigns a weight to each result in each state of nature, and should choose the action which produces the best sum; this generalizes to the best value of the integral. This is the prior Bayes approach.

Let us simplify to the usual Bayes model; it can be generalized to include more, and in fact, must be for the infinite parametric (usually called non-parametric) problems encountered.

So the quantity to be minimized is

$$\int \int L(\omega, q(x))d\mu(x|\omega)d\xi(\omega).$$

If this integral is finite, and if, for example, L is positive, the integration can be interchanged and the result written as

$$\int \int L(\omega, q(x))d\phi(\omega|x)dm(x),$$

and if the properties of q for different x are unrestricted, one can (hopefully) use the usual Bayes proceduce of minimizing the inner integral.

But this can require a huge amount of computing power, possibly even exceeding the capacity of the universe,and sufficient assurance that one has a sufficiently good approximation to the loss-prior combination. One use this latter because it is only the product of L and ξ which is relevant. One can try to approximate, but posterior robustness results are hard, and often impossible, to come by.

On the other hand, the prior Bayes approach can show what is and what is not important, and can, in many cases, provide methods which are not much worse than full Bayes methods, or at least the approximations made. One might question how this can be shown, considering that the full Bayes procedure cannot be calculated; however, a smaller problem with one which can be calculated might be shown to come close. Also, one can get procedures which are good with considerable uncertainty about the prior distribution of the parameter.

Results and Examples

Some early approaches, some by those not believing in the Bayes approach, were the empirical Bayes results of Robbins and his followers. Empirical Bayes extends much farther now, and it is in the spirit of prior Bayes, as the performance of the procedures is what is considered. There is a large literature on this, and I will not go into it in any great detail.

Suppose we consider the case of the usual test of a point null, when the distribution is normal. If we assume the prior is symmetric, we need only consider procedures which accept if the mean is close enough to the null, and reject otherwise. If we assume that the prior gives a point mass at the null, and is otherwise given by a smooth density, the prior Bayes risk is

$$\xi(\{0\})P(|X > c|0) + \int Q(\omega)h(\omega)P(|X| \le c|\omega)d\omega,$$

where Q is the loss of incorrect acceptance.

This shows that the tails of the prior distribution are essentially irrelevant if the variance is at all small, so the prior probability of the null is NOT an important consideration. Only the ratio of the probability of the null to the density at the null of the alternative is important. This shows that the large-sample results of Rubin and Sethuraman (1965) can be a good approximation for moderate, or even small, samples. It also shows how the "p-value" should change with the sample size. An expository paper is in preparation.

If the null is not a point null, this is still a good approximation if the width of the null is smaller than the standard deviation of the observations; if it is not, the problem is much harder, and the form of the loss-prior combination under the null becomes of considerable importance. Again, the prior Bayes approach shows where the problem lies, in the behavior of the loss-prior combination near the division point between accepptance and rejection. In "standard" units, a substantial part of the parameter space may be involved.

For another example, consider the problem of estimating an infinite dimensional normal mean with the covariance matrix a multiple of the identity, or the similar problem of estimating a spectral density function. The first problem was considereed by Rubin (1987b), and was corrrected in Hui Xu's thesis in 2008. If one assumes the prior mean square of the kth mean, or the corresponding Fourier coefficient, is non-increasing, an empirical Bayes type result can be obtained, and can be shown to be asymptotically optimal in the class of "simple" procedures, which are the ones currently being used with more stringent assumptions, and which are generally not as good. The results are good even if the precise form is not known, while picking a kernel is much more restrictive and generally not even asymptotically optimal.

For the latter problem, obtaining a reasonable prior seems to be extremely difficult, but the simple procedures obtained are likely to be rather good even if one is found. The positive definiteness of the Toeplitz matrix of the covariances is a difficult condition to work with.

I see a major set of applications to non-parametric problems, properly called infinite parametric, problems such as density estimation, testing with infinite dimensional alternatives, etc. With a large number of dimensions, the classical approach does not work well, and the only "simple Bayesian" approaches use priors which look highly constrained to me.

About the Author

"Professor Rubin has contributed in deep and original ways to statistical theory and philosophy. The statistical community has been vastly enriched by his contributions through his own research and through his influence, direct or indirect on the research and thinking of others." Erich Marchard and William Strawdeman, *A festschrift for Herman Rubin*, A. DasGupta (ed.), p. 21. "He is well known for his broad ranging mathematical research interests and for fundamental contributions in Bayesian decision theory, in set theory, in estimations for simultaneous equations, in probability and in asymptotic statistics." Mary Ellen Block, Ibid. p. 408. Professor Rubin is a Fellow of the IMS, and a Fellow of the AAAS (American Association for the Advancement of Science).

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Model Selection
►Moderate Deviations
►Statistics: An Overview
►Statistics: Nelder's View

References and Further Reading

Rubin H (1987a) A weak system of axioms for "rational" behavior and the non-separability of utility from prior. Stat Decisions 5:47–58
Rubin H (1987b) Robustness in generalized ridge regression and related topics. Third Valencia Symp Bayesian Stat 3:403–410
Rubin H, Sethuraman J (1965) Bayes risk efficiency. Sankhya A 27:347–356
von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton university press, Princeton
Xu H (2008) Some applications of the prior Bayes approach. Unpublished thesis

Probabilistic Network Models

Ove Frank
Professor Emeritus
Stockholm University, Stockholm, Sweden

A network on vertex set V is represented by a function y on the set V^2 of ordered vertex pairs. The function can be univariate or multivariate and its variables can be numerical or categorical. A graph G with vertex set V and edge set E in V^2 is represented by a binary function $y = \{(u, v, y_{uv}) : (u, v) \in V^2\}$ where y_{uv} indicates whether $(u, v) \in E$. If $V = \{1, \ldots, N\}$ and vertices are ordered according to their labels, y can be given as an N by N adjacency matrix $y = (y_{uv})$. Simple undirected graphs have $y_{uv} = y_{vu}$ and $y_{vv} = 0$ for all u and v. Colored graphs have a categorical variable y with the categories labeled by colors. Graphs with more general variables y are called valued graphs or networks. If Y is a random variable with outcomes y representing networks in a specified family of networks, the probability distribution induced on this family is a probabilistic network model and Y is a representation of a random network.

Simple random graphs are defined with uniform distributions or Bernoulli distributions. Uniform models assign

equal probabilities to all graphs in a specified finite family of graphs, such as all graphs of order N and size M, or all connected graphs of order N, or all trees of order N. Bernoulli graphs (Bernoulli digraphs) have edge indicators that are independent Bernoulli variables for all unordered (ordered) vertex pairs. There is an extensive literature on such random graphs, especially on the simplest Bernoulli (p) graph, which has a common edge probability p for all vertex pairs. An extension of a fixed graph G to a Bernoulli (G, α, β) graph is a Bernoulli graph that is obtained by independently removing edges in G with probability α and independently inserting edges in the complement of G with probability β. Such models have been applied to study reliability problems in communication networks. Attempts to model the web have recently contributed to an interest in random graph models with specified degree distributions and random graph processes for very large dynamically changing graphs.

The literature on social networks describes models for finite random digraphs on $V = \{1, \ldots, N\}$ in which dyads (Y_{uv}, Y_{vu}) for $u < v$ are independent and have probabilities that depend on parameters governing in- and out-edges of each vertex and mutual edges of each vertex pair. Special cases of such models with independent dyads are obtained by assuming homogeneity for the parameters of different vertices or different groups of vertices. Extensions to models with dependent dyads include Markov graphs that allow dependence between incident dyads. Other extensions are log-linear models that assume that the log-likelihood function is a linear function of specified network statistics chosen to reflect various properties of interest in the network.

Statistical analysis of network data comprise exploratory tools for selecting appropriate probabilistic network models as well as confirmatory tools for estimating and testing various models. Many of these tools use computer intensive methods.

Applications of probabilistic network models appear in many different areas in which relationships between the units studied are essential for an understanding of their properties and characteristics. The social and behavioral sciences have contributed to the development of many network models for the study of social interaction, friendship, dominance, co-operation and competition. There are applications to criminal networks and co-offending, communication and transportation networks, vaccination programs in epidemiology, information retrieval and organizational systems, particle systems in physics, biometric cell systems. Random graphs and random fields are also theoretically developed in computer science, mathematics, and statistics. There is an exciting interplay between model development and new applications in a variety of important areas.

Many references to the literature on graphs, random graphs, and random networks are provided by the following sources.

About the Author

For biography *see* the entry ►Network Sampling.

Cross References

►Graphical Markov Models
►Network Models in Probability and Statistics
►Network Sampling
►Social Network Analysis
►Uniform Distribution in Statistics

References and Further Reading

Bonato A (2008) A course on the web graph. American Mathematical Society, Providence
Carrington P, Scott J, Wasserman S (eds) (2005) Models and methods in social network analysis. Cambridge University Press, New York
Diestel R (2005) Graph theory. Springer, Berlin/Heidelberg
Durrett R (2007) Random graph dynamics. Cambridge University Press, New York
Kolaczyk E (2009) Statistical analysis of network data. Springer, New York
Meyers R (ed) (2009) Encyclopedia of complexity and systems science. Springer, New York

P

Probability on Compact Lie Groups

David Applebaum
Professor, Head
University of Sheffield, Sheffield, UK

Introduction

Probability on groups enables us to study the interaction between chance and symmetry. In this article I'll focus on the case where symmetry is generated by continuous groups, specifically compact Lie groups. This class contains many examples such as the n-torus, special orthogonal groups $SO(n)$ and special unitary groups $SU(n)$ which are important in physics and engineering applications. It is also a very good context to demonstrate the key role played by non-commutative harmonic analysis via group representations. The classic treatise (Heyer 1977) by Heyer gives a systematic mathematical introduction to this topic while

Diaconis (1988) presents a wealth of concrete examples in both probability and statistics.

For motivation, let ρ be a probability measure on the real line. Its characteristic function $\widehat{\rho}$ is the Fourier transform $\widehat{\rho}(u) = \int_{\mathbb{R}} e^{iux}\rho(dx)$ and $\widehat{\rho}$ uniquely determines ρ. Note that the mappings $x \to e^{iux}$ are the irreducible unitary representations of $\mathbb{R}$.

Now let G be a compact Lie group and ρ be a probability measure defined on G. The group law of G will be written multiplicatively. If we are given a probability space $(\Omega, \mathcal{F}, P)$ then ρ might be the law of a G-valued random variable defined on Ω. The *convolution* of two such measures ρ_1 and ρ_2 is the unique probability measure $\rho_1 * \rho_2$ on G such that

$$\int_G f(\sigma)(\rho_1 * \rho_2)(d\sigma) = \int_G \int_G f(\sigma\tau)\rho_1(d\sigma)\rho_2(d\tau),$$

for all continuous functions f defined on G. If X_1 and X_2 are independent G-valued random variables with laws ρ_1 and ρ_2 (respectively), then $\rho_1 * \rho_2$ is the law of X_1X_2.

Characteristic Functions

Let $\widehat{G}$ be the set of all irreducible unitary representations of G. Since G is compact, $\widehat{G}$ is countable. For each $\pi \in \widehat{G}, \sigma \in G, \pi(\sigma)$ is a unitary (square) matrix acting on a finite dimensional complex inner product space V_π having dimension d_π. Every group has the trivial representation δ acting on $\mathbb{C}$ by $\delta(\sigma) = 1$ for all $\sigma \in G$. The *characteristic function* of the probability measure ρ is the matrix-valued function $\widehat{\rho}$ on $\widehat{G}$ defined uniquely by

$$\langle u, \widehat{\rho}(\pi)v\rangle = \int_G \langle u, \pi(\tau)v\rangle\rho(d\tau),$$

for all $u, v \in V_\pi$. $\widehat{\rho}$ has a number of desirable properties (Siebert 1981). It determines ρ uniquely and for all $\pi \in \widehat{G}$:

$$\widehat{\rho_1 * \rho_2}(\pi) = \widehat{\rho}_1(\pi)\widehat{\rho}_2(\pi).$$

In particular $\widehat{\delta} = 1$.

Lo and Ng (1988) considered a family of matrices $(C_\pi, \pi \in \widehat{G})$ and asked when there is a probability measure ρ on G such that $C_\pi = \widehat{\rho}(\pi)$. They found a necessary and sufficient condition to be that $C_\delta = 1$ and that the following non-negativity condition holds: for all families of matrices $(B_\pi, \pi \in \widehat{G})$ where B_π acts on V_π and for which $\sum_{\pi\in S_\pi} d_\pi \mathrm{tr}(\pi(\sigma)B_\pi) \geq 0$ for all $\sigma \in G$ and all finite subsets S_π of V_π we must have $\sum_{\pi\in S_\pi} d_\pi \mathrm{tr}(\pi(\sigma)C_\pi B_\pi) \geq 0$.

Densities

Every compact group has a bi-invariant finite Haar measure which plays the role of Lebesgue measure on $\mathbb{R}^d$ and which is unique up to multiplication by a positive real number. It is convenient to normalise this measure (so it has total mass 1) and denote it by $d\tau$ inside integrals of functions of τ. We say that a probability measure ρ has a *density* f if $\rho(A) = \int_A f(\tau)d\tau$ for all Borel sets A in G. To investigate existence of densities we need the *Peter–Weyl theorem* that the set of functions $\left\{d_\pi^{\frac{1}{2}}\pi_{ij}; 1 \leq i, j \leq d_\pi, \pi \in \widehat{G}\right\}$ are a complete orthonormal basis for $L^2(G, \mathbb{C})$. So any $f \in L^2(G, \mathbb{C})$ can be written

$$f(\sigma) = \sum_{\pi\in\widehat{G}} d_\pi \mathrm{tr}(\pi(\sigma)\widehat{f}(\pi)), \tag{1}$$

where $\widehat{f}(\pi) = \int_G f(\tau)\pi(\tau^{-1})d\tau$ is the Fourier transform. In Applebaum (2008) it was shown that ρ has a square-integrable density f (which then has an expansion as in (1)) if and only if $\sum_{\pi\in\widehat{G}} d_\pi \mathrm{tr}(\widehat{\rho}(\pi)\widehat{\rho}(\pi)^*) < \infty$ where * is the usual matrix adjoint. A sufficient condition for ρ to have a continuous density is that $\sum_{\pi\in\widehat{G}} d_\pi^{\frac{3}{2}} |\mathrm{tr}(\widehat{\rho}(\pi)\widehat{\rho}(\pi)^*)|^{\frac{1}{2}} < \infty$ in which case the series on the right hand side of (0.1) converges absolutely and uniformly (see Proposition 6.6.1 on pp. 117–118 of Faraut [2008]).

Conjugate Invariant Probabilities

Many interesting examples of probability measures are conjugate invariant, i.e., $\rho(\sigma A\sigma^{-1}) = \rho(A)$ for all $\sigma \in G$. In this case there exists $c_\pi \in \mathbb{C}$ such that $\widehat{\rho}(\pi) = c_\pi I_\pi$ where I_π is the identity matrix in V_π (Said et al. 2010). If a density exists it takes the form $f(\sigma) = \sum_{\pi\in\widehat{G}} d_\pi \overline{c_\pi}\chi_\pi(\sigma)$, where $\chi_\pi(\sigma) = \mathrm{tr}(\pi(\sigma))$ is the group character.

Example 1 *Gauss Measure.* Here $c_\pi = e^{\sigma^2\kappa_\pi}$ where $\kappa_\pi \leq 0$ is the eigenvalue of the group Laplacian corresponding to the Casimir operator $\kappa_\pi I_\pi$ on V_π and $\sigma > 0$. For example if $G = SU(2)$ then it can be parametrized by the Euler angles ψ, ϕ and θ, $\widehat{G} = \mathbb{Z}_+$, $\kappa_m = -m(m+2)$ and we have a continuous density depending only on $0 \leq \theta \leq \frac{\pi}{2}$:

$$f(\theta) = \sum_{m=0}^{\infty} (m+1)e^{-\sigma^2 m(m+2)} \frac{\sin((m+1)\theta)}{\sin(\theta)}.$$

Example 2 *Laplace Distribution.* This is a generalization of the double exponential distribution on $\mathbb{R}$ (with equal parameters). In this case $c_\pi = (1 - \beta^2\kappa_\pi)^{-1}$ where $\beta > 0$ and κ_π is as above.

Infinite Divisibility

A probability measure ρ on G is *infinitely divisible* if for each $n \in \mathbb{N}$ there exists a probability measure $\rho^{\frac{1}{n}}$ on G such that the nth convolution power $\left(\rho^{\frac{1}{n}}\right)^{*n} = \rho$. Equivalently $\widehat{\rho}(\pi) = \widehat{\rho^{\frac{1}{n}}}(\pi)^n$ for all $\pi \in \widehat{G}$. If G is connected as well as compact any such ρ can be realised as μ_1 in

a weakly continuous convolution semigroup of probability measures $(\mu_t, t \geq 0)$. For a general Lie group, such an *embedding* may not be possible and the investigation of this question has generated much research over more than 30 years (McCrudden 1998). The structure of convolution semigroups has been intensely analyzed. These give the laws of group-valued ►*Lévy processes*, i.e., processes with stationary and independent increments (Liao 2004). In particular there is a Lévy–Khintchine type formula (originally due to G.A.Hunt) which classifies these in terms of the structure of the infinitesimal generator of the associated Markov semigroup that acts on the space of continuous functions. One of the most important examples is Brownian motion (see ►Brownian Motion and Diffusions) and this has a Gaussian distribution. Another important example is the *compound Poisson process* (see ►Poisson Processes)

$$Y(t) = X_1 X_2 \cdots X_{N(t)} \tag{2}$$

where $(X_n, n \in \mathbb{N})$ is a sequence of i.i.d. random variables having common law ν (say) and $(N(t), t \geq 0)$ is an independent Poisson process of intensity $\lambda > 0$. In this case $\mu_t = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \nu^{*n}$. Note that μ_t does not have a density and it is conjugate invariant if ν is.

Applications

There is intense interest among statisticians and engineers in the *deconvolution problem* on groups. The problem is to estimate the signal density f_X from the observed density f_Y when the former is corrupted by independent noise having density f_ϵ, so the model is $Y = X\epsilon$ and the inverse problem is to untangle $f_Y = f_X * f_\epsilon$. Inverting the characteristic function enables the construction of non-parametric estimators for f_X and optimal rates of convergence are known for these when f_ϵ has certain smoothness properties (Kim and Richards 2001; Koo and Kim 2008). In Said et al. (2010) the authors consider the problem of *decompounding*, i.e., to obtain non-parametric estimates of the density of X_1 in (2) based on i.i.d. observations of a noisy version of Y: $Z(t) = \epsilon Y(t)$, where ϵ is independent of $Y(t)$. This is applied to multiple scattering of waves from complex media by working with the group $SO(3)$ which acts as rotations on the sphere.

About the Author

David Applebaum is Professor of Probability and Statistics and is currently Head of the Department of Probability and Statistics, University of Sheffield UK. He has published over 60 research papers and is the author of two books, *Probability and Information* (second edition), Cambridge University Press (1996, 2008) and *Lévy Processes and Stochastic Calculus* (second edition), Cambridge University Press (2004, 2009). He is joint managing editor of the *Tbilsi Mathematical Journal* and associate editor for *Bernoulli*, *Methodology and Computing in Applied Probability*, *Journal of Stochastic Analysis and Applications* and *Communications on Stochastic Analysis*. He is a member of the Atlantis Press Advisory Board for Probability and Statistics Studies.

Cross References

►Brownian Motion and Diffusions
►Characteristic Functions
►Lévy Processes
►Poisson Processes

References and Further Reading

Applebaum D (2008) Probability measures on compact groups which have square-integrable densities. Bull Lond Math Sci 40:1038–1044
Diaconis P (1988) Group representations in probability and statistics, Lecture Notes – Monograph Series Volume 11. Institute of Mathematical Statistics, Hayward
Faraut J (2008) Analysis on Lie groups. Cambridge University Press, Cambridge
Heyer H (1977) Probability measures on locally compact groups. Springer, Berlin/Heidelberg
Kim PT, Richards DS (2001) Deconvolution density estimators on compact Lie groups. Contemp Math 287:155–171
Koo J-Y, Kim PT (2008) Asymptotic minimax bounds for stochastic deconvolution over groups. IEEE Trans Inf Theory 54:289–298
Liao M (2004) Lévy processes in Lie groups. Cambridge University Press, Cambridge
Lo JT-H, Ng S-K (1988) Characterizing Fourier series representations of probability distributions on compact Lie groups. Siam J Appl Math 48:222–228
McCrudden M (1998) An introduction to the embedding problem for probabilities on locally compact groups. In: Hilgert J, Lawson JD, Neeb K-H, Vinberg EB (eds) Positivity in Lie theory: open problems. Walter de Gruyter, Berlin/New York, pp 147–164
Said S, Lageman C, LeBihan N, Manton JH (2010) Decompounding on compact Lie groups. IEEE Trans Inf Theory 56(6):2766–2777
Siebert E (1981) Fourier analysis and limit theorems for convolution semigroups on a locally compact group. Adv Math 39:111–154

Probability Theory: An Outline

Tamás Rudas
Professor, Head of Department of Statistics, Faculty of Social Sciences
Eötvös Loránd University, Budapest, Hungary

Sources of Uncertainty in Statistics

Statistics is often defined as the science of the methods of data collection and analysis, but from a somewhat more

conceptual perspective, statistics is also the science of methods dealing with uncertainty. The sources of uncertainty in statistics may be divided into two groups. Uncertainty associated with the data one has and uncertainty associated with respect to the mechanism which produced the data. These kinds of uncertainty are often interrelated in practice, yet it is useful to distinguish them.

Uncertainties in Data

One may distinguish two main sources of uncertainty with respect to data. One is related to measurement error, the other to sampling error.

Measurement error is the difference between the actual measurement obtained and the true value of what was measured. This applies to both cases when a numerical measurement taken (typical in the physical, biological, medical, psychological sciences) but also to qualitative observations when subjects are classified (typical in the social and behavioral sciences), except that in the latter case, instead of the numerical value of the error, its lack or presence is considered. In the former case, it is often observed that the errors are approximately distributed as normal, even if very precise and expensive measuring instruments are used. In the latter case, the existence of measurement error, that is misclassification, is often attributed to self-reflection of the human beings observed. In both situations, the lack of understanding of the precise mechanisms behind measurement errors suggest applying a stochastic model assuming that the result of the measurement is the sum of the true value plus a random error.

Sampling error stems from the uncertainty of how our results would differ, if a sample that is different from the actual one were observed. It is usually associated with the entire sample (and not with the individual observations) and is measured as the difference between the estimates obtained from the actual sample, and the census value that could be obtained if the same data collection methods were applied to the entire population. The census value may or may not be equal to the true population parameter of interest. For example, if the measurement error is assumed to be constant, then the census value differs from the population value by this quantity. Usually, the likely size of the sampling error is characterized by the standard deviation (standard error) of the estimates. Under many common random sampling schemes, the distribution of the estimates is normal, and the choice of the standard error, as a characteristic quantity, is well justified.

Uncertainties in Modeling

While uncertainties associated with the data seem to be inherent characteristics, uncertainties related to modeling are more determined by our choice of models, which depends very often on the existing knowledge regarding the research problem at hand. The most common assumption is that the quantity of interest has a specified, though unknown to the researcher, distribution, that may or may not be assumed to belong to some parametric family of distributions. A further possible choice, gaining increasing popularity during the recent decades, is that the distribution of interest belongs to a parametric family, though not with a specified parameter value, rather characterized by a probability distribution (the prior distribution) on the possible parameter values. The former view is adopted in frequentist statistics and the latter view is the Bayesian approach to statistics.

To model uncertainty, frequentist statistics uses frequentist or classical probability theory, while ▶Bayesian statistics often relies on a subjective concept of probability.

Classical and Frequentist Probability

Historically, there are two sources of modern probability theory. One is the theory of gambling, where the main goal was to determine how probable certain outcomes were in a game of chance. These problems could be appropriately handled under the assumptions that all possible outcomes of an experiment (rolling a die, for example) are equally likely and probabilities could be determined as the ratio of the number of outcomes with a certain characteristic, to the total number of outcomes. This interpretation of probability is called classical probability. Questions related to gambling also made important contributions to developing the concepts of Boolean algebra (the algebra of events associated with an experiment), conditional probability and infinite sequences of random variables (which play an important role in the frequentist interpretation of probability see below). The other source of modern probability theory is the analysis of errors associated with a measurement. This led, among others, to the understanding of the central role played by the normal distribution.

It is remarkable, that the main concepts and results of these two apparently very different fields, all may be based on one set of axioms, proposed by Kolmogorov and given in the next section.

The Kolmogorov Axioms

The axioms, summarizing the concepts developed within the classical and frequentist approaches, apply to experiments that may be repeated infinitely many times, where all circumstances of the experiment are supposed to remain constant. An experiment may be identified with its possible outcomes. Certain subsets of outcomes are called events, with he assumption that no outcome (the impossible event) and all the outcomes (the certain event) are

events and countable unions or intersections of events are also events. This means that the set of events associated with an experiment form a sigma-field. Then the Kolmogorov axioms (basic assumptions that are accepted to be true) are the following:

For any event A, its probability

$$P(A) \geq 0$$

For the certain event Ω

$$P(\Omega) = 1$$

For a series $A_i, i = 1, 2, \ldots$ of pairwise disjoint events

$$\Sigma_{i=1,2,\ldots} P(A_i) = P(\Sigma_{i=1,2,\ldots} A_i)$$

The heuristic interpretation is that the probability of an event manifests itself via the relative frequency of this event over long series of repetitions of the experiment. This is why this approach to probability is often called frequentist probability. Indeed, the axioms are true for relative frequencies instead of probabilities.

The Laws of Large Numbers

The link between the heuristic notion of probability and the mathematical theory of probability is established by the result that if $f_n(A)$ denotes the frequency of event A after n repetitions of an experiment, then

$$f_n(A)/n \to P(A),$$

where the convergence $\to$ may be given various interpretations. More generally, if X is a random variable (that is, such a function that $X \in I$ is an event for every interval I), then for the average of n independent observations of X, $\overline{X}_n$,

$$\overline{X}_n \to E(X),$$

where $E(X)$ is the expected value of X. Here, convergence is in probability (weak law) or almost surely (strong law) (See also ►Laws of Large Numbers).

The Central Limit Theorem

This fundamental result explains why the normal distribution plays such a central role of statistics. Many of the statistics are sample averages and for their asymptotic distributions the following result holds. If $V(X)$ denotes the variance of X, then the asymptotic distribution of

$$\frac{\overline{X}_n - E(X)}{\sqrt{V(X)/n}}$$

is standard normal (see also ►Central Limit Theorems).

Subjective Probability

This interpretation of the concept of probability associates it with the strength of trust or belief that a person has in the occurrence of an event. Such beliefs manifest themselves, for example, in betting preferences: out of two events, a rational person would have a betting preference for the one with which he/she associates a larger subjective probability. A fundamental difference between frequentist and subjective probability is that the latter may also be applied to experiments and events that may not be repeated many times. Of course, the subjective probabilities of different individuals may be drastically different from each other and it has been demonstrated repeatedly that the subjective probabilities an individual associates with different events, may not be logically consistent. Bayesian statistics sometimes employs the elicitation of such subjective probabilities to construct a prior distribution.

About the Author

Professor Rudas was the Founding Dean of the Faculty of Social Sciences, Eötvös Loránd University, from 2003 to 2009. He is also an Affiliate Professor in the Department of Statistics, University of Washington, Seattle, and a Recurrent Visiting Professor in the Central European University, Budapest. He is Vice President, European Association of Methodology, 2008–. Professor Rudas has been awarded the Erdei Prize of the Hungarian Sociological Association for the application of statistical methods in sociology (1988), and the Golden Memorial Medal of the Eötvös Loránd University (2009).

Cross References

►Axioms of Probability
►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Central Limit Theorems
►Convergence of Random Variables
►Foundations of Probability
►Fuzzy Set Theory and Probability Theory: What is the Relationship?
►Laws of Large Numbers
►Limit Theorems of Probability Theory
►Measure Theory in Probability
►Philosophy of Probability
►Probability, History of
►Statistics and Gambling

References and Further Reading

Billingsley P (1995) Probability and measure, 3rd edn. Wiley, New York

Kolmogorov AN (1950) Foundations of the theory of probability. Chelsey, New York (Original work: Grundbegriffe der Wahrscheinlichkeits Rechnung, 1933, Berlin: Springer-Verlag)
Rudas T (ed) (2008) Handbook of probability: theory and applications. Sage, Thousand Oaks

Probability, History of

Jordi Vallverdú
Universitat Autònoma de Barcelona, Barcelona, Spain

Five thousand years ago dice were invented in India (David 1998). This fact implies that their users had at least a common sense approach to the idea of probability. Those dice were not the contemporary cubical standard dice, but fruit stones or animal bones (Dandoy 2006). They must surely have been used for fun and gambling as well as for fortune-telling practices. The worries about the future and the absurd idea that the world was causally guided by supernatural forces led those people to a belief in the explanatory power of rolling dice.

In fact, cosmogonical answers were the first attempt to explain in a causal way the existence of things and beings. The Greek creation myth involved a game of dice between Zeus, Poseidon, and Hades. Also in the classic Hindu book *Mahabharata* (section "Sabha-parva"), we can find the use of dice for gambling. But in both cases there is no theory regarding probability in dice, just their use "for fun."

Later, and beyond myths, Aristotle was the strongest defender of the causal and empirical approach to reality (*Physics*, II, 4–6) although he considered the possibility of chance, especially the problem of the game of dice (*On Heavens*, II, 292a30) and probabilities implied in it. These ideas had nothing to do with those about atomistic chance by Leucippus and Democritus nor Lucrecius' controversial *clinamen*'s theory. Hald (1988) affirms the existence of probabilistic rather than mathematical thought in Classical Antiquity; we can accept that some authors (like Aristotle) were worried about the idea of chance (as well as about the primordial emptiness and other types of conceptual *cul-de-sac*), but they made no formal analysis of it. Later, we can find traces of interest in the moral aspects of gambling with dice in Talmudic (*Babylonian Talmud, Book* 8*: Tract Sanhedrin, chapter* III*, Mishnas* I *to* III) and Rabbinical texts, and we know that in 960, Bishop Wibolf of Cambrai calculated 56 diverse ways of playing with three dice. *De Vetula*, a Latin poem from the thirteenth century, tells us of 216 possibilities. But the first occurrence of combinatorics per se arose from Chinese interest in future prediction through the 64 hexagrams of the *I Ching* (previously eight trigrams derived from four binary combinations of two elemental forces, *yin* and *yang*).

In 1494 Luca Paccioli defined the basic principles of algebra and multiplication tables up to 60 × 60 in his book *Summa de arithmetica, geometria, proportioni e proportionalita*. He posed the first serious statistical problem of two men playing a game called "balla," which is to end when one of them has won six rounds. However, when they stop playing *A* has only won five rounds and *B* three. How should they divide the wager? It would be another 200 years before this problem was solved.

In 1545 Girolamo Cardano wrote the books *Ars magna* (the great art) and *Liber de ludo aleae* (the book on games of chance). This was the first attempt to use mathematics to describe statistics and probability, and accurately described the probabilities of throwing various numbers with dice. Galileo expanded on this by calculating probabilities using two dice. At the same time the quantification of all aspects of daily life (art, music, time, space) between the years 1250 and 1600 made possible the numerical analysis of nature and, consequently, the discovery of the distribution of events and their rules (Crosby 1996).

It was finally Blaise Pascal who refined the theories of statistics and, with Pierre de Fermat, solved the "balla" problem of Paccioli (Devlin 2008). All these paved the way for modern statistics, which essentially began with the use of actuarial tables to determine insurance for merchant ships (Hacking 1984, 1990). Pascal was also the first to apply probability studies to the theory of decision (see his *Pensées*, 233), curiously, in the field of religious decisions. It is in this historical moment that the Latin term "probabilis" acquires its actual meaning, evolving from "worthy of approbation" to "numerical assessment of likelihood on a determined scale" (Moussy 2005).

In 1662, Antoine Arnauld and Pierre Nicole published the influential *La logique ou l'art de penser*, where we can find statistical probabilities. Games and their statistical roots worried people like Cardano, Pascal, Fermat, and Huygens (Weatherford 1982), although all of them were immersed in a strict mechanistic paradigm. Huygens is considered the first scientist interested in scientific probability, and in 1657 he published *De ratiotiniis in aleae ludo*. In 1708 Pierre Raymond de Montmort published his *Essay d'Analyse sur les Jeux de Hazard*, probably the first comprehensive text on probability theory. It was the next step after Pascal's work on combinatorics and its application to the solution of problems on games of chance. Later, De Moivre wrote the influential *De mensura sortis* (1711), and 78 years later, Laplace published

his *Philosophical Essay About Probability*. In the 1730s, Daniel Bernoulli (Jacob Bernoulli's nephew) developed the idea of utility as the mathematical combination of the quantity and perception of risk. Gottfried Leibniz at the beginning of the eighteenth century argued in several of his writings against the idea of chance, defending deterministic theories. According to him, chance was not part of the true nature of reality but the result of our incomplete knowledge. In this sense, probability is the estimation of facts that could be completely known and predicted, not the basic nature of things. Even morality was guided by natural laws, as Immanuel Kant argued in his *Foundations of the Metaphysics of Morals* (1785).

In 1763 an influential paper written by the Reverend Thomas Bayes was published posthumously. Richard Price, who was a friend of his, worked on the results of his efforts to find the solution to the problem of computing a distribution for the parameter of a ▶binomial distribution: *An Essay towards solving a Problem in the Doctrine of Chances*. Proposition 9 in the essay represented the main result of Bayes. Degrees of belief are therein considered as a basis for statistical practice. In a nutshell, Bayes proposed a theorem in which "probability" is defined as an index of subjective confidence, at the same time taking into account the relationships that exist within an array of simple and conditional probabilities. ▶Bayes' theorem is a tool for assessing how probable evidence can make a given hypothesis (Swinburne 2005). So, we can revise predictions in the light of relevant evidence and make a Bayesian inference, based on the assignment of some a priori distribution of a parameter under investigation (Stigler 1990). The classical formula of Bayes' rule is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

where our *posterior* belief $P(A|B)$ is calculated by multiplying our *prior* belief $P(A)$ by the *likelihood* $P(B|A)$ that B will occur if A is true. This classical version of Bayesianism had a long history, beginning with Bayes and continuing through Laplace to Jeffreys, Keynes, and Carnap in the twentieth century. Later, in the 1930s, a new type of Bayesianism appeared, the "subjective Bayesianism" of Ramsey and De Finetti (Ramsey 1931; de Finetti 1937; Savage 1954).

At the end of the nineteenth century, a lot of things were changing in the scientific and philosophical arena. The end of the idea of "causality" and the conflicts about observation lay at the heart of the debate. Gödel attacked Hilbert's axiomatic approach to mathematics and Bertrand Russell, as clever as ever, told us: "The law of causality (. . .) is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm (. . .) The principle "same cause, same effect," which philosophers imagine to be vital to science, is therefore utterly otiose" (Suppes 1970, p. 5). Nevertheless, scientists like Einstein were reluctant to accept the loss of determinism in favor of a purely random Universe; Einstein's words "God does not play dice" are the example of the difficulty of considering the whole world as a world of probabilities, with no inner intentionality, nor moral direction. On the other hand, scientists like Monod (*Chance and Necessity*, 1970) accepted this situation. In both cases, there is a deep consideration of the role of probability and chance in the construction of the philosophical and scientific meaning about reality.

In the 1920s there arose from theworks of Fisher (1922) and Neyman and Pearson (1928) the *classic* statistical paradigm: frequentism. They use the relative frequency concept, that is, you must perform one experiment many times and measure the proportion where you get a positive result. This proportion, if you perform the experiment enough times, is the probability. If Neyman and Pearson wrote their first joint paper and presented their approach as *one among alternatives*, Fisher, with his null hypothesis testing, gave a different message: his statistics was the formal solution of the problem of inductive inference (Gigerenzer 1990, p. 228).

From then on, these two main schools, Bayesian and Frequentist, were fighting each other to demonstrate that theirs was the superior and only valid approach (Vallverdú 2008).

Finally, with the advent of the information era and all the (super)computer scientific simulations, Bayesianism has again achieved a higher status inside the community of experts on probability. Bayesian inference also allows intelligent and real-time monitoring of computational clusters, and its application in belief networks has proved to be a good technique for diagnosis, forecasting, and decision analysis tasks. This fact has contributed to the increasing application of parallel techniques for large Bayesian networks in expert systems (automated causal discovery, AI…) (Korb and Nicholson 2003).

Acknowledgments

This research was supported by the project "El diseño del espacio en entornos de cognición distribuida: plantillas y affordances," MCI [FFI2008-01559/FISO].

About the Author

Jordi Vallverdú is a lecturer professor of philosophy in science and computing at Universitat Autonoma de Barcelona. He holds a Ph.D. in philosophy of science (UAB) and a master's degree in history of sciences (UAB).

After a short research stay as a fellowship researcher at Glaxo-Wellcome Institute for the History of Medicine, London (1997), and a research assistant of Dr. Jasanoff at J.F.K. School of Government, Harvard University (2000), he worked in computing epistemology issues and bioethic and synthetic emotions. He is listed as an EU Biosociety Research Expert and is a member of the E-CAP' Steering (http://www.ia-cap.org/administration.php). He leads a research group, SETE (Synthetic Emotions in Technological Environments), which has published about computational models of synthetic emotions and their implementation into social robotic systems. He is Editor-in-Chief of the *International Journal of Synthetic Emotions* (IJSE) and has edited (and written as an included author) the following books: *Handbook of research on synthetic emotions and sociable robotics: new applications in affective computing and artificial intelligence* (with D. Casacuberta, Information Science Publishing, 2009) and *Thinking machines and the philosophy of computer science: concepts and principles* (Ed., 2010).

Cross References

►Actuarial Methods
►Bayes' Theorem
►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Foundations of Probability
►Philosophy of Probability
►Probability Theory: An Outline
►Statistics and Gambling
►Statistics, History of

References and Further Reading

Crosby AW (1996) The measure of reality: quantification in Western Europe 1250–1600. Cambridge University Press, Cambridge
Dandoy JR (2006) Astragali through time. In: Maltby M (ed) Integrating zooarchaeology. Oxbow Books, Oxford, pp 131–137
David FN (1998) Games, gods and gambling, a history of probability and statistical ideas. Dover, Mineola
De Finetti B (1937) La Prevision: Ses Lois Logiques, Ses Sources Subjectives. Annals de l'Institute Henri Poincaré 7:1–68
Devlin K (2008) The unfinished game: Pascal, Fermat, and the seventeenth-century letter that made the world modern. Basic Books, New York
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos trans Roy Soc London Ser A 222:309–368
Gigerenzer G et al (1990) The Empire of chance. How probability changed science and everyday life. Cambridge University Press, Cambridge
Hacking I (1984) The emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference. Cambridge University Press, Cambridge
Hacking I (1990) The taming of chance. Cambridge University Press, Cambridge
Hald A (1988) A history of probability and statistics and their applications before 1750. Wiley, New York
Korb KB, Nicholson AE (2003) Bayesian artificial intelligence. CRC Press, Boca Raton
Moussy C (2005) Probare, probatio, probabilis' dans de vocabulaire de la démonstration. Pallas 69:31–42
Neymanv J, Pearson ES (1928) On the use and interpretation certain test criteria for purposes of statistical inferences. Biometrika 20(A):175–240, 263–294
Ramsey FP (1931) Truth and probability (1926). In: Braithwaite RB (ed) The foundations of mathematics and other logical essays, Ch. VII. Kegan Paul, Trench, Trubner/Harcourt, Brace, London/New York, pp 156–198
Savage LJ (1954) The foundations of statistics. Wiley, New York
Stigler SM (1990) The history of statistics: the measurement of uncertainty before 1900. Harvard University Press, Cambridge
Suppes P (1970) A probabilistic theory of causality. North-Holland, Helsinki
Swinburne R (ed) (2005) Bayes's theorem. In: Proceedings of the British academy, vol 113. Oxford University Press, London
Vallverdú J (2008) The false Dilemma: Bayesian vs. Frequentist. E – LOGOS Electron J Philos 1–17. http://e-logos.vse.cz/index.php?target=indexyear
Weatherford R (1982) Philosophical foundations of probability theory. Routledge & Kegan Paul, London

Probit Analysis

Tiberiu Postelnicu
Professor Emeritus
Romanian Academy, Bucharest, Romania

Introduction

The idea of probit analysis was originally published in *Science* by Chester Ittner Bliss (1899–1979) in 1934. He was primarily concerned with finding an effective pesticide to control insects that fed on grape leaves. By plotting the response of the insects to various concentrations of pesticides, he could visually see that each pesticide affected the insects at different concentrations, but he did not have a statistical method to compare this difference. The most logical approach would be to fit a regression of the response versus the concentration or dose and compare between the different pesticides. The relationship of response to dose was sigmoid in nature and at that time regression was only used on linear data. Therefore, Bliss developed the idea of transforming the sigmoid dose–response curve to a straight line. When biological responses are plotted against their causal stimuli (or their logarithms) they often form a

sigmoid curve. Sigmoid relationships can be linearized by transformations such as logit, probit, and angular. For most systems the probit (normal sigmoid) and logit (logistic sigmoid) give the most closely fitting result. Logistic methods are useful in epidemiology, but in biological assay work, probit analysis is preferred. David Finney, from the University of Edinburgh, took Bliss' idea and wrote a book entitled *Probit Analysis* in 1947, and since this year it is still the preferred statistical method in understanding dose–response relationships.

Probit

In probability theory and statistics, the *probit function* is the inverse cumulative distribution function (CDF), or *quantile function* associated with the standard normal distribution. It has applications in exploratory statistical graphics and specialized regression modeling of binary response variables. For the standard normal distribution, the CDF is commonly denoted $\Phi(z)$, which is a continuous, monotone-increasing sigmoid function whose domain is the real line and range is $(0,1)$. The probit function gives the inverse computation, generating a value of an $N(0,1)$ random variable, associated with specified cumulative probability. Formally, the probit function is the inverse of $\Phi(z)$, denoted $\Phi^{-1}(p)$. In general, we have $\Phi(\text{probit}(p)) = p$ and $\text{probit}(\Phi(z)) = z$. Bliss proposed transforming the percentage into "*probability unit*" (or "*probit*") and included a table to aid researchers to convert their kill percentages to his probit, which they could then plot against the logarithm of the dose and thereby, it was hoped, obtain a more or less straight line. Such a probit model is still important in toxicology, as well as in other fields. It should be observed that probit methodology, including numerical optimization for fitting of probit functions, was introduced before widespread availability of electronic computing and, therefore, it was convenient to have probits uniformly positive.

Related Topics

The probit function is useful in statistical analysis for diagnosing deviation from normality, according to the method of Q–Q plotting. If a set of data is actually a sample of a normal distribution, a plot of the values against their probit scores will be approximately linear. Specific deviation from normality such as asymmetry, heavy tails, or bimodality can be diagnosed based on the detection of specific deviations from linearity. While the $Q-Q$ plot can be used for comparison with any distribution family (not only the normal), the normal Q–Q plot is a relatively standard exploratory data analysis procedure because the assumption of normality is often a starting point for analysis.

The normal distribution CDF and its inverse are not available in closed form, and computation requires careful use of numerical procedures. However, the functions are widely available in software for statistics and probability modeling, and also in spreadsheets. In computing environments where numerical implementations of the inverse error function are available, the probit function may be obtained as $\text{probit}(p) = \sqrt{2}\text{erf}^{-1}(2p-1)$. An example is MATLAB, where an "erfinv" function is available and the language MATHEMATICA implements "InverseErf." Other environments directly implement the probit function in the R programming language.

Closely related to the probit function is the *logit function* using the "odds" $p/(1-p)$, where p is the proportional response, i.e., r out of n responded, so there is $p = r/n$ and $\text{logit}(p) = \text{log odds} = \log(p/(1-p))$. Analogously to the probit model, it is possible to assume that such a quantity is related linearly to a set of predictors, resulting in the logit model, the basis in particular of logistic regression model (see ►Logistic Regression), the most prevalent form of regression analysis for binary response data. In current statistical practice, probit and logit regression models are often handled as cases of the generalized linear model (see ►Generalized Linear Models).

Probit Model

In statistics and related fields, a *probit model* is a specification for a binary response model that employs a probit link function. This model is most often estimated using the standard maximum likelihood procedure; such an estimation is called *probit regression*. A fast method for computing maximum likelihood estimates for probit models was proposed by Ronald Fisher in an Appendix to the article of Bliss in 1935.

Probit analysis is a method of analyzing the relationship between a stimulus (dose) and the quantal (all or nothing) response. Quantitative responses are almost always preferred, but in many situations they are not practical. In these cases, it is only possible to determine if a certain response has occurred. In a typical quantal response experiment, groups of animals are given different doses of a drug. The percent dying at each dose level is recorded. These data may then be analyzed using probit analysis. StatPlus includes two different methods of probit analysis, but the Finney method is the most important and useful. The probit model assumes that the percent response is related to the log dose as the cumulative normal distribution, that is, the log doses may be used as variables to read the percent dying from the cumulative normal. Using the

normal distribution, rather than other probability distributions, influences the predicted response rate at the high and low ends of possible doses, but has little influence near the middle. Much of the comparison of different drugs is done using response rates of 50%. The probit model may be expressed mathematically as follows:

$$P = a + b(\log(\text{Dose})),$$

where P is five plus the inverse normal transform of the response rate (called the probit). The five is added to reduce the possibility of negative probits, a situation that caused confusion when solving the problem by hand.

Suppose the response variable Y is *binary*, that is, it can have only two possible outcomes, which we will denote as 1 and 0. For example, Y may represent presence/absence of a certain condition, success/failure of some device, and answer yes/no on a survey. We also have a vector of regression X, which are assumed to influence the outcome Y. Specifically, we assume that the model takes the form

$$P[Y = 1|X] = \Phi(X'\beta),$$

where P is the probability and Φ is the probit function – the CDF of the standard normal distribution. The parameters β are typically estimated by maximum likelihood. For more complex probit analysis, such as the calculation of relative potencies from several related dose–response curves, consider nonlinear optimization software or specialist dose–response analysis software. The latter is a FORTRAN routine written by David Finney and Ian Craigie from Edinburgh University Computing Center. MLP or GENSTAT can be used for a more general nonlinear model fitting. We must take into account that the standard probit analysis is designed to handle only quantal responses with binomial error distributions. Quantal data, such as the number of subjects responding versus the total number of subjects tested, usually have binomial error distributions. We should not use continuous data, such as percent maximal response, with probit analysis as these data are likely to require regression methods that assume a different error distribution.

Applications

Probit analysis is used to analyze many kinds of dose–response or binomial response experiments in a variety of fields. It is commonly used in toxicology to determine the relative toxicity of chemicals to living organisms. This is done by testing the response of an organism under various concentrations of each of the chemicals in question and then comparing the concentrations at which one encounters a response. The response is always binomial and the relationship between the response and the various concentrations is always sigmoid. Probit analysis acts as a transformation from sigmoid to linear and then runs a regression on the relationship. Once the regression is run, we can use the output of the probit analysis to compare the amount of chemical required to create the same response in each of the various chemicals. There are many points used to compare the differing toxicities of chemicals, but the LC50 (liquids) or LD50 (solids) are the most widely used outcomes of the modern dose–response experiments. The LC50/LD50 represent the concentration (LC50) or dose (LD50) at which 50% of the population responds. It is possible to use probit analysis with various methods such as statistical packages SPSS, SAS, R, or S, but it is good to see the history of the methodology to get a thorough understanding of the material. We must take care that probit analysis assumes that the relationship between number responding (non-percent response) and concentration is normally distributed; if not, logit is preferred.

The properties of the estimates given by probit analysis have been studied also by Ola Hertzberg (1974). The up-and-down technique is the best known among staircase methods for estimating the parameters in quantal response curves (QRC). Some small sample properties of probit analysis are considered and in the estimate distribution the medians are used as a measure of location.

About the Author

Tiberiu Postelnicu (born on June 15, 1930, Campina), received his Ph.D. in Mathematics, University of Bucharest, in 1957. He was Head of Department of Biostatistics, Carol Davila University of Medicine and Pharmacy, Bucharest, and Department of Biometrics, Centre of Mathematical Statistics of the Romanian Academy, Bucharest. He is a member of the International Statistical Institute, Biometric Society, Bernoulli Society for Mathematical Statistics and Probability, Italian Society for Statistics, and the New York Academy of Science. Dr. Postelnicu was a member of the editorial boards of the *Biometrical Journal* and *Journal for Statistical Planning and Inference*. He was awarded the Gheorghe Lazar Prize for Mathematics, Romanian Academy (1972). Currently, Professor Postelnicu is President of the Commission for Biometrics of the Romanian Academy.

Cross References

►Agriculture, Statistics in
►Econometrics
►Generalized Linear Models
►Logistic Regression

►Nonlinear Models
►Normal Distribution, Univariate

References and Further Reading

Bliss CI (1934) The method of probit. Science 79(2037):38–39
Bliss CI (1935) The calculation of the dosage–mortality curve. Ann Appl Biol 22:1, 134–167
Bliss CI (1938) The determination of the dosage-mortality curve from small numbers. Q J Pharmacol 11:192–216
Collett D (1991) Modelling binary data. Chapman & Hall, London
Finney DJ (1947) Probit analysis, 1st edn. Cambridge University Press, Cambridge
Finney DJ, Stevens WL (1948) A table for the calculation of working probits and weights in probit analysis. Biometrika 35(1–2): 191–201
Finney DJ (1971) Probit analysis, 3rd edn. Cambridge University Press, Cambridge
Greenberg BG (1980) Chester I Bliss, 1899–1979. Int Stat Rev 8(1):135–136
Hertzberg JO (1974) On small sample properties of probit analysis. In: Proceedings of the 8th International Biometric Conference. Constantza, Romania, pp 153–162
McCullagh P, Nelder J (1989) Generalized linear models. Chapman & Hall, London

Promoting, Fostering and Development of Statistics in Developing Countries

Noël H. Fonton[1], Norbert Hounkonnou[2]
[1]Professor, Head of Centre of Biostatistics and Director Laboratory of Modeling of Biological Phenomena University of Abomey-Calavi, Cotonou, Benin Republic
[2]Professor, Chairman of International Chair of Mathematical Physics and Applications (ICMPA UNESCO Chair) Cotonou, Benin Republic

Statistical methods are universal and hence their applicability depends neither on geographical area nor on a people's culture. Promoting and increasing the use of statistics in developing countries can help to find solutions to the needs of their citizens. Developing countries are confronted with endemic poverty that requires implementable solutions for alleviating suffering. Such poverty is a signal call to the world to meet fundamental human needs—food, adequate shelter, access to education and healthcare, protection from violence, and freedom. Statistics and statistical tools, the matching between hypothesis, data collection, and statistical method, are necessary as development strategies in the developing counties are formulated to address these needs.

First, a reliable basis for the implementation of strategies against poverty and achievement of the Millennium Development Goals require good statistics, an essential element of good governance. Therefore, important indicators to inform and monitor development policies are often derived from household surveys, which have become a dominant form of data collection in developing countries. Such surveys are an important source of socioeconomic data. Azouvi (2001) has proposed a low-cost statistical program in four areas: statistical coordination, national accounts, economic and social conjuncture, and dissemination. To increase awareness that good statistics are important for achieving better development results, the Marrakech Action Plan for Statistics (MPS) was developed in 2004. This global plan for improving development statistics was agreed upon at a second round-table for best managing development results (World Bank, 2004). The idea is that better data are needed for better results in order to improve development statistics. One indicator to ensure the application of this MPS is the full participation of developing countries in the 2010 census round. Additionally, funds allocated by the World Bank from the Development Grant Facility and technical assistance from national universities are critical.

Second, national capacity-building in statistics is very limited. Even in developed countries, secondary school students, together with their teachers, rarely see the applicability and the challenge of statistical thinking (Boland, 2002). This situation exists to a greater degree in the university training systems of many developing countries. It is suggested that statistical programs should be reviewed and executed by statisticians and examples of a local nature should be used. This is possible with sufficient number of statisticians in various fields. According to Lo (2009), there are 5 to 10 holders of doctoral degrees in statistics per country, with higher numbers in some countries. There is a low number of statisticians in developing countries due to the lack of master's degree programs in statistics. In sub-Saharan, French-speaking African countries, the "Statistiques pour l'Afrique Francophone et Applications au vivant" (STAFAV) project is being implemented in three parts: a network for master's training in applied statistics at Cotonou (Benin), Saint-Louis (Senegal), and Yaounde (Cameroon); some PhD candidates jointly supervised by scientists from African universities and French universities; and the development of statistical research through an African Network of Mathematical Statistics and Applications (RASMA). STAFAV constitutes a good means for increasing statistical capacity in

developing countries. With the launch of the Statistical Pan African Society (SPAS), more visibility for research and the use of statistics and probability may be achieved via the African Visibility Program in Statistics and Probability. This society is an antidote to the very isolated work of statistics researchers and allows researchers to share experiences and scientific work.

Third, statistics are a tool for solving the problems of developing countries, but the development of research activities is a challenge. Many of these countries are located in tropical areas; therefore, there are major differences between them and other countries due to high biological variability and the probability distribution of the studied phenomena is frequently misunderstood. Control of biological variability requires wide use of probability theory. Because of the disparate populations and subpopulations in the experimental data, the use of one probability distribution should be called into question. Lacking sophisticated statistical methods, development of statistical methods of mix-distributions becomes a challenge. Economic loss, agriculture, water policies, and health (malaria, HIV/AIDS, and recently H1N1) are the major areas for research programs in which statistics have a major role to play. Development of statistical research programs directed at the well-being of local people is necessary.

The sustainability of statistical development is another important issue in the field. Graduate statisticians, when returning to their native countries, often do not have facilities for continuing education, documentation, or statistics software packages. To foster statistics in developing countries, national statistics institutes, universities, and research centers need to increase funds allocated for subscribing to statistical journals, mainly online, and software, with a staff properly trained on those fields. Statisticians must be offered opportunities to attend annual conferences and to do research and/or professional training in a statistical institute outside of their countries.

Contributions of statisticians from developed countries, working or teaching in developing countries, are welcome. Specifically, they can join research teams in developing countries, share experiences with them, help acquire funding, and teach.

Cross References

►African Population Censuses
►Careers in Statistics
►Learning Statistics in a Foreign Language
►National Account Statistics
►Online Statistics Education
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Role of Statistics: Developing Country Perspective
►Selection of Appropriate Statistical Methods in Developing Countries
►Statistics and Climate Change
►Statistics Education

References and Further Reading

Azouvi A (2001) Proposals for a minimum programme for statistics in developing countries. Int Stat Rev 69(2):333–343
Boland PJ (2002) Promoting statistics thinking amongst secondary school students in national context. ICOTS6, p 6
Lo GS (2009) Probability and statistics in Africa. IMS Bull 38(7):8
World Bank (2004) The marrakech action plan for statistics. Second international roundtable on managing for development results. Morocco, p 19

Properties of Estimators

Paul H. Garthwaite
Professor of Statistics
The Open University, Milton Keynes, UK

Estimation is a primary task of statistics and estimators play many roles. Interval estimators, such as confidence intervals or prediction intervals, aim to give a range of plausible values for an unknown quantity. Density estimators aim to approximate a probability distribution. These and other varied roles of estimators are discussed in other sections. Here attention is restricted to point estimation, where the aim is to calculate from data a single value that is a good estimate of an unknown parameter.

We will denote the unknown parameter by θ, which is assumed to be a scalar. In the standard situation there is a statistic T whose value, t, is determined by sample data. T is a random variable and it is referred to as a (point) estimator of θ if t is an estimate of θ. Usually there will be a variety of possible estimators so criteria are needed to separate good estimators from poor ones. There are a number of desirable properties which we would like estimators to possess, though a property will not necessarily identify a unique "best" estimator and rarely will there be an estimator that has all the properties mentioned here. Also, caution must be exercised in using the properties as a reasonable property will occasionally lead to an estimator that is unreasonable.

One property that is generally useful is unbiasedness. T is an unbiased estimator of θ if, for any θ, $E(T) = \theta$. Thus T is unbiased if, on average, it tends neither to be bigger nor smaller than the quantity it estimates, regardless of

the actual value of the quantity. The bias of T is defined to be $E(T) - \theta$. Obviously a parameter can have more than one unbiased estimator. For example, if θ is the mean of a symmetric distribution from which a random sample is taken, then T is an unbiased estimator if it is the mean, median or mid-range of the sample. It is also the case that sometimes a unique unbiased estimator is not sensible. For example, Cox and Hinkley (1974, p. 253) show that if a single observation is taken from a geometric distribution with parameter θ, then there is only one unbiased estimator and its estimate of θ is either 1 (if the observation's value is 1) or 0 (if the observation's value is greater than 1). In most circumstances these are not good estimates.

It is desirable, almost by definition, that the estimate t should be close to θ. Hence the quality of an estimator might be judged by its expected absolute error, $E(|T - \theta|)$, or its mean squared error, $E[(T - \theta)^2]$. The latter is used far more commonly, partly because of its relationship to the mean and variance of T:

$$\text{mean squared error} = \text{variance} + (\text{bias})^2. \qquad (1)$$

If the aim is to find an estimator with small mean squared error (MSE), clearly unbiasedness is desirable, as then the last term in Eq. (1) vanishes. However, unbiasedness is not essential and trading a small amount of bias for a large reduction in variance will reduce the MSE. Perhaps the best known biased estimators are the regression coefficients given by ridge regression (see ►Ridge and Surrogate Ridge Regressions), which handles multicollinearities in a regression problem by allowing a small amount of bias in the coefficient estimates, thereby reducing the variance of the estimates.

It may seem natural to try to find estimators which minimize MSE, but this is often difficult to do. Moreover, given any estimator, there is usually some value of θ for which that estimator's MSE is greater than the MSE of some other estimator. Hence the existence of an estimator with a *uniformly* minimum MSE is generally in doubt. For example, consider the trivial and rather stupid estimator that ignores the data and chooses some constant θ_0 as the estimator of θ. Should θ actually equal θ_0, then this estimator has an MSE of 0 and other estimators will seldom match it. Thus other estimators will not have a uniformly smaller MSE than this trivial estimator.

Restricting attention to unbiased estimators solves many of the difficulties of working with MSE. The task of minimizing MSE reduces to that of minimizing variance and substantial theory has been developed about minimum variance unbiased estimators (MVUEs). This includes two well-known results, the *Cramér–Rao lower bound* and the ►*Rao–Blackwell theorem.* The Cramér–Rao lower bound is I_θ^{-1}, where I_θ is the Fisher information about θ. (I_θ is determined from the likelihood for θ.) Subject to certain regularity conditions, the Cramér–Rao lower bound is a lower bound to the variance of any unbiased estimator of θ.

A benefit of the Cramér–Rao lower bound is that it provides a numerical scale-free measure for judging an estimator: the *efficiency* of an unbiased estimator is defined as the ratio of the Cramér–Rao lower bound to the variance of the estimator. Also, an unbiased estimator is said to have the property of being efficient if its variance equals the Cramér–Rao lower bound. Efficient estimators are not uncommon. For example, the sample mean is an efficient estimator of the population mean when sampling is from a normal distribution or a Poisson distribution, and there are many others. By definition, only an MVUE might be efficient.

Sufficiency is a property of a statistic that can lead to good estimators. A statistic S (which may be a vector) is sufficient for θ if it captures all the information about θ that the data contain. For example, the sample variance is sufficient for the population variance when data are a random sample from a normal distribution – hence to make inferences about the population variance we only need to know the sample variance and not the individual data values. The definition of sufficiency is a little more transparent in ►Bayesian statistics than in classical statistics (though the definitions are equivalent). In the Bayesian approach, S is sufficient for θ if the distribution of θ, given the value of S, is the same as θ's distribution given all the data. i.e. $g(\theta \mid S) = g(\theta \mid \text{data})$, where $g(\,.\,)$ is the p.d.f. of θ. In the classical definition (where θ cannot be considered to have a distribution), S is sufficient for θ if the conditional distribution of the data, given the value of S, does not depend on θ. A sufficient statistic may contain much superfluous information along with the information about θ, so the concept of a *minimal sufficient statistic* is also useful. A statistic is minimal sufficient if it can be expressed as a function of every other sufficient statistic.

The Rao–Blackwell theorem shows the importance of sufficient statistics when seeking unbiased estimators with small variance. It states that if $\widehat{\theta}$ is an unbiased estimator of θ and S is a sufficient statistic, then

1. $T_S = E(\widehat{\theta} \mid S)$ is a function of S alone and is an unbiased estimator of θ.
2. $\text{Var}(T_S) \le \text{var}(\widehat{\theta})$.

The theorem means that we can try to improve on any unbiased estimator by taking its expectation conditional on a sufficient statistic – the resulting estimator will also be unbiased and its variance will be smaller than, or equal

to, the variance of the original estimator. Stronger results hold if a minimal sufficient statistic is also *complete*: S is complete if $E[h(S)]$ cannot equal 0 for all θ unless $h(S) \equiv 0$ almost everywhere (where h is any function). If S is complete there is at most one function of S that is an unbiased estimator of θ. Suppose now, that S is a complete minimal sufficient statistic for θ. An important result is that if $h(S)$ is an unbiased estimator of θ, then $h(S)$ is an MVUE for θ, if an MVUE exists. The consequence is that, when searching for an MVUE, attention can be confined to functions of a complete sufficient statistic.

Turning to asymptotic properties, suppose that data consist of a ►simple random sample of size n and consider the behavior of an estimator T as $n \to \infty$. An almost essential property is that the estimator should be consistent: T is a consistent estimator of θ if T converges to θ in probability as $n \to \infty$. Consistency implies that, as the sample size increases, any bias in T tends to 0 and the variance of T also tends to 0.

Two useful properties that do not relate directly to the accuracy of an estimator are ►*asymptotic normality* and *invariance*. When sample sizes are large, confidence intervals and hypothesis tests are often based on the assumption that the distribution of an estimator is approximately normal. Hence asymptotic normality is a useful property in an estimator, especially if approximate normality holds quite well for modest sample sizes. Invariance of estimators relates to the method of forming them. It is the notion that if we take a transformation of a parameter, then ideally its estimator should transform in the same way. For example, let $\phi = g(\theta)$, where ϕ is a one-to-one function of θ. Then if a method of forming estimators gives t_1 and t_2 as estimates of ϕ and θ, invariance would imply that t_1 necessarily equalled $g(t_2)$. Maximum likelihood estimators are invariant.

We have assumed that the unknown parameter (θ) is a scalar. Concepts such as unbiasedness, sufficiency, consistency, invariance and asymptotic normality extend very naturally to the case where the unknown parameter is a vector. If θ is a vector but an estimate of just one of its components is required, then a vector-form of the Cramer–Rao lower bound yields a minimum variance for any unbiased estimator of the component. Simultaneous estimation of more than one component, however, raises new challenges unless estimating each component separately and combining the estimates is optimal.

While a search for MVUEs has been a focus of one area of statistics, other branches of statistics want different properties in estimators. Robust methods want point estimators that are comparatively insensitive to a few aberrant observations or the odd outlier. Nonparametric methods want to estimate a population mean or variance, say, without making strong assumptions about the population distribution. These branches of statistics do not place paramount importance on unbiasedness or minimal variance, but they nevertheless typically seek estimators with low bias and small variance – it is just that their estimators must also satisfy other requirements. In contrast, Bayesian statistics uses a markedly different framework for choosing estimators. In its basic form the parameters of the sampling model are given a prior distribution, while a loss function specifies the penalty for inaccuracy in estimating a parameter. The task is then to select an estimator or decision rule that will minimize the expected loss, so minimizing expected loss is the property of dominant importance.

Many other sections of this encyclopedia also consider point estimators and point estimation methods. These include sections on nonparametrics, robust estimation, Bayesian methods and decision theory. The focus in this section has been the classical properties of point estimators. Deeper discussion of this topic and proofs of results are given in most advanced textbooks on statistical inference or theoretical statistics, such as Bickel and Doksum (2000), Cox and Hinkley (1974), Garthwaite et al. (2002), and Lehmann and Casella (1998).

About the Author

Paul Garthwaite is Professor of Statistics at the Open University, UK, where he was Head of the Department of Statistics from 2001–2004 and again in 2006. He worked for twenty years in the University of Aberdeen and has held visiting positions at universities in New York State, Minnesota, Brisbane and Sydney. In 1983 he was awarded the L J Savage Prize, a prize now under the auspices of the International Society for Bayesian Analysis. He has published 80 journal papers and is co-author of two books, *Statistical Inference* (Oxford University Press, 2002) and *Uncertain Judgements: Eliciting Expert' Probabilities* (Willey, 2006).

Cross References

►Approximations for Densities of Sufficient Estimators
►Asymptotic Normality
►Asymptotic Relative Efficiency in Estimation
►Bayesian Statistics
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Cramér–Rao Inequality
►Estimation
►Estimation: An Overview
►Minimum Variance Unbiased
►Rao–Blackwell Theorem

►Sufficient Statistics
►Unbiased Estimators and Their Applications

References and Further Reading

Bickel PJ, Doksum KA (2000) Mathematical statistics: basic ideas and selected topics, 2nd edn. Prentice Hall, London
Cox DR, Hinkley DV (1974) Theoretical statistics. Wiley, New York
Garthwaite PH, Jolliffe IT, Jones B (2002) Statistical inference, 2nd edn. Oxford University Press, Oxford
Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer, New York

Proportions, Inferences, and Comparisons

George A. F. Seber
Emeritus Professor of Statistics
Auckland University, Auckland, New Zealand

A common problem in statistics, and especially in sample surveys, is how to estimate the proportion p $(= 1 - q)$ of people with a given characteristic (e.g., being left-handed) in a population of known size N. If there are M left-handed people in the population, then $p = M/N$. The usual method of estimating p is to take a ►simple random sample (SRS), that is, a random sample without replacement, of size n and count the number of left-handed people, x, in the sample. If the sample is representative, then p can be estimated by the sample proportion $\hat{p} = x/n$. To make inferences about p we need to use the probability distribution of x, namely the Hypergeometric distribution (see ►Hypergeometric Distribution and Its Application in Statistics), a distribution that is difficult to use. From this distribution we can get the mean and variance of x and hence of $\hat{p}$, namely

$$\mu_{\hat{p}} = p \quad \text{and} \quad \sigma^2_{\hat{p}} = r\frac{pq}{n},$$

where $r = (N - n)/(N - 1) \approx 1 - f$, and f is the sampling fraction n/N, which can be ignored if it is less that 0.1 (or better 0.05). Fortunately, for sufficiently large N, M and n, x and $\hat{p}$ are approximately normal so that $z = (\hat{p} - p)/\sigma_{\hat{p}}$ has an approximate standard normal distribution (with mean 0 and variance 1). This approximation will still hold if we replace p by $\hat{p}$ in the denominator of $\sigma_{\hat{p}}$ to get $\hat{\sigma}_{\hat{p}}$ giving us an approximate 95% confidence interval $\hat{p} \pm 1.96\hat{\sigma}_{\hat{p}}$.

Inverse sampling can also be used to estimate p, especially when the characteristic is rare. Random sampling is continued until x units of the given characteristic are selected, n now being random, and Haldane in 1945 gave the estimate $\hat{p}_I = (x - 1)/(n - 1)$. This has variance estimate $\hat{\sigma}^2_{\hat{p}_I} = r_I(\hat{p}_I\hat{q}_I)/(n - 2)$, where $r_I = 1 - (n - 1)/N$ for without replacement and $r_I = 1$ for with replacement, and an approximate 95% confidence interval for p is $\hat{p}_I \pm 1.96\hat{\sigma}_{\hat{p}_I}$ (Salehi and Seber 2001).

Another application of this theory is in the case where M consists of a known number of marked animals released into a population of unknown size, but with f known to be sufficiently small so that we can set $r = 1$. The confidence interval for p can then be rearranged to give a confidence interval for N. This simple idea has lead to a very large literature on capture-recapture (Seber 2002).

Returning to our example relating to left-handed people, when we choose the first person from the population, the probability of getting a left-handed person will be p so that the terms "probability" and "proportion" tend to be used interchangeably in the literature, although they are distinct concepts. They can be brought even closer together if sampling is with replacement for then the probability of getting a left-handed person at each selection remains at p and x now has a ►Binomial distribution, as we have n independent trials with probability of "success" being p. The above formulas for means and variances and confidence interval are still the same except that r is now exactly 1. This is not surprising as we expect sampling with replacement to be a good approximation for sampling without replacement when a small proportion of a population is sampled. Confidence intervals for the Binomial distribution have been studied for many years and a variety of approximations and modifications have been considered, for example, Newcombe (1998a). "Exact" confidence intervals, usually referred to as the Clopper–Pearson intervals, can also be computed using the so-called "tail" probabilities of the binomial distribution, which are related to a ►Beta distribution (cf. Agresti and Coull 1998). We can also use the above theory to test null hypotheses like H_0: $p = p_0$, though such hypotheses apply more to probabilities than proportions.

When it comes to comparing two proportions, there are three different experimental situations that need to be considered. Our example for explaining these relates to voting preferences. Suppose we wish to compare the proportions, say p_i $(i = 1, 2)$, of people voting for a particular candidate in two separate areas and we do so by taking an SRS of size n_i in each area and computing $\hat{p}_i$ for each area. In comparing the areas we will be interested in estimating $\theta = p_1 - p_2$ using $\hat{\theta} = \hat{p}_1 - \hat{p}_2$. As the two estimates are statistically independent, and assuming f can be neglected for each sample, we have

$$\sigma^2_{\hat{\theta}} = \frac{p_1q_1}{n_1} + \frac{p_1q_2}{n_2},$$

which we can estimate by replacing each p_i by its estimate. Assuming the normal approximation is valid for each sample, we have the usual approximate 95% confidence interval $\hat{\theta} \pm 1.96\hat{\sigma}_{\hat{\theta}}$ for θ. This theory also applies to comparing two Binomial probabilities and, as with a single probability, a number of methods have been proposed (see Newcombe 1998b). Several procedures for testing the null hypothesis H_0: $\theta = 0$ are available including an "exact" test due to Fisher and an approximate Chi-square test with or without a correction for continuity.

A second situation is when we have a single population and the p_is now refer to two different candidates. The estimates $\hat{p}_i$ $(i = 1, 2)$ are no longer independent and we now find, from Scott and Seber (1983), that

$$\sigma^2_{\hat{\theta}} = \frac{1}{n}[p_1 + p_2 - (p_1 - p_2)^2],$$

where n is the sample size. Once again we can replace the unknown p_i by their estimates and, assuming f can be ignored, we can obtain an approximate confidence interval for θ, as before.

A third situation occurs when two proportions from the same populations overlap in some way. Suppose we carry out a sample survey ▶questionnaire of n questions that have answers "Yes" and "No." Considering the first two questions, Q_1 and Q_2, let p_{11} be the proportion of people who say "Yes" to both questions, p_{12} the proportion who say "Yes" to Q_1 and "No" to Q_2, p_{21} the proportion who say "No" to Q_1 and "Yes" to Q_2, and p_{22} the proportion who say "No" to both questions. Then $p_1 = p_{11}+p_{12}$ is the proportion saying "Yes " to Q_1 and $p_2 = p_{11} + p_{21}$ the proportion saying "Yes" to Q_2. We want to estimate $\theta = p_1 - p_2$, as before. If x_{ij} are observed in the category with probability p_{ij} and $x_1 = x_{11} + x_{12}$ and $x_2 = x_{11} + x_{21}$, then, from Wild and Seber (1993),

$$\sigma^2_{\hat{\theta}} = \frac{1}{n}[p_{12} + p_{21} - (p_{12} - p_{21})^2].$$

To estimate the above variance, we would replace each p_{ij} by its estimate $\hat{p}_{ij}$. In many surveys though, only x_1 and x_2 are recorded so that the only parameters we can estimate are the p_i using $\hat{p}_i = x_i/n$. However, we can use these estimates in the following bounds

$$\frac{1}{n}d(1-d) \le \sigma^2_{\hat{\theta}} \le \frac{1}{n}[min(p_1 + p_2, q_1 + q_2) - (p_1 - p_2)^2],$$

where $d = |p_{12} - p_{21}| = |p_1 - p_2|$. Further comments about constructing confidence intervals, testing hypotheses, and dealing with non-responses to the questions are given in the above paper.

For an elementary discussion of some of the above ideas see Wild and Seber (2000).

About the Author

For biography *see* the entry ▶Adaptive Sampling.

Cross References

▶Asymptotic Normality
▶Binomial Distribution
▶Fisher Exact Test
▶Hypergeometric Distribution and Its Application in Statistics
▶Inverse Sampling
▶Statistical Inference
▶Statistical Inference in Ecology
▶Statistical Inference: An Overview

References and Further Reading

Agresti A, Coull BA (1998) Approximate is better than 'exact' for interval estimation of binomial proportions. Am Stat 52:119–126
Brown LD, Cai TT, DasGupta A (2001) Interval estimation for a binomial proportion. Stat Sci 16(2):101–133 (Followed by a discussion by several authors)
Newcombe R (1998a) Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 17(2): 857–872
Newcombe R (1998b) Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med 17(2):873–890 (Correction: 1999, 18, 1293)
Salehi MM, Seber GAF (2001) A new proof of Murthy's estimator which applies to sequential sampling. Aust N Z J Stat 43: 901–906
Seber GAF (2002) The estimation of animal abundance, 2nd edn. Blackburn, Caldwell (Reprinted from the 1982 edition).
Scott AJ, Seber GAF (1983) Difference of proportions from the same survey. Am Stat 37(4):319–320
Wild CJ, Seber GAF (1993) Comparing two proportions from the same survey. Am Stat 47(3):178–181 (Correction: 1994, 48(3), 269)
Wild CJ, Seber GAF (2000) Chance encounters: a first course in data analysis and inference. Wiley, New York

Psychiatry, Statistics in

Graham Dunn
Professor of Biomedical Statistics and Head of the Health Methodology Research Group
University of Manchester, Manchester, UK

Why "Statistics in Psychiatry"? What makes statistics in psychiatry a particularly interesting intellectual challenge? Why is it not merely a sub-discipline of ▶medical statistics such as the application of statistics in rheumatology or statistics in cardiology? It is in the nature of mental illness and of mental health. Mental illness extends beyond

medicine into the realms of the social and behavioral sciences. Similarly, statistics in psychiatry owes as much, or more, to developments in social and behavioral statistics as it does to medical statistics. Statisticians in this area typically user a much wider variety of multivariate statistical methods than do medical statisticians elsewhere. Scientific psychiatry has always taken the problems of measurement much more seriously than appears to be the case in other clinical specialties. This is partly due to the fact that the measurement problems in psychiatry are obviously rather complex, but partly also because the other clinical fields appear to have been a bit backward by comparison. It is also an academic discipline where, at its best, there is fruitful interplay between the ideas typical of the 'medical model' of disease and those coming from the psychometric traditions of, say, educationalists and personality theorists.

"Mental diseases have both psychological, sociological and biological aspects and their study requires a combination of the approaches of the psychologist, the sociologist and the biologist, using the last word rather than physician since the latter must be all three. In *each* of these aspects statistical reasoning plays a part, whether it be in the future planning of hospitals, the classification of the various forms of such illnesses, the study of causation or the evaluation of methods of treatment." (Moran 1969 – my italics)

"The etiology of mental illness inevitably involves complex and inadequately understood interactions between social stressors and genetically and socially determined vulnerabilities – the whole area being overlaid by a thick carpet of measurement and misclassification errors." (Dunn 2000)

Do the varieties of mental illness fall into discrete, theoretically justified, diagnostic categories? Or are the boundaries entirely pragmatic and artificial? Should we be considering dimensions (matters of degree) or categories? Is the borderline between bipolar depression and schizophrenia, for example, real or entirely arbitrary? The same question even applies to the existence of mental illness itself. What distinguishes illness from social deviance and eccentricity? Establishing the validity and utility of psychiatric diagnoses has been, and still is, a major application of statistical thinking involving a whole range of complex multivariate methods (factor analysis being one of the most prominent). Once psychiatrists have created a diagnostic system they also need to be able to demonstrate its reliability. They need to be confident that psychiatrists will consistently allocate patients with a particular profile of symptoms to the same diagnostic group. They need to know that the category "schizophrenia" means the same thing and is used in the same way by all scientific psychiatrists, whether they work in the USA, China or Uganda. Here, again, statistical methods (kappa coefficients, for example) hold the centre stage.

The development of rating scales and the evaluation of the associated measurement errors form the central core of statistics in psychiatry. It is the problem of measurement that makes psychiatry stand apart from the other medical specialties. Scientific studies in all forms of medicine (including psychiatry) need to take account of confounding and selection effects. Demonstration of treatment efficacy for mental illness, like treatment efficacy elsewhere in medicine, always needs the well-designed controlled randomized trial. But measurement and measurement error make psychiatry stand out. Here, the statistician's world has long been populated by latent variable models of all sorts – finite mixture and latent class models, factor analysis and item response models and, of course, structural equations models (SEM) allowing one to investigate associations and - with luck and careful design - causal links between these latent variables.

About the Author

Graham Dunn has been Professor of Biomedical Statistics and Head of the Biostatistics Group at the University of Manchester since 1996. Before that he was Head of the Department of Biostatistics and Computing at the Institute of Psychiatry. Graham's research is primarily focussed on the design and analysis of randomised trials of complex interventions, specialising on the evaluation of cognitive behavioural and other psychological approaches to the treatment of psychosis, depression and other mental health problems. Of particular interest is the design and analysis of multi-centre explanatory trials from which it is possible to test for and estimate the effects of mediation and moderation, and for the effects of dose (sessions attended) and the quality of the therapy provided (including therapist effects). He also has interests in the design and analysis of measurement reliability studies. A key methodological component of both of these fields of applied research is the development and implementation of econometric methods such as the use of instrumental variables. He is author or co-author of seven statistics books, including *Statistical Evaluation of Measurement Errors* (Wiley, 2nd edition 2004), *Statistics in Psychiatry* (Hodder Arnold, 2000) and (with Brian Everitt) *Applied Multivariate Data Analysis* (Wiley, 2nd edition 2001).

Cross References

►Factor Analysis and Latent Variable Modelling
►Kappa Coefficient of Agreement
►Medical Statistics

►Rating Scales
►Structural Equation Models

References and Further Reading

Dunn G (2000) Statistics in psychiatry. Arnold, London
Moran PAP (1969) Statistical methods in psychiatric research (with discussion). J Roy Stat Soc A 132:484–525

Psychological Testing Theory

Vesna Buško
Associate Professor, Faculty of Humanities and Social Sciences
University of Zagreb, Zagreb, Croatia

Testing and assessment of individual differences have been a critical part of the professional work of scientists and practitioners in psychology and related disciplines. It is generally acknowledged that psychological tests, along with the existing conceptualizations of measurements of human potential, are among the most valuable contributions of the behavioral sciences to society. Testing practice is for many reasons an extremely sensitive issue, and is not only a professional but also a public issue. As the decisions based on test results and their interpretations often entail important individual and societal consequences, psychological testing has been the target of substantial public attention and long-standing criticism (AERA, APA, and NCME 2006).

The theory of psychological tests and measurement, or, as typically referred to, test theory or psychometric theory, offers a general framework and a set of techniques for evaluating the development and use of psychological tests. Due to their latent nature, the majority of psychological constructs are typically measured indirectly, i.e., by observing behavior on appropriate tasks or responses to test items. Different test theories have been proposed to provide rationales for behaviorally based measurement.

Classical test theory (CTT) has been the foundation of psychological test development since the turn of the twentieth century (Lord and Novick 1968). It comprises a number of psychometric models and techniques intended to estimate theoretical parameters, including the description of different psychometric properties, such as the derivation of reliability estimates and ways to assess the validity of use of test. This knowledge is crucial if we are to make sound inferences and interpretations from the test scores.

The central notion of CTT is that any observed test score (X) can be decomposed into two additive components: a *true score* (T) and a random *measurement error* term (e). Different models of CTT have been proposed, each defined by specific sets of assumptions that determine the circumstances under which the model may be reasonably applied. Some assumptions are associated with properties of measurement error as random discrepancies between true and observed test scores, whereas others include variations of the assumption that the two tests measure the same attribute. The latter assumption is essential for deducing test reliability, i.e., the ratio of true score variance to observed score variance, from the discrepancy between two measurements of the same attribute in the same person.

CTT and its applications have been criticized for various weaknesses, such as population dependence of its parameters, focus on a single undifferentiated random error, or arbitrary definition of test score variables. Generalizability theory (see Brennan 2001) was developed as an extension of the classical test theory approach, providing a framework for estimating the effects of multiple sources of error or other factors determining test scores. Another generalization of CTT has been put forward within the formulation of the Latent State-Trait Theory (see Steyer 2003). Formal definitions of states and traits have been introduced, and models allowing the separation of persons, situations, and/or interaction effects from measurement error components of the test scores are presented.

A more recent development in psychometric theory, item response theory (IRT), emerged to address some of the limitations of the classical test theory (Embretson and Reise 1999). The core assumption of IRT is that the probability of a person's expected response to an item is the joint function of that person's ability, or her/his position on the latent trait, and one or more parameters characterizing the item. The response probability is presented in the form of an item characteristic curve as a function of the latent trait.

Despite the controversies and criticisms surrounding CTT, and the important advances and challenges in the field of IRT, both classical and modern test theories appear today to be widely used and are complementary in designing and evaluating psychological and educational tests.

Cross References

►Psychology, Statistics in

References and Further Reading

AERA, APA, NCME (2006) Standardi za pedagoško i psihološko testiranje (Standards for Educational and Psychological Testing). Naklada Slap, Jastrebasko

Brennan RL (2001) Generalizability theory. Springer, New York

Embretson SE, Reise SP (1999) Item response theory for psychologists. LEA, Mahwah

Lord FC, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley Publishing Company, Reading

Steyer R (2003) Wahrscheinlichkeit und regression. Springer, Berlin

Psychology, Statistics in

Joseph S. Rossi
Professor, Director of Research at the Cancer Prevention Research Center
University of Rhode Island, Kingston, RI, USA

The use of quantitative methods in psychology is present essentially at its beginning as an independent discipline, and many of the early developers of statistical methods, such as Galton, Pearson, and Yule, are generally considered by psychologists as among the major contributors to the development of psychology itself. In addition, many early psychologists made major contributions to the development of statistical methods, often in the context of psychometric measurement theory and multivariate methods (e.g., Spearman, Thurstone). Among the techniques that psychologists developed or helped to develop during the early part of the twentieth century are the correlation coefficient, the chi-square test, regression analysis, factor analysis (see ►Factor Analysis and Latent Variable Modelling), ►principal components analysis, and various multivariate procedures. The use of the ►analysis of variance (ANOVA) in psychology did not begin until about 1940 and quickly became widespread.

During the decades of the 1940s and 1950s, a kind of schism arose among psychologists, with experimental psychologists favoring the use of ANOVA techniques and psychologists interested in measurement and individual differences favoring correlation and regression techniques, culminating in Cronbach's famous declaration concerning the "two disciplines" of scientific psychology. That these procedures were both aspects of the general linear model (see ►General Linear Models) and essentially equivalent mathematically did not become widely known among psychologists until about 1968. A similar sort of schism with respect to models of statistical inference has been resolved with a kind of hybrid model that accommodates both the Fisher and Neyman-Pearson approaches, although in this case, most researchers in psychology are completely unaware that such a schism ever existed, and that the models of statistical decision-making espoused in their textbooks and in common everyday use represent a combination of views thought completely antithetical by their original proponents. Bayesian approaches, while not unknown in psychology, remain vastly underutilized.

Statistical methods currently in common use in psychology include: Pearson product-moment correlation coefficient, chi-square test (see ►Chi-Square Tests), *t* test, univariate and multivariate analysis of variance (see ►Analysis of Variance and ►Multivariate Analysis of Variance (MANOVA)) and covariance with associated follow-up procedures (e.g., Tukey test), multiple regression, factor analysis (see ►Factor Analysis and Latent Variable Modelling) and ►principal components analysis, discriminant function analysis, path analysis, and structural equation modeling (see ►Structural Equation Models). Psychologists have been instrumental in the continued development and refinement of many of these procedures, particularly for measurement oriented procedures, such as item response theory, and structural equation modeling techniques, including confirmatory factor analysis, latent growth curve modeling, multiple group structural invariance modeling, and models to detect mediation and moderation effects. There is considerable emphasis on group level data analysis using parametric statistical procedures and the assumptions of univariate and multivariate normality. The use of nonparametric procedures, once fairly common, has declined substantially in recent decades. The use of more modern nonparametric techniques and robust methods is almost unknown among applied researchers.

The Null Hypothesis Significance Testing Controversy

Common to many of the procedures in use in psychology is an emphasis on null hypothesis significance testing (NHST) and concomitant reliance on statistical test *p* values for assessing the merit of scientific hypotheses. Considering its still dominant position, the use of the NHST paradigm in psychology and related disciplines has been subject to numerous criticisms over a surprisingly long period of time, starting at least 70 years ago. Until recently, these criticisms have not gained much traction. Common objections raised against the NHST paradigm include the following:

- The null is not a meaningful hypothesis and is essentially always false.

- Rejection of the null hypothesis provides only weak support for the alternative hypothesis.
- Failure to reject the null hypothesis does not mean that the null can be accepted, so that null results are inconclusive.
- Significance test p values are misleading in that they depend largely on sample size and consequently do not indicate the magnitude or importance of the obtained effect.
- The obtained p value is unrelated to, but frequently confused with both the study alpha (α) level and $1 - \alpha$.
- Reliance on p values has led to an overemphasis on the type I error rate and to the neglect of the type II error rate.
- Statistical significance is not the same as scientific or practical significance.
- The NHST approach encourages an emphasis on point estimates of parameter values rather than confidence intervals.
- The use of the $p < 0.05$ criterion for ►statistical significance is arbitrary and has led to dichotomous decision making with regard to the acceptance/rejection of study hypotheses. This has resulted in the phenomenon of "publication bias," which is the tendency for studies that report statistical significance to be published while those that do not are not published, despite the overall quality or merit of the research.
- The dichotomous decision making approach inherent to the NHST paradigm has seriously compromised the ability of researchers to accumulate data and evidence across studies. This has hindered the development of theories in many areas of psychology, since a few negative results tend to be accorded more weight than numerous positive results.

The extent and seriousness of these criticisms has led some to suggest an outright ban on the use of significance testing in psychology. Once inconceivable, this position is receiving serious consideration in numerous journal articles in the most prestigious journals in psychology, has been discussed by recent working groups and task forces on quantitative methods and reporting standards in psychology, and is even the subject of one recent book. Even among those not willing to discard significance testing entirely, there is widespread agreement on a number of alternative approaches that would reduce reliance on p values. These include the use of confidence intervals, effect size indices, ►power analysis, and ►meta-analysis. Confidence intervals (see ►Confidence Interval) provide useful information beyond that supplied by point estimates of parameters and p values. In psychology, the most frequently used has been the 95% confidence interval. Despite its simplicity, confidence intervals are still not widely used and reported or even that well understood by many applied researchers. For example, some recent studies have indicated that even when error bars are shown on graphs, it is not at all clear if authors intended to show standard deviations, standard errors, or confidence intervals.

Measures of effect size have been recommended as supplements to or even as substitutes for reporting p values. These provide a more direct index of the magnitude of study results and are not directly influenced by study sample size. Measures of effect size fall into two broad categories. The most common historically are measures of the proportion of shared variance between two (or more) variables, such as r^2 and R^2. These indices have a long history of use in research employing correlational and regression methods, particularly within the tradition of individual differences research. Within the tradition of experimental psychology, comparable measures have been available for many years but have not seen widespread use until relatively recently. Most commonly used for the t test and ANOVA are η^2 and ω^2, which index the proportion of variance in the dependent variable that is accounted for by the independent variable. Again, under the general linear model (see ►General Linear Models), these indices are essentially equivalent mathematically and retain their separate identities primarily for historical purposes. Multivariate analogs exist but are not well known and not often used.

A second and more recent approach to measuring the magnitude of study outcomes independent of p values is represented by standardized measures of effect size. These were developed by Cohen as early as 1960 but have not seen widespread use until relatively recently. While he developed a wide range of such indices designed to cover numerous study designs and statistical methods, by far the most widely known and used is Cohen's d. An advantage of d is its simplicity:

$$d = (M_1 - M_2)/s_p$$

where M_1 and M_2 represent the means of two independent groups and s_p is the pooled within-groups standard deviation. It is also easily related to proportion of variance measures:

$$\eta^2 = d^2/(d^2 + 4).$$

A disadvantage of d is that is it applicable only to the comparison of two groups. A generalized version applicable to the ANOVA F test is Cohen's f. However, this measure is much less well-known and not often utilized. As a guide to use and interpretation, Cohen developed a simple rubric

for categorizing the magnitudes of effect sizes. For *d*, small, medium, and large effect sizes are defined as 0.20, 0.50, and 0.80, respectively. The equivalent magnitudes for proportion of variance accounted for are 0.01, 0.06, and 0.14, respectively. Cohen recognized these definitions as arbitrary, but subsequent research suggests they hold up well across a broad range of research areas in the "softer" areas of psychology.

Although methods for aggregating data across independent studies have been in use for more than 100 years, a more formal and systematic approach did not begin to take shape until 1976 with the independent work of Rosenthal and Glass, who coined the term meta-analysis. While very controversial at first, and to a lesser extent still, the technique caught on rapidly as an efficient way to summarize quantitatively the results of a large number of studies, thus overcoming the heavy reliance on *p* values used in the more traditional narrative literature review. In principle any outcome measure can be employed; however, in practice meta-analysis relies heavily on the use of effect sizes as the common metric integrated across studies. Many studies employ the Pearson correlation coefficient *r* for this purpose, although Cohen's *d* is without question the most frequently used, primarily due to its simplicity and easy applicability to a wide range of focused two-group comparisons characteristic of many studies in psychology (e.g., control vs treatment, men vs women). It is probably fair to say that the rise of meta-analysis over the past 20–30 years has greatly facilitated and popularized the concept of the effect size in psychology. As a consequence, a great deal of work has been conducted on *d* to investigate its properties as a statistical estimator. This has resulted in substantial advances in meta-analysis as a statistical procedure. Most early meta-analyses employed simple two-group comparisons of effect size across studies using a fixed effects model approach (often implicitly). Most recent applications have emphasized a regression model approach in which numerous study-level variables are quantified and used as predictors of effect size (e.g., subject characteristics, study setting, measures and methods used, study design quality, funding source, publication status and date, author characteristics, etc.). Fixed effects models still predominate, but there is growing recognition that random (or mixed) effects models may be more appropriate in many cases. A considerable array of follow-on procedures have been developed as aids in the interpretation of meta-analysis results (e.g., effect size heterogeneity test *Q*, funnel and forest plots, assessment of publication bias, fail-safe number, power analysis, etc.). When done well, meta-analysis not only summarizes the literature, it identifies gaps and provides clear suggestions for future research.

Current Trends and Future Directions

Quantitative specialists in psychology continue to work on methods and design in a number of areas. These include data descriptive and exploratory procedures and alternatives to parametric methods, such as ▶exploratory data analysis and cluster analysis (see ▶Cluster Analysis: An Introduction), robust methods, and computer-intensive methods. Work focusing on design includes alternatives to randomized designs, methods for field experiments and quasi-experimental designs, and the use of fractional ANOVA designs. Structural equation modeling continues to receive a lot of attention, including work on latent growth curve modeling, latent transition analysis, intensive longitudinal modeling, invariance modeling, multiple group models, multilevel models, hierarchical linear modeling, and models to detect mediation and moderation effects.

Missing data analysis and multiple imputation methods (see ▶Multiple Imputation), especially for longitudinal designs, is also receiving considerable emphasis. The increased interest in longitudinal approaches has not been limited to group designs. The single subject/idiographic approach, also known as the person-specific paradigm, has been the focus of much recent work. This approach focuses on change over time at the individual level, exemplified by time series analysis, intensive longitudinal modeling, dynamic factor analysis, and dynamic cluster analysis.

Psychologists also continue to conduct a great deal of work on meta-analysis and integrative data analysis, particularly on random effects and hierarchical model approaches, as well as on investigations of the properties of various effect size indices as statistical estimators and the application and development of effect size indices for a wider range of study designs. Increased use of meta-analysis by applied researchers is encouraging the use of alternatives to null hypothesis testing, including the specification of non-zero null hypotheses, and the use of alternative hypotheses that predict the magnitude of the expected effect sizes.

About the Author

Joseph S. Rossi, Ph.D., is Professor and Director of the Behavioral Science Ph.D. program in the Department of Psychology, and Director of Research at the Cancer Prevention Research Center, at the University of Rhode Island. He has been principal investigator or co-investigator on more than 50 grants and has published more than 150 papers and chapters. In 1996, the American Psychological Society and the Institute for Scientific Information listed Dr. Rossi 5th in author impact (citations/paper) and 12th in number of citations (*APS Observer*, January

1996, pp. 14–18). In 2006, he was named one of the most highly cited researchers in the world in the fields of Psychology/Psychiatry by Thomson Reuters (http://isihighlycited.com/). He won the University of Rhode Island's Scholarly Excellence Award in 2003, was elected to membership in the Society of Multivariate Experimental Psychology in 1995, and is a fellow of Division 5 of the American Psychological Association. Dr. Rossi is one of the principal developers of the transtheoretical model of health behavior change. His areas of interest include quantitative psychology, health promotion and disease prevention, and expert system development for health behavior change. Dr. Rossi is a member of the University of Rhode Island's Institutional Review Board.

Cross References

►Confidence Interval
►Cross Classified and Multiple Membership Multilevel Models
►Effect Size
►Factor Analysis and Latent Variable Modelling
►Frequentist Hypothesis Testing: A Defense
►Meta-Analysis
►Moderating and Mediating Variables in Psychological Research
►Multidimensional Scaling
►Multidimensional Scaling: An Introduction
►Null-Hypothesis Significance Testing: Misconceptions
►Psychological Testing Theory
►Sociology, Statistics in
►Statistics: Controversies in Practice

References and Further Reading

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008) Reporting standards for research in psychology: why do we need them? What might they be? Am Psychol 63:839–851
Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum, Hillsdale, NJ
Cohen J (1990) Things I have learned (so far). Am Psychol 45: 1304–1312
Cohen J (1994) The earth is round ($p < .05$). Am Psychol 49: 997–1003
Cowles M (1989) Statistics in psychology: an historical perspective. Lawrence Erlbaum, Hillsdale, NJ
Cronbach LJ (1957) The two disciplines of scientific psychology. Am Psychol 12:671–684
Cumming G, Finch S (2005) Inference by eye: confidence intervals and how to read pictures of data. Am Psychol 60:170–180
Cumming G, Fidler F, Leonard M, Kalinowski P, Christiansen A, Kleinig A, Lo J, McMenamin N, Wilson S (2007) Statistical reform in psychology: is anything changing? Psychol Sci 18: 230–232
Grissom RJ, Kim JJ (2005) Effect sizes for research: a broad practical approach. Lawrence Erlbaum, Hillsdale, NJ
Harlow LL, Mulaik SA, Steiger JH (eds) (1997) What if there were no significance tests? Lawrence Erlbaum, Hillsdale, NJ
Kline RB (2004) Beyond significance testing: Reforming data analysis methods in behavioral research. American Psychological Association, Washington, DC
Lipsey MW, Wilson DB (2000) Practical meta-analysis. Sage, Thousand Oaks, CA
MacKinnon DP (2008) Introduction to statistical mediation analysis. Lawrence Erlbaum, New York
Maxwell SE (2004) The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychol Meth 9:147–163
Meehl PE (1978) Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. J Consult Clin Psychol 46:806–834
Molenaar P, Campbell CG (2009) The new person-specific paradigm in psychology. Current Directions in Psychological Science 18:112–117
Nickerson RS (2000) Null hypothesis significance testing: A review of an old and continuing controversy. Psychol Meth 5:241–301
Shadish WR, Cook TD (2009) The renaissance of field experimentation in evaluating interventions. Annual Review of Psychology 60:607–629
Wilkinson L, and the Task Force on Statistical Inference (1999) Statistical methods in psychology journals: Guidelines and explanations. American Psychologist 54:594–604

Public Opinion Polls

Christopher Wlezien
Professor of Political Sciences
Temple University, Philadelphia, PA, USA

A public opinion poll is a survey of the views of a sample of people. It is what we use to measure public opinion in the modern day. This was not always true, of course, as nonrandom "straw polls" have been in regular use at least since the early nineteenth century. We thus had information even back then about what the public thought and wanted, though it was not very reliable. The development of probability sampling, and its application by George Gallup, Archibald Crossley, and Elmo Roper, changed things in important ways, as we now had more reliable information about public opinion (see Geer 1996). The explosion of survey data since that time has fueled the growth in research on attitudes and opinion and behavior that continues today.

There are various forms of probability sampling. In simple random sampling respondents are selected purely at random from the population. This is the most basic form of probability sampling and does a good job representing

the population particularly as the sample size increases and sampling error declines. In stratified random sampling, the population is divided into strata, e.g., racial or ethnic groups, and respondents are selected randomly from within the strata. This approach helps reduce sampling error across groups, which can result from simple random sampling. Traditionally, most survey organizations have relied on ▶cluster sampling. Here the population is divided into geographic clusters, and the survey researcher draws a sample of these clusters and then samples randomly from within them. This is particularly useful when respondents are geographically disbursed. Survey organizations using any of these methods traditionally have relied on face-to-face interviews.

The technology of public opinion polling has changed quite dramatically over time. The invention of random digit dialing (RDD) had an especially significant impact, as interviewing could be done over the telephone based on lists of randomly-generated phone numbers. The more recent introduction of internet polling is having a similar impact. These developments have clear and increasing advantages in cost and speed, and have made it much easier to conduct polls. Witness the growth in the number of pre-election trial-heat polls in presidential election years in the United States (US) (Wlezien and Erikson 2002). Consider also that National Annenberg Election Survey (NAES) conducted over 100,000 telephone interviews during the 2000 presidential election campaign, with similar numbers in 2004 and 2008. In the same election years Knowledge Networks' conducted repeated interviews with 29,000 individuals via the internet. Similar developments can be seen in other countries, including Canada and the United Kingdom. These numbers would be almost inconceivable using face-to-face interviews.

The developments also come with disadvantages. To begin with, there is coverage error. Not everyone has a telephone, and the number relying solely on a cell phone – which poses special challenges for telephone surveys – is growing. Fewer have access to the internet and we cannot randomly e-mail them (note that this precludes calculations of sampling error, and thus confidence intervals. Internet polls do have a number of advantages for scholarly research, however (Clarke et al. 2008)). Even among those we can reach, nonresponse is a problem. The issue here is that respondents who select out of surveys may not be representative of the population. Survey organizations and scholars have long relied on weighting devices to address coverage and nonresponse error (besides sampling error and coverage and nonresponse error, all polls are subject to measurement error, which reflects flaws in the survey instrument itself, including question wording, order, interviewer training and other things. For a treatment of the different forms of survey error, see Weissberg (2005)). In recent years more complicated approaches have begun to be used, including "propensity scores" (see, e.g., Terhanian 2008).

A recent analysis of polling methods in the 2008 US presidential election campaign suggested that the survey mode had little effect on poll estimates (AAPOR 2009). The extent to which the weighting fixes used by survey organizations succeed is the subject of ongoing research – see, e.g., work on internet polls by Malhotra and Krosnick (2007) and Sanders et al. (2007). It is of special importance given the appeal of internet surveys owing to the speed with which they can be conducted and their comparatively low cost.

Despite the difficulties, polls have performed very well. Pre-election polls have proved very accurate at predicting the final vote, particularly at the end of the campaign (Traugott 2005). They also predict well earlier on, though it is not an identity relation, e.g., in US presidential elections, early leads tend to decline by Election Day (Wlezien and Erikson 2002). Pre-election polls in recent election years have provided more information about the election outcome than highly-touted election prediction markets (Erikson and Wlezien 2008). Polls also tell quite a lot about public opinion (see, e.g., Stimson 1991; Page and Shapiro 1992; Erikson and Tedin 2009). Policymakers now have reliable information about the preferences of those with a stake in the policies they make. It also appears to make a difference to what they actually do (Geer 1996).

Acknowledgments

I thank my colleague Michael G. Hagen for very helpful comments.

About the Author

Dr. Christopher Wlezien is Professor, Department of Political Science, Temple University, Philadelphia. While at Oxford University, he co-founded the Spring School in Quantitative Methods for Social Research. He is an elected member of the International Statistical Institute (2006), and holds or has held visiting positions at Columbia University, the European University Institute, Institute Empresa, Juan March Institute, McGill University, L'Institut d'Etudes Politiques de Paris, and the University of Manchester. He has authored or co-authored many papers and books, including *Degrees of Democracy* (Cambridge, 2010), in which he develops and tests a thermostatic model of public opinion and policy. Currently, he is founding co-editor of the *Journal of Elections, Public Opinion and Parties* and Associate editor of *Public Opinion Quarterly*.

Cross References

►Margin of Error
►Nonresponse in Surveys
►Questionnaire
►Social Statistics
►Sociology, Statistics in
►Telephone Sampling: Frames and Selection Techniques

References and Further Reading

American Association for Public Opinion Research (2009) An evaluation of the methodology of the 2008 pre-election primary polls. Available at http://aapor.org/uploads/AAPOR_Rept_FINAL-Rev-4-13-09.pdf

Clarke HD, Sanders D, Stewart MC, Whiteley P (2008) Internet surveys and national election studies: a symposium. Journal of Elections, Public Opinion and Parties 18:327–330

Erikson RS, Tedin KL (2009) American public opinion. Longman, New York

Erikson RS, Wlezien C (2008) Are political markets really superior to polls as election predictions? Public Opin Quart 72:190–215

Geer J (1996) From tea leaves to public opinion polls. Columbia University Press, New York

Malhotra N, Krosnick JA (2007) The effect of survey mode on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to internet surveys with non-probability samples. Polit Anal 15:286–323

Page B, Shapiro RY (1992) The rational public. University of Chicago Press, Chicago

Sanders D, Clarke HD, Stewart MC, Whiteley P (2007) Does mode matter for modeling political choice: evidence from the british election study. Polit Anal 15:257–285

Stimson JN (1991) Public opinion in american: moods, cycles and swings. Westview Press, Boulder, CO

Terhanian G (2008) Changing times, changing modes: the future of public opinion polling. Journal of Elections, Public Opinion and Parties 18:331–342

Traugott MW (2005) The accuracy of the pre-election polls in the 2004 presidential election. Public Opin Quart 69:642–654

Weissberg H (2005) The total survey error approach. University of Chicago Press, Chicago

Wlezien C, Erikson RS (2002) The timeline of presidential election campaigns. J Polit 64(4):969–993

P–Values

Raymond Hubbard
Thomas F. Sheehan Distinguished Professor of Marketing
Drake University, Des Moines, IA, USA

The origin of the *p*-value is credited to Karl Pearson (1900), who introduced it in connection with his chi-square test (see ►Chi-Square Tests). However, it was Sir Ronald Fisher who popularized significance tests and *p*-values in the multiple editions of his hugely influential books *Statistical Methods for Research Workers* and *The Design of Experiments*, first published in 1925 and 1935, respectively. Fisher used divergencies in the data to reject the null hypothesis by calculating the probability of the data on a true null hypothesis, or $\Pr(x|H_0)$. More formally, $p = \Pr(T(X) \geq T(x)|H_0)$. The *p*-value is the probability of getting a test statistic $T(X)$ larger than or equal to the observed result, $T(x)$, as well as more extreme ones, assuming a true null hypothesis, H_0, of no effect or relationship. Thus, the *p*-value is an index of the (im)plausibility of the actual observations (together with more extreme, unobserved ones) if the null is true, and is a random variable whose distribution is uniform over the interval $[0,1]$.

The reasoning is that if the data are viewed as being rare or extremely unlikely under H_0, this constitutes *inductive evidence* against the null hypothesis. Fisher (1966, p. 13) immortalized a *p*-value of 0.05 for rejecting the null: "It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard." Consequently, values like $p < 0.01$, $p < 0.001$, and so on, are said to furnish even stronger evidence against H_0. So Fisher considered *p*-values to play an important epistemological role (Hubbard and Bayarri 2003).

Moreover, Fisher (1959, p. 43) saw the *p*-value as an *objective* measure for judging the (im)plausibility of H_0:

> "…the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to and verifiable by other rational minds. The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief [in the null hypothesis] it engenders."

Researchers across the world have enthusiastically adopted *p*-values as a "scientific" and "objective" criterion for certifying knowledge claims.

Unfortunately, the *p*-value is neither an objective nor very useful measure of evidence in statistical significance testing (see Hubbard and Lindsay 2008, and the references therein). In particular, *p*-values exaggerate the evidence against H_0. Because of this, the validity of much published research with comparatively small (including 0.05) *p*-values must be called into question.

Of great concern, though obviously no fault of the index itself, members of the research community insist on investing *p*-values with capabilities they do not possess (for critiques of this, see, among others, Carver 1978; Nickerson 2000). Some common misconceptions regarding the *p*-value are that it denotes an objective measure of:

- The probability of the null hypothesis being true
- The probability (in the sense of $1 - p$) of the alternative hypothesis being true
- The probability (again, in the sense of $1 - p$) that the results will replicate
- The magnitude of an effect
- The substantive or practical significance of a result
- The Type I error rate
- The generalizability of a result

Despite its ubiquity, the *p*-value is of very limited use. Indeed, I agree with Nelder's (1999, p. 261) assertion that the most important task in developing a helpful statistical science is "to demolish the *P*-value culture."

About the Author

Dr. Raymond Hubbard is Thomas F. Sheehan Distinguished Professor of Marketing in the College of Business and Public Administration, Drake University, Des Moines, Iowa, USA. He has served as the Chair of the Marketing Department (1988–1989; 1992–1994; 2000–2003). He is a member of the American Marketing Association, Academy of Marketing Science, and the Association for Consumer Research. He has authored or coauthored over 50 journal articles, many of them methodological in nature. He is presently working on a book (with R. Murray Lindsay), tentatively titled "From Significant Difference to Significant Sameness: A Proposal for a Paradigm Shift in Managerial and Social Science Research."

Cross References

►Bayesian P-Values
►Effect Size
►False Discovery Rate
►Marginal Probability: Its Use in Bayesian Statistics as Model Evidence
►Misuse of Statistics
►Null-Hypothesis Significance Testing: Misconceptions
►Psychology, Statistics in
►P-Values, Combining of
►Role of Statistics
►Significance Testing: An Overview
►Significance Tests, History and Logic of
►Significance Tests: A Critique
►Statistical Evidence
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Inference: An Overview
►Statistical Significance
►Statistics: Controversies in Practice

References and Further Reading

Carver RP (1978) The case against statistical significance testing. Harvard Educ Rev 48:378–399
Fisher RA (1959) Statistical methods and scientific inference, 2nd edn. Oliver and Boyd, Edinburgh. Revised
Fisher RA (1966) The design of experiments, 8th edn. Oliver and Boyd, Edinburgh
Hubbard R, Bayarri MJ (2003) Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing (with comments). Am Stat 57:171–182
Hubbard R, Lindsay RM (2008) Why *P*-values are not a useful measure of evidence in statistical significance testing. Theory Psychol 18:69–88
Nelder JA (1999) From statistics to statistical science (with comments). Stat 48:257–269
Nickerson RS (2000) Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Meth 5: 241–301
Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London Edinburgh Dublin Philos Mag J Sci 50:157–175

P-Values, Combining of

Dinis Pestana
Professor, Faculty of Sciences
Universidade de Lisboa, DEIO, and CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, Lisboa, Portugal

Let us assume that the *p*-values p_k are known for testing H_{0k} versus H_{Ak}, $k = 1, \ldots, n$, in n independent studies on some common issue, and our aim is to achieve a decision on the overall question H_0^*: all the H_{0k} are true *versus* H_A^*: some of the H_{Ak} are true. As there are many different ways in which H_0^* can be false, selecting an appropriate test is in general unfeasible. On the other hand, combining the available p_k's so that $T(p_1, \ldots, p_n)$ is the observed value of a random variable whose sampling distribution under H_0^* is known is a simple issue, since under H_0^*, $\boldsymbol{p}$ is the observed value of a random sample $\boldsymbol{P} = (P_1, \ldots P_n)$ from a $Uniform(0,1)$ population. In fact, several different sensible combined testing procedures are often used.

A rational combined procedure should of course be *monotone*, in the sense that if one set of *p*-values $\boldsymbol{p} = (p_1, \ldots, p_n)$ leads to rejection of the overall null hypothesis H_0^*, any set of componentwise smaller *p*-values $\boldsymbol{p}' = (p_1', \ldots, p_n')$, $p_k' \leq p_k$, $k = 1, \ldots, n$, must also reject H_0^*.

Tippett (1931) used the fact that $P_{1:n} = \min\{P_1, \ldots, P_n\} \underset{|_{H_0^*}}{\sim} Beta(1,n)$ to reject H_0^* if the minimum observed p-value $p_{1:n} < 1-(1-\alpha)^{1/n}$. This *Tippett's minimum method* is a special case of *Wilkinson's method* (Wilkinson, 1951), advising rejection of H_0^* when some low rank order statistic $p_{k:n} < c$; as $P_{k:n} \sim Beta(k, n+1-k)$, to reject H_0^* at level α the cut-of-point c is the solution of $\int_0^c u^{k-1}(1-u)^{n-k}\mathrm{d}u = \alpha\, B(k, n+1-k)$.

The exact distribution of $\overline{P}_n = \frac{1}{n}\sum_{k=1}^n P_k$ is cumbersome, but for large n an approximation based on the central limit theorem (see ▶Central Limit Theorems) can be used to perform an overall test on H_0^* vs. H_A^*. On the other hand, the probability density function of the ▶geometric mean $G_n = \left(\prod_{k=1}^n P_k\right)^{\frac{1}{n}}$ of n independent uniform random variables is readily computed, $f_{G_n}(x) = \dfrac{n\,(-n\,x\,\ln\,x)^{n-1}}{\Gamma(n)}\mathrm{I}_{(0,1)}(x)$, leading to a more powerful test; see, however, the discussion below on publication bias.

Another way of constructing combined p-values is to use additive properties of simple functions of uniform random variables. Fisher (1932) used the fact that $P_k \sim Uniform(0,1) \Longrightarrow -2\ln P_k \sim \chi^2_2$, and therefore, $-2\sum_{k=1}^n \ln P_k \underset{|_{H_0^*}}{\sim} \chi^2_{2n}$. Then H_0^* is rejected at the significance level α if the $-2\sum_{k=1}^n \ln p_k > \chi^2_{2n,1-\alpha}$. Stouffer et al. (1949) used as test statistic $\sum_{k=1}^n \dfrac{\Phi^{-1}(P_k)}{\sqrt{n}} \underset{|_{H_0^*}}{\sim} Gaussian(0,1)$, where Φ^{-1} denotes the inverse of the distribution function of the standard Gaussian, rejecting H_0^* at level α if $\left|\sum_{k=1}^n \frac{\Phi^{-1}(P_k)}{\sqrt{n}}\right| > z_{1-\alpha}$.

Another simple transformation of uniform random variables P_k is the logit transformation, $\ln\frac{P_k}{1-P_k} \sim Logistic(0,1)$. As $\sum_{k=1}^n \dfrac{\ln\frac{P_k}{1-P_k}}{\sqrt{n\,\frac{\pi^2(5n+2)}{3(5n+4)}}} \approx t_{5n+4}$, reject H_0^* at the significance level α if $-\sum_{k=1}^n \dfrac{\ln\frac{p_k}{1-p_k}}{\sqrt{n\,\frac{\pi^2(5n+2)}{3(5n+4}}} > t_{5n+4,1-\alpha}$.

Birnbaum (1954) has shown that every monotone combined test procedure is *admissible*, i.e., provides a most powerful test against some alternative hypothesis for combining some collection of tests, and is therefore optimal for some combined testing situation whose goal is to harmonize eventually conflicting evidence, or to pool inconclusive evidence. In the context of social sciences Mosteller and Bush (1954) recommend Stouffer's method, but Littel and Folks (1971, 1973) have shown that under mild conditions Fisher's method is optimal for combining independent tests.

As in many other techniques used in ▶meta-analysis, publication bias can easily lead to erroneous conclusions. In fact, the set of available p-values comes only from studies considered worth publishing because the observed p-values were small, seeming to point out significant results. Thus the assumption that the p_k's are observations from independent $Uniform(0,1)$ random variables is questionable, since in general they are in fact a set of low order statistics, given that p-values greater than 0.05 have not been recorded. For instance, $\mathbb{E}\left(G_n^k\right) = \left(\frac{1}{1+\frac{k}{n}}\right)^n \underset{n\to\infty}{\longrightarrow} \mathrm{e}^{-k}$, and in particular $\mathbb{E}\left(G_n\right) = \left(\frac{n}{n+1}\right)^n \underset{n\to\infty}{\downarrow} \frac{1}{\mathrm{e}} \approx 0.3679$, the standard deviation decreases to zero, the skewness steadily decreases after a maximum 0.2645 for $n = 5$, and the kurtosis increases from -0.8541 (for $n = 2$) towards 0. Whenever $p_{n:n}$ falls below the critical rejection point, this test will lead to the rejection of H_0^*, but $p_{n:n}$ smaller than the critical point (for $n \geq 14$, the expected value of G_n is greater than 0.36 and the standard deviation is smaller than 0.1) is what should be expected as a consequence of publication bias.

Another important issue: H_A^* states that some of the H_{Ak} are true, and so a meta-decision on H_0^* implicitly assumes that some of the P_k may have non-uniform distribution, cf. Hartung et al. (2008, pp. 81–84) and Kulinskaya et al. (2008, pp. 117–119), and references therein, on the promising concepts of generalized and of random p-values. Gomes et al. (2009) investigated the effect of augmenting the available set of p-values with uniform and with non uniform pseudo-p-values, using results such as: Let X_{m_1} and X_{m_2} be independent random variables, X_m denoting a random variable with probability density function $f_m(x) = \left(mx + \frac{2-m}{2}\right)\mathrm{I}_{(0,1)}(x)$, $m \in [-2,2]$, i.e., a convex mixture of uniform and $Beta(1,2)$ (if $m \in [-2,0)$), thus favoring pseudo-p-values near 0 the sharper the slope is, the slope $m = 0$ corresponding to standard uniform, or of uniform and $Beta(2,1)$ (if $m \in (0,2]$), in this case favoring the occurrence of p-values near 1. Then $\min\left(\frac{X_{m_1}}{X_{m_2}}, \frac{1-X_{m_1}}{1-X_{m_2}}\right)$ is a member of the same family – more precisely $X_{\frac{m_1 m_2}{6}}$. In particular, if either $m_1 = 0$ or $m_2 = 0$, then $\min\left(\frac{X_{m_1}}{X_{m_2}}, \frac{1-X_{m_1}}{1-X_{m_2}}\right)$ can be used to generate a new set of uniform random variables, which moreover are independent of the ones used to generate them.

Extensive simulation, namely with computationally augmented samples of p-values (Gomes et al. 2009;

Brilhante et al. 2010) led to the conclusion that in what concerns decreasing power, and increasing number of unreported cases needed to reverse the overall conclusion of a meta-analysis, the methods of combining *p*-values rank as follows:

1. Arithmetic mean
2. ►Geometric mean
3. Chi-square transformation (Fisher's method)
4. Logistic transformation
5. Gaussian transformation (Stouffer's method)
6. Selected order statistics (Wilkinson's method)
7. Minimum (Tippett's method)

About the Author

Professor Dinis Pestana has been President of the Department of Statistics and Operations Research, University of Lisbon, for two consecutive terms (1986–1989), President of the Faculty of Sciences of Lisbon extension in Madeira (1985–1988) before the University of Madeira has been founded, President of the Center of Statistics and Applications, Lisbon University, for three consecutive terms (1981–1987). He supervised the Ph. D. studies of many students that played a leading role in the development of Statistics at the Portuguese universities, and is co-author of 40 papers published in international journals or as chapters of books, and many other papers, namely explaining Statistics to the layman. He launched the annual meetings of the Portuguese Statistical Society, and had a leading role on the local organization of international events, such as the 2001 European Meeting of Statisticians in Funchal.

Cross References

►Bayesian P-Values
►Meta-Analysis
►P-Values

References and Further Reading

Birnbaum A (1954) Combining independent tests of significance. J Amer Stat Assoc 49:559–575

Brilhante MF, Pestana D, Sequeira F (2010) Combining *p*-values and random *p*-values. In: Luzar-Stiffler V, Jarec I, Bekic Z (eds) Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, IEEE CFP10498-PRT, 515–520

Fisher RA (1932) Statistical methods for research workers, 4th edn. Oliver and Boyd, London

Gomes MI, Pestana D, Sequeira F, Mendonça, S, Velosa S (2009) Uniformity of offsprings from uniform and non-uniform parents. In: Luzar-Stiffler V, Jarec I, Bekic Z (eds) Proceedings of the ITI 2009, 31st International Conference on Information Technology Interfaces, pp 243–248

Hartung J, Knapp G, Sinha BK (2008) Statistical meta-analysis with applications. Wiley, New York

Kulinskaya E, Morgenthaler S, Staudte RG (2008) Meta analysis. a guide to calibrating and combining statistical evidence. Wiley, Chichester

Littel RC, Folks LJ (1971, 1973) Asymptotic optimality of Fisher's method of combining independent tests, I and II. J Am Stat Assoc 66:802–806, 68:193–194

Mosteller F, Bush R (1954) Selected quantitative techniques In: Lidsey G (ed) Handbook of social psychology: theory and methods, vol I. Addison-Wesley, Cambridge

Stouffer SA, Schuman EA, DeVinney LC, Star S, Williams RM (1949) The American Soldier, vol I: Adjustment during army life. Princeton University Press, Princeton

Tippett LHC (1931) The methods of statistics. Williams and Norgate, London

Wilkinson B (1951) A statistical consideration in psychological research. Psychol Bull 48:156–158

Pyramid Schemes

Robert T. Smythe
Professor
Oregon State University, Corvallis, OR, USA

A pyramid scheme is a business model in which payment is made primarily for enrolling other people into the scheme. Some schemes involve a legitimate business venture, but in others no product or services are delivered. A typical pyramid scheme combines a plausible business opportunity (such as a dealership) with a recruiting operation that promises substantial rewards. A recruited individual makes an initial payment, and can earn money by recruiting others who also make a payment; the recruiter receives part of these receipts, and a cut of future payments as the new recruits go on to recruit others. In reality, because of the geometrical progression of (hypothetical) recruits, few participants in a pyramid scheme will be able to recruit enough others to recover their initial investment, let alone make a profit, because the pool of potential recruits is rapidly exhausted.

Although they are illegal in many countries, pyramid schemes have existed for over a century. As recently as November 2008, riots broke out in several towns in Colombia after the collapse of several pyramid schemes, and in 2006 Ireland launched a website to better educate consumers to pyramid fraud after a series of schemes were perpetrated in Cork and Galway.

Perhaps the best-known type of pyramid scheme is a *chain letter*, which often does not involve even a fictitious product. A chain letter may contain k names; purchasers of the letter invest \$2x, with \$$x$ paid to the name at the top of the letter and \$$x$ to the seller of the letter. The purchaser deletes the name at the top of the list, adds his own at the bottom, and sells the letter to new recruits. The promoter's pitch is that if the purchaser, and each subsequent recruit for k−1 stages, sells just two letters, there will be 2^{k-1} people selling 2^k letters featuring the purchaser's name at the top of the list, so that the participant would net $\$2^k x$ from the venture. Many variants of this basic "get rich quick" scheme have been, and continue to be, promoted.

A structure that can be used to model many pyramid schemes is that of recursive trees. A tree with n vertices labeled $1, 2, \ldots, n$ is a *recursive tree* if node 1 is distinguished as the *root*, and for each j with $2 \le j \le n$, the labels of the vertices in the unique path from the root to node j form an increasing sequence. The special case of *random* or *uniform* recursive trees, in which all trees in the set of trees of given order n are equally probable, has been extensively analyzed (cf. Smythe and Mahmoud (1994), for example); however, most pyramid schemes or chain letters in practice have restrictions making their probability models non-uniform. The number of places where the next node may join the tree is then a random variable, unlike the uniform case. This complicates the analysis considerably (and may account for the relative sparsity of mathematical analysis of the properties of pyramid schemes).

Bhattacharya and Gastwirth (1983) analyze a chain letter scheme allowing reentry, in which each purchaser may sell only two letters, unless he purchases a new letter to re-enter the chain. In terms of recursive trees, this means that a node of the tree is *saturated* once it has two offspring nodes, and no further nodes can attach to it. It is further assumed that at each stage, participants who have not yet sold two letters all have an equal chance to make the next sale, i.e., all unsaturated nodes of the recursive tree have an equal chance of being the "parent" of the next node to be added. If L_n denotes the number of leaves of the recursive tree (nodes with no offspring) at stage n under this growth rule, L_n/n corresponds to the proportion of "shutouts" (those receiving no revenue) in this chain letter scheme. The analysis of Bhattacharya and Gastwirth sets up a nonhomogeneous Markov chain model and derives a diffusion approximation for large n. They find that L_n/n converges to 0.382 in this model and that the (centered and scaled) number of shutouts has a normally distributed limit. Mahmoud (1994) considers the height h_n of the tree of order n in this same "random pyramid" scheme and show that it converges with probability 1 to 3.98912; the proof involves embedding the discrete-time growth process of the pyramid in a continuous time birth-and-death process. Mahmoud notes that a similar analysis could be carried out for schemes permitting the sale of m letters, provided that the probabilistic behavior of the total number of shutouts could be derived (as it was in the binary case by Bhattacharya and Gastwirth).

Gastwirth (1977) and Gastwirth and Bhattacharya (1984) analyze another variant of pyramid schemes, known as a quota scheme. This places a limit on the maximum number of participants, so that the scheme corresponds to a recursive tree of some fixed size n. This scheme derives from a case in a Connecticut court (Naruk 1975) in which people bought dealerships in a "Golden Book of Values," then were paid to recruit other dealers. In this scheme, each participant receives a commission from all of his descendants; thus for the jth participant, the size of the branch of the tree rooted at j determines his profit. If S_j denotes the size of this branch, and j/n converges to a limit θ, Gastwirth and Bhattacharya showed that the distribution of S_j converges to the geometric law

$$P(S_j = i+1) = \theta(1-\theta)^i \quad \text{for } i = 0, 1, 2, \ldots$$

(It was later shown (Mahmoud and Smythe (1991)) that if j is fixed, the limiting distribution of S_j/n is $Beta(1, j-1)$.) Calculations made by Gastwirth and Bhattacharya show that, for example, when n is fixed at 270, the 135th entry has probability only about 0.15 of recruiting two or more new entrants, and probability 0.03 of three or more recruits. Gastwirth (1977) shows that for large n, the expected proportion of all participants who are able to recruit at least r persons is 2^{-r}.

Other variants of pyramid schemes include the "8-Ball Model" and the "2-Up System" (http://www.mathmotivation.com/money/pyramid-scheme.html). In the eight-Ball model, the participant again recruits two new entrants, but does not receive any payment until two further levels have been successfully recruited. Thus a person at any level in the scheme would theoretically receive $2^3 = 8$ times his "participation fee," providing incentive to help those in lower levels succeed. In the two-Up scheme, the income from a participant's first two recruits goes to the individual who recruited the participant; if the participant succeeds in recruiting three or more new entrants, the income received from these goes to the participant, along with the income from the first two sales made by each subsequent recruit. This scheme creates considerable incentive to pursue the potentially lucrative third recruit. For both of these schemes, it is easily calculated that when the pool of prospective recruits is exhausted, the majority of the participants in the scheme end up losing money.

About the Author

Robert Smythe is Professor of Statistics at Oregon State University. He was Chair of the Department of Statistics for ten years and previously chaired the Department of Statistics at George Washington University for eight years. He has written in many areas of probability and statistics, and is a Fellow of the American Statistical Association and a Fellow of the Institute for Mathematical Statistics.

Cross References

►Beta Distribution
►Geometric and Negative Binomial Distributions
►Markov Chains

References and Further Reading

Bhattacharya P, Gastwirth J (1983) A non-homogeneous Markov model of a chain letter scheme. In: Rizvi M, Rustagi J, Siegmund D (eds) Recent advances in statistics: papers in honor of Herman Chernoff. Academic, New York, pp 143–174

Gastwirth J (1977) A probability model of a pyramid scheme. Am Stat 31:79–82

Gastwirth J, Bhattacharya P (1984) Two probability models of pyramids or chain letter schemes demonstrating that their promotional claims are unreliable. Oper Res 32:527–536

Mahmoud H (1994) A strong law for the height of random binary pyramids. Ann Appl Probab 4:923–932

Mahmoud H, Smythe RT (1991) On the distribution of leaves in rooted subtrees of recursive trees. Ann Prob 1: 406–418

http://www.mathmotivation.com/money/pyramid-scheme.html

Naruk H (1975) Memorandum of decision: State of Connecticut versus Bull Investment Group, 32 Conn. Sup. 279

Smythe RT, Mahmoud H (1994) A survey of recursive trees. Theor Probab Math Stat 51:1–27

Q

Quantitative Risk Management

Paul Embrechts
Professor of Mathematics
ETH Zurich, Zurich, Switzerland

Introduction

Quantitative Risk Management (QRM for short) is a relatively new field of mathematical research on the scientific firmament. As a field of science, QRM concentrates on the axiomatisation, the measurement and the analysis of risk in a rather broad context. Examples range from the construction of dykes, over the development of new medical compounds to the calculation of risk capital for insurance companies and banks. The quantification of risk and the societal challenges concerning "living with risks" are well known to all; how high do we need to build a sea dyke in order to protect a geographic area and its inhabitants, or what are prudent regulatory guidelines in order to safeguard a stable financial system? It is immediately clear that an overview of the field is out of the question, the various acronyms encountered stand proof of this: besides QRM, ERM (Enterprise-wide Risk Management), GRM (Global RM), IRM (Integrative RM), and no doubt others. A broad overview of the various faces of RM can for instance be obtained from (Melnick and Everitt 2008). In this paper we will restrict to a rather specific, though important interpretation, that of McNeil et al. (2005); the latter is mainly derived from the field of banking and insurance, though the concepts, techniques and tools are much more widely applicable. First of all, we will look at risk as related to randomness or uncertainty. As such, the main tools discussed in McNeil et al. (2005) concern the field of stochastics: probability theory, statistics and the theory of stochastic processes. I am very well aware that this restriction is a shortcoming but so be it. It suffices to stress over and over again that in any application of importance and substance, it pays for the quantitative risk modeller to show a fair amount of humbleness. Also, that same modeller must realise that such "real problems" only can be solved in an interdisciplinary context, and that the quantitative analyst (often referred to as "quant" in the world of banking) should be very well aware of the (often disturbing) importance of the qualitative, the irrational, the human factor. Having said all that, in the following paragraphs we concentrate only on the quantitative, stochastic side of risk, hence referred to as QRM, and this as summarised in McNeil et al. (2005). Examples will mainly come from the banking world.

The Basic Set-Up: Risk Measures

The first question one needs to consider is whether risk is to be measured in a dynamic way (as a process in time) or in a static, one-period way, say. For reason of space, we concentrate on the latter. Hence risk is modelled initially by a real-valued random variable (rv) X defined on some probability space $(\Omega, \mathcal{F}, P)$

$$X : (\Omega, \mathcal{F}, P) \to \mathbb{R}\,.$$

The distributional properties of X are captured by its distribution function (df) F_X, or F for short,

$$F_X(x) = P(X \le x)\,, \quad \overline{F}_X(x) = P(X > x)\,.$$

For notational convenience, we shall only concentrate on the upper tail $\overline{F} = 1 - F$ of the df F. A risk measure $\mathcal{R}$ now maps the rv X (in our case the df F_X) onto a real number $\mathcal{R}(X)$ satisfying certain axiomatic properties: in applications to finance, these are referred to as the axioms of *coherence* (see Artzner et al. 1999):

(C1) (homogeneity) $\forall \lambda > 0 : \mathcal{R}(\lambda X) = \lambda \mathcal{R}(X)$
(C2) (translation invariance) $\forall a \in \mathcal{R} : \mathcal{R}(X + a) = \mathcal{R}(X) + a$
(C3) (subadditivity) $\mathcal{R}(X_1 + X_2) \le \mathcal{R}(X_1) + \mathcal{R}(X_2)$
(C4) (monotonicity) $X_1 \le X_2 \Rightarrow \mathcal{R}(X_1) \le \mathcal{R}(X_2)$.

A risk measure $\mathcal{R}$ satisfying (C1–C4) is called a coherent risk measure. Especially (C3) is disputed (it is related to the notions of diversification and risk aggregation) and indeed several alternative axioms have been proposed; see for instance Föllmer and Schied (2004). The key questions to be asked for any axiomatic system are:

(Q1) Construct examples satisfying (C1–C4);
(Q2) Do risk measures used in practice satisfy (C1–C4), and
(Q3) How can such risk measures be estimated.

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

Concerning (Q1) one can show that any coherent risk measure is a so called generalized scenario measure; see McNeil et al. (2005, p.244). A commonly used risk measure in the world of banking regulation is the so-called Value-at-Risk measure $\mathrm{VaR}_\alpha(X)$ for $\alpha \in (0,1)$, typically $\alpha \in \{0.99, 0.995, 0.999, 0.9997\}$, i.e., α is close to 1,

$$\mathrm{VaR}_\alpha(X) = F_X^{\leftarrow}(\alpha)$$

where $F_X^{\leftarrow}$ stands for the (generalized) inverse, or quantile function of F_X; see McNeil et al. (2005, p. 38). In general, VaR is *not* coherent. A consequence is that $P(X > \mathrm{VaR}_\alpha(X)) = 1 - \alpha$. It is coherent for so-called elliptical dfs like the multivariate normal or the multivariate Student-t, but coherence (in particular (C3)) typically fails for either very skew, very heavy-tailed (infinite mean models) or multivariate dfs exhibiting a very special dependence structure; see McNeil et al. (2005 pp. 241, 242). A widely used risk measure that is coherent is the so-called Expected Shortfall.

$$\mathrm{ES}_\alpha(X) = E(X \mid X > \mathrm{VaR}_\alpha(x))$$

which for continuous dfs F_X is coherent; see McNeil et al. (2005, p. 243). A very readable paper on this topic is Acerbi and Tasche (2002). An obvious weakness of VaR, as compared to ES, is that it only contains a *frequency* estimate ("once every ...") whereas ES yields *severity* information ("what if ...").

Multivariate Models

So far, I have mainly looked at one-dimensional risk rvs (though in (C3) the two-dimensional distribution of the vector $(X_1, X_2)'$ plays a crucial role). QRM therefore devotes a considerable amount of effort to determining useful models for d-dimensional risk vectors $\mathbf{X} = (X_1, \ldots, X_d)'$. In the case of continuous rvs, key models are:

(M1) The multivariate normal $N_d(\boldsymbol{\mu}, \Sigma)$;
(M2) The multivariate Student-t $t_{\nu,d}(\Sigma)$;
(M3) The class of generalised hyperbolic dfs, and
(M4) The class of elliptical dfs.

The latter class (M4) is a particularly useful one. It is defined (in a slightly restricted form) as follows: suppose $\mathbf{Z} \sim N_d(\mathbf{0}, I_d)$, hence the components $Z_1, \ldots, Z_d$ of $\mathbf{Z}$ are iid $N(0,1)$ and $\boldsymbol{\mu} \in \mathbb{R}^d$, $\mathbb{A}$ is a $d \times d$ matrix, then

$$\mathbf{X} = \boldsymbol{\mu} + \mathbb{A}\mathbf{Z}$$

is called *elliptical*. From a RM point of view, these dfs exhibit several very nice properties, for instance, VaR is coherent in this class; see McNeil et al. (2005, p. 242). Also, linear combinations of the components $X_1 \ldots, X_d$ are of the same type; see McNeil et al. (2005, p. 95). An interesting discussion relevant for applications to finance is to be found in Bingham and Kiesel (2002). As already stated, the world of elliptical models is fully understood (and very well behaved) when it comes to risk management. An interesting research topic concerns "how do such nice properties change if one moves away from ellipticality?"

A final model construction we need to mention here is the so-called *copula* construction; see McNeil et al. (2005) (Chap. 5). A copula is a df C on $[0,1]^d$ with uniform marginals. Suppose now that we are given marginal dfs $F_1, \ldots, F_d$, then the following construction always leads to a valid d-dimensional df with marginals $F_1, \ldots, F_d$:

$$\text{(M5)} \quad F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)).$$

Reading (M5) from left to right yields, for a given df F, at least one copula C such that (M5) holds; whenever the F_i's are continuous, then C (or better C_F) is unique. This is the content of Sklar's Theorem; see for instance McNeil et al. (2005, p. 186), or Nelsen (2007) and Joe (1997). This leads for instance to widely used models like the Gauss- (or normal-) copula or the t-copula. In the former case, F corresponds to a multivariate normal distribution (see ▶Multivariate Normal Distributions), whereas in the latter case F stands for a multivariate t. In (M5) we can inject any copula function on the right-hand side leading to numerous families like the Clayton, Gumbel, Frank, Archimedean, ... copulas. All with their specific properties which may make them useful for specific QRM applications. Since their appearance on the QRM scene in the late 1990s, ▶copulas have attracted a considerable amount of interest; the reader should consult Embrechts (2009) for details on their use and which basic papers to read for a start.

Statistical Estimation

Given a problem where risk has to be estimated, the whole of statistics may enter. However, as many of the risk measures used (like VaR_α and ES_α) concern rare or extreme events, α is close to 1, it is natural that Extreme Value Theory (EVT) plays an important role. Numerous textbooks of a varying degree of complexity yield an introduction to EVT; see for instance McNeil et al. (2005, Chap. 7), Embrechts et al. (1997), Coles (2001) and the references therein. Many other statistical techniques enter at this stage, in particular such fields as rare event estimation/simulation and resampling techniques are very useful here. Just to show that EVT quickly links to further relevant QRM questions, consider the following

$$\lambda_u = \lim_{\alpha \uparrow 1} P(X_2 > \mathrm{VaR}_\alpha(X_2) \mid X_1 > \mathrm{VaR}_\alpha(X_1)),$$

the so-called (upper) asymptotic dependence measure. This number (given that the limit exists, and this is indeed a condition) yields so-called spillover or contagion information from a high loss on X_1 to X_2. It turns out that λ_u *only* depends on the copula of the joint df of X_1 and X_2. Also, for the Gaussian copula model with correlation $\rho < 1$, $\lambda_u = 0$; this mathematically explains why the Gaussian copula model never yields sufficient joint extremal events. The latter issue played a role in the credit crisis starting in 2007 where such models were used to price and hedge complicated credit derivatives; see Li (2001). For the t-copula, $\lambda_u > 0$, hence such a model would allow for joint extremes. At this point, one could combine multivariate EVT with questions of risk aggregation and diversification in QRM, a topic of considerable methodological as well as practical importance; see for instance Embrechts et al. (2009) for a first impression.

An Example: Operational Risk

In order to end with a brief example showing how some of the above may be applied, consider the so-called class of Operational Risk (OR): "OR is defined as the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events. This definition includes legal risk, but excludes strategic and reputational risk." See McNeil et al. (2005, Chap. 10) and Panjer (2006). The so-called Basel II regulatory guidelines for larger international banks require the estimation of risk capital for OR based on a 1-year, 99.9% VaR, i.e., a 1 in 1,000 year event. Internal OR data is typically aggregated over a number d (often 8) of business lines each yielding their OR loss rvs $X_1, \ldots, X_d$ and the corresponding (estimated) risk measures $\mathrm{VaR}_{0.999}(X_1), \ldots, \mathrm{VaR}_{0.999}(X_2)$. At a next step one adds these numbers yielding

$$\mathrm{VaR}_{0.999}^{\mathrm{sum}} = \sum_{i=1}^{d} \mathrm{VaR}_{0.999}(X_i).$$

The bank in question then has the possibility to bring a so–called diversification factor $\delta \in [0,1]$ into account so that the resulting risk capital becomes

$$\mathrm{RC}(\mathrm{OR}) = (1-\delta)\mathrm{VaR}_{0.999}^{\mathrm{sum}}.$$

Of course, non-coherence would yield $\delta < 0$! As such one sees that several of the topics very briefly touched upon enter immediately, and this in a very fundamental way. See the above references for further reading on this topic.

Conclusion

It is clear that the above paragraphs only give a tiny view on the new, emerging field of QRM. A good (quantitative) risk manager has to be master of several trades, including being an excellent communicator. The literature on the field is exploding. The handbook McNeil et al. (2005) is definitely an excellent start on the topic from a more mathematical point of view. The references given no doubt will help in digging deeper and hopefully also yield a much broader view of this fast evolving field.

About the Author

Paul Embrechts is Professor of Mathematics at the ETH Zurich specialising in actuarial mathematics and quantitative risk management. Previous academic positions include the Universities of Leuven, Limburg and London (Imperial College). Dr. Embrechts has held visiting professorships at the University of Strasbourg, ESSEC Paris, the Scuola Normale in Pisa (Cattedra Galileiana), the London School of Economics (Centennial Professor of Finance), the University of Vienna, Paris 1 (Panthéon-Sorbonne), the National University of Singapore and has an Honorary Doctorate from the University of Waterloo. He is an Elected Fellow of the Institute of Mathematical Statistics, Actuary-SAA, Honorary Fellow of the Institute and the Faculty of Actuaries, Corresponding Member of the Italian Institute of Actuaries, Member Honoris Causa of the Belgian Institute of Actuaries and is on the editorial board of numerous scientific journals. He belongs to various national and international research and academic advisory committees. He co-authored the influential books *Modelling of Extremal Events for Insurance and Finance* (Springer, 1997) and *Quantitative Risk Management: Concepts, Techniques and Tools* (Princeton UP, 2005). Dr. Embrechts consults on issues in quantitative risk management for financial institutions, insurance companies and international regulatory authorities.

Cross References

►Actuarial Methods
►Banking, Statistics in
►Copulas in Finance
►Extreme Value Distributions
►Insurance, Statistics in
►Multivariate Statistical Distributions
►Risk Analysis
►Simulation Based Bayes Procedures for Model Structures with Non-Elliptical Posteriors
►Statistical Modeling of Financial Markets
►Statistics of Extremes

References and Further Reading

Acerbi C, Tasche D (2002) On the coherence of expected shortfall. J Banking Finance 26:1487–1503

Artzner P, Delbaen F, Eber J-M, Heath D (1999) Coherent measures of risk. Math Finance 9:203–228

Bingham NH, Kiesel R (2002) Semi-parametric modelling in finance: theoretical foundations. Quan Finance 2: 241–250
Coles S (2001) An introduction to statistical modeling of extreme values. Springer, London
Embrechts P (2009) Copulas: a personal view. J Risk Insur 76: 639–650
Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events for insurance and finance. Springer, Berlin
Embrechts P, Lambrigger DD, Wüthrich MV (2009) Multivariate extremes and the aggregation of dependent risks: examples and counter-examples. Extremes 12:107–127
Föllmer H, Schied A (2004) Stochastic finance: an introduction in discrete time, 2nd edn (de Gruyter Studies in Mathematics 27). Walter de Gruyter, Berlin
Joe H (1997) Multivariate models and dependence concepts. Chapman & Hall, London
Li D (2001) On default correlation: a copula function approach. J Fixed Income 9:43–54
McNeil AJ, Frey R, Embrechts P (2005) Quantitative risk management: concepts, techniques, tools. Princeton University Press, Princeton
Melnick EL, Everitt BS (eds) (2008) Encyclopedia of quantitative risk analysis and assessment, vol 4. Wiley, Chichester
Nelsen RB (2007) An introduction to copulas, 2nd edn. Springer, New York
Panjer HH (2006) Operational risk: modeling analytics. Wiley, London

Questionnaire

Jasna Horvat
Professor, Faculty of Economics in Osijek
J.J. Strossmayer University, Osijek, Croatia

A questionnaire is a helpful tool for collecting a wide range of information from a large number of respondents. The questionnaire was invented by Sir Francis Galton in about 1870. Containing structured groups of questions, it can be used to examine the general characteristics of a population, to compare attitudes of different groups, and to test theories. Questionnaires are any written instruments that present respondents with a series of questions or statements to which they are to react, either by writing out their answers or selecting from among existing answers (Brown 2001). It is important to emphasize that the process of developing a questionnaire involves several steps, starting with problem definition and ending with analysis and interpretation.

A questionnaire should be organized so as to be easy to conduct, fill in, and respond to. It should start with a general introduction to a topic, followed by questions from the least sensitive to the most sensitive. Its structure must in no way influence the responses. A frequently used technique when going from general to specific topics is called the *funnel approach*, as it begins with broader (more general) questions and then asks narrower (more specific) questions, reflecting the shape of a funnel (Grover and Vriens 2006). A questionnaire should ask the "right questions" (by careful use of wording and language) in such a way as to make it easily understandable to the respondent.

There are a great number of questionnaire types. Churchill (2006) classified questionnaires by the method of administration, describing the following types: *personal* (face-to-face conversation between the interviewer and the respondent), *telephone* (the conversation with respondents occurs over the telephone), and *mail questionnaire* (mailing the questionnaire to designated respondents with a covering letter). With the advent of computers and their increasing usage in research processes, the development of different forms of questionnaires adapted to CADAC (Computer Assisted Data Collection) methods has taken place. In other words, every CADAC method requires a different design and logic of the questionnaire (Table 1).

In recent years, there have been publications covering tested measurement instruments that allow researchers to use the most appropriate questionnaire for the defined research problem. Manuals such as *The Handbook of Marketing Scales* or *The Marketing Scale Handbook of Scaling Procedures: Issues and Application* provide invaluable assistance when creating questionnaires. Such handbooks provide an insight into all the new uses of previously developed scales in consumer behavior and advertising, presented along with a description, the origin of the scale, reliability, validity, and other useful information.

Information on respondents collected by a questionnaire is divided into factual, behavioral, and attitudinal information (Dörnyei 2003). This assumption is expanded by *knowledge questions*, the group of information intended to measure respondents' knowledge about a topic.

Factual questions. These query basic information about the respondents, demographic questions (gender, age. . .), and other background information relevant for the interpretation of the questionnaire results (level of education, socioeconomic status, religion, workplace. . .).

Behavioral questions. These are used to investigate current and past behaviors of respondents (lifestyles, habits, and actions).

Attitudinal questions. These depend on the topical problems under investigation and include information about attitudes, opinions, beliefs, interests, and values.

Questionnaire. Table 1 Specific question computer assisted form (de Leeuw and Nicholls II 1996)

Face-to-face questionnaire	CAPI	Computer Assisted Personal Interviewing
Telephone questionnaire	CATI	Computer Assisted Telephone Interviewing
Self-administered questionnaire	CASI	Computer Assisted Self Interviewing
	CSAQ	Computerized Self-Administered Questionnaire
Questionnaire where interviewer is present	CASI of CASIIP	Computer Assisted Self-Interviewing With Interviewer Present
	CASI-V	Question Text On Screen: Visual
	CASI-A	Text On Screen And On Audio
Mail questionnaire	DBM	Disk By Mail
	EMS	Electronic Mail Survey
Panel research questionnaire	CAPAR	Computer Assisted Panel Research
	Teleinterview	Electronic Diaries
Various questionnaires (no interviewer)	TDE	Touchtone Data Entry
	VR	Voice Recognition
	ASR	Automatic Speech Recognition

Knowledge questions. These are used to assess what respondents know about a particular topic, as the level of knowledge.

Forms of questions. Questions are basically divided into closed-ended questions (respondents choose from a set of predetermined answers) and open-ended questions (respondents can answer in their own words).

Closed-ended questions have to be exhaustive and mutually exclusive. They include multiple choice, yes-no answers, and questions with a numerical rating scale (e.g., 1 – strongly disagree, 5 – strongly agree). Closed-ended questionnaires are suitable for processing massive quantities of data, they are easier to answer, and it takes less time to respond to more questions. Respondents pick out answers without any possibility of intervention and need to understand both questions and answers.

Open-ended questions are appropriate in an exploratory phase of research or to obtain specific comments or answers that cannot be expressed as a numerical code. The answers are more difficult to tabulate and analyze, but provide more information and uncover the respondents' knowledge without being reminded. Eventually the answers must be adjusted (open categories are transformed into closed ones) to allow for statistical analysis, but this process is costly, time consuming, and subject to error.

Advantages of questionnaires are primarily in their low costs and relatively quick collection of data from a large portion of a group (particularly compared to face-to-face interviews). The data collected with questionnaires are easy to analyze. Data entry and tabulation can be simply done with various computer software packages. The most popular questionnaires are conducted by large organizations such as the European Commission (Eurobarometer) and UNESCO among others, who regularly publish their research and make it publicly accessible.

Disadvantages of questionnaires. Questionnaires are standardized forms of data collecting, so it is not possible to explain any points in the questions that participants might misinterpret. If researchers inadvertently omit a question, it is not usually possible to go back to respondents, especially if they are anonymous. Open-ended questions can generate large amounts of data that can take a long time to process and analyze.

Sometimes questionnaires can seem impersonal and respondents have even been known to ignore certain questions, particularly if they are not interested in the topic. Another drawback is a potentially low response rate, which can dramatically decrease the confidence in the results. For example, perhaps more responses would be received from participants having a strong opinion on the subject matter and are thereby motivated to respond, while less would be received from potential participants who are indifferent to the topic. The results would not necessarily be accurately representative if applied to the wider

population. Additionally, the results would be less reliable if some questions were misunderstood.

Cross References

►African Population Censuses
►Business Surveys
►Census
►Internet Survey Methodology: Recent Trends and Developments
►Nonresponse in Web Surveys
►Nonsampling Errors in Surveys
►Public Opinion Polls
►Sample Survey Methods
►Statistical Fallacies
►Validity of Scales

References and Further Reading

Bearden WO, Netemeyer RG (1999) Handbook of marketing scales: multi-item measures for marketing and consumer behaviour research, 2nd edn. Sage Publications, Thousand Oaks, CA

Brace I (2008) How to plan, structure and write survey material for effective market research, 2nd edn. Market research practice

Brown JD (2001) Using surveys in language programs. Cambridge University Press, Cambridge

Churchill GA (2006) Marketing research: methodological foundations, 8th edn. Dryden Press, New York

Dörnyei Z (2003) Questionnaires in second language research: construction, administration, and processing. Lawrence Erlbaum, Mahwah, NJ

Grover R, Vriens M (2006) Handbook of marketing research: uses, misuses, and future advances. Sage Publications, Thousand Oaks, CA

de Leeuw E, Nicholls W II (1996) Technological innovations in data collection: acceptance, data quality and costs, Sociological Research Online, vol 1(4)

Imber J, Toffler B (2008) Dictionary of marketing terms, 3rd edn. Barron's Educational Series

Eurobarometer. http://ec.europa.eu/public_opinion/index_en.htm

Queueing Theory

U. Narayan Bhat
Professor Emeritus
Southern Methodist University, Dallas, TX, USA

Queueing is essential to manage congestion in traffic of any type in the modern technological world. This does not mean it is a new phenomenon. More than one hundred years ago, recognizing its importance to telephone traffic, Danish mathematician A.K. Erlang (1909) showed for the first time how probability theory can be used to provide a mathematical model for telephone conversations. From then on, slowly in the first 3 decades, moderately in the next 3 decades, and tremendously in the last 4 decades, the probabilistic approach to modeling queueing phenomena when it is appropriate has grown and contributed significantly to the technological progress. For a historical perspective of the growth of queueing theory see Chapter 1 of Bhat (2008).

Queueing theory describes probabilistically and mathematically the interaction between the arrival process of customers and the service provided to them in order to manage the system in an efficient manner. The term customer is used in a generic sense representing a unit, human or otherwise, demanding service. The unit providing service is known as the server. Some examples of a queueing system are: a communication system with voice and data traffic demanding transmission; a manufacturing system with several work stations; patients arriving at a doctor's office; vehicles requiring service; and so on.

Since the arrival process and service are random phenomena we start with a probabilistic model (also known as a stochastic model) of a queueing system. If we analyze such models using mathematical techniques we can derive its properties that can be used in understanding its behavior and managing it for its efficient use.

In order to build a probabilistic model, first we describe the arrival process (called the input process) using probability distributions. For example, the arrival of customers could be in a Poisson process (see ►Poisson Processes) i.e., the number of customers arriving in a set period of time has a Poisson distribution. Its parameter, say λ, gives the mean number of customers arriving during a unit time. The known distribution now identifies the arrival process. The amount of service provided by the facility is represented by a random variable since it could be random. The distribution of the random variable identifies the service process. When we talk about service we have to take into consideration the mode of service such as service provided with several servers, service provided in a network of servers, etc. Also we must include factors such as queue discipline (e.g., first come, first served (FCFS), also known as first in, first out (FIFO); last come, first served (LCFS or LIFO); group service; priority service; etc.). Another factor that complicates the model is the system capacity, which may be finite or infinite.

Because of the multitude of factors involved in a queueing system, we use a three or four element symbolic representation in discussing various types of systems. The basic structure of the representation is to use symbols or numbers for the three elements: input/service/number of servers. When the system capacity is finite an additional element is added. The commonly used symbols for distributions are: M for Poisson or exponential, E_k

for Erlangian with k phases (gamma distribution with an integer scale parameter k), D for deterministic, and G for a general (also GI for general independent) or an unspecified distribution. Thus $M/G/1$ represents a Poisson arrival, general service, and a single server system, and $M/G/1/N$ has the same description as above with a capacity restriction of N customers in the system.

When the arrival process is represented by a random variable with an index parameter t, define $A(t)$ as the number of customers arriving and $D(t)$ the number of customers leaving the system during a time period $(0, t)$. Let the number of customers in the system at time t be $Q(t)$. Then $Q(t) = A(t) - D(t)$. In order to manage the system efficiently one has to understand how the process $Q(t)$ behaves over time. Note that all $A(t), D(t)$, and $Q(t)$ are stochastic processes (which are sequences of random variables indexed by the time parameter t.) Since the total number of customers leaving the system at t is dependent on the number customers arriving during that time, the mode of their arrival (e.g., there may be time periods with no customers in the system, commonly called idle periods), the service mechanism, queue discipline (when some customers get preferred treatment) and other factors that affect the operation of the system (e.g., service breakdowns), to analyze $Q(t)$, all these factors need to be taken into account in the model.

In the analysis of a queueing system the stochastic process $W(t)$ representing the waiting time of a customer to get served, and the random variable, say B, representing the busy period (the amount of time the system is continuously busy at a stretch) are also used. The objective of the analysis is to get the distributional properties of the stochastic processes $Q(t)$ and $W(t)$ and the random variable B for use in decision making. Analyzing stochastic processes in finite time t often becomes very complex. When the constituent elements such as arrival and service are not time-dependent we can derive the distributions of the limit random variables $Q = \lim_{t\to\infty} Q(t)$ and $W = \lim_{t\to\infty} W(t)$ when they exist. The ratio arrival/service rate is commonly known as the traffic intensity of the queueing system (say ρ). The property $\rho < 1$ is generally the requirement for the existence of the limit distributions of the stochastic processes $Q(t)$ and $W(t)$, when they are time-independent. The behavioral performance measures of interest in a queueing system are transition probability distributions of $Q(t)$ and $W(t)$, probability distributions of Q, W and B, and their means and variances.

In addition to the behavioral problems of underlying stochastic processes mentioned above, we are also interested in inference problems such as estimation and tests of hypotheses regarding basic parameters and performance measures, and optimization problems for assistance in decision making. An introduction to these topics and the necessary references may be found in Bhat (2008).

In order to provide an illustration of the behavioral analysis of a queueing system we consider below a system with Poisson arrivals, exponential service, and single server, symbolically known as an $M/M/1$ queue. This is the simplest and the most used system in applications. As systems include more complicated features more advanced techniques will need to be employed to analyze the corresponding stochastic models.

Let customers arrive in a Poisson process with rate λ. This means that the number $A(t)$ of the customers arriving in $(0, t)$ has a Poisson distribution

$$P[A(t) = j] = e^{-\lambda t}\frac{(\lambda t)^j}{j!},\ j = 0, 1, 2, \ldots$$

It also means that the interarrival times have an exponential distribution with probability density $a(x) = \lambda e^{-\lambda x}$ $(x > 0)$. We assume the service times to have an exponential distribution with probability density $b(x) = \mu e^{-\mu x}$ $(x > 0)$. With these assumptions we have $E[\text{inter-arrival time}] = (1/\lambda) = 1/\text{arrival rate}$ and $E[\text{service time}] = (1/\mu) = 1/\text{service rate}$. The ratio of arrival rate to service rate is the traffic intensity $\rho = \lambda/\mu$. Note that we have assumed the processes to be time-independent.

Let $Q(t)$ be the number of customers in the system at time t and its transition probability distribution be defined as

$$P_{ij}(t) = P[Q(t) = j | Q(0) = i]$$

because of the Poisson arrival process and the exponential service distribution, $Q(t)$ can be modeled as a birth and death process (a class of stochastic processes with major properties (a) probability of more than one state change during an infinitesimal interval of time is close to zero; (b) the rate of change in a unit time is constant and (c) changes occurring in non-overlapping intervals of time are independent of each other) governed by the following difference-differential equations.

$$\begin{aligned} P'_{i0}(t) &= -\lambda P_{i0}(t) + \mu P_{i1}(t) \\ P'_{in}(t) &= -(\lambda + \mu)P_{in}(t) + \lambda P_{i,n-1}(t) \\ &\quad + \mu P_{i,n+1}(t) \quad n = 1, 2, \ldots \end{aligned}$$

with $P_{in}(0) = 1$ when $n = i$ and $= 0$ otherwise. Solving these equations to obtain $P_{in}(t)$ is not very simple. Readers may refer to Gross et al. (2008) and its earlier editions for their solutions.

When $\rho < 1$, the limit $\lim_{t\to\infty} P_{ij}(t) = p_j$ exists and is independent of the initial state i. It can be obtained easily from the following equations that result by letting $t \to \infty$ in the above set of difference-differential equations.

$$\lambda p_0 = \mu p_1$$
$$(\lambda + \mu)p_n = \lambda p_{n-1} + \mu p_{n+1} \quad n = 1, 2, \ldots$$

along with $\sum_{n=0}^{\infty} p_n = 1$. We get $p_0 = 1 - \rho$, $p_n = (1 - \rho)\rho^n (n = 0, 1, 2, \ldots)$. The mean $E(Q)$ and variance $V(Q)$ of this distribution can be obtained as $E(Q) = \rho/(1 - \rho)$ and $V(Q) = \rho/(1 - \rho)^2$.

The waiting time of an arriving customer, when the queue discipline is FCFS, is the total amount of time required to serve the customers who are already in the system and this total time has an Erlangian distribution. Let us denote it as T_q (we use T as the random variable representing the total time the customer is in the system, also known as total workload.) Accordingly we get

$$P[T_q \le t] = 1 - \rho e^{-\mu(1-\rho)t}$$
$$E[T_q] = \rho/\mu(1 - \rho) \text{ and } E[T] = 1/\mu(1 - \rho).$$

Let $E(Q) = L$ and $E[T] = W$. Looking at the above results we can see that $L = \lambda W$ showing how L and W are related in this system. This property is known as Little's Law and it holds in more complex systems under certain general conditions. Another property is the exponential nature of the limiting waiting time distribution shown above which holds in more general queueing systems as well.

The derivation of the distribution of the busy period B is more complicated even in this simple system. We may refer the reader to Gross et al. (2008) for its derivation.

The literature on queueing theory is vast and it is impossible to cover all facets of the analysis of queueing systems using various modeling and sophisticated mathematical techniques in a short article in an encyclopedia. The following references and bibliographies given in them provide the basic understanding of the subject at two levels: Bhat (2008) for those who have a background only in probability and statistics and Gross and Harris (1998) or Gross et al. (2008) for those who have some background in stochastic processes.

About the Author

Professor Bhat retired in May 2005 after serving as SMU's dean of research and graduate studies and professor of statistical science and operations research in Dedman College of Humanities and Sciences at SMU. He joined the SMU faculty in 1969. He received his Ph.D. degree from the University of Western Australia. In addition to teaching, he has served in numerous administrative roles, including chair of several academic departments, vice provost and dean of graduate studies, associate dean for academic affairs and dean ad interim of Dedman College, as well as associate dean of the School of Engineering. Dr. Bhat received the SMU Trustees' Distinguished Service Award in May 2004. He was associate editor for following journals: *Operations Research* (1970–1975), *Management Science* (1969–1974), *OPSEARCH* (1968–1974), and *Queueing Systems: Theory And Applications* (1985–2000). He was President, Omega Rho, The International Honor Society (1994–1996). He has authored and co-authored six books, in addition to technical papers, the latest book being *Introduction to queueing theory: Modeling and Analysis in Applications*, published by Birkhauser in 2008.

Cross References

►Erlang's Formulas
►Geometric and Negative Binomial Distributions
►Markov Processes
►Poisson Processes
►Renewal Processes

References and Further Reading

Bhat UN (2008) An introduction to queueing theory. Birkhauser, Boston
Erlang AK (1909) The theory of probabilities and telephone conversations. Nyt Tidsskrift for Matematik B 20:33
Gross D, Harris CM (1998) Fundamentals of queueing theory, 3rd edn. Wiley, New York
Gross D, Shortle JF, Thompson JM, Harris CM (2008) Fundamentals of queueing theory, 4th edn. Wiley, New York

R

R Language

Robert Gentleman[1], Wolfgang Huber[2],
Vincent J. Carey[3]
[1]Genentech, South San Francisco, CA, USA
[2]Heidelberg, Germany
[3]Brigham and Women's Hospital, Boston, MA, USA

R is a widely used open source language for scientific computing and visualization. It is based on the S language (S: An Interactive Environment for Data Analysis and Graphics, R. A. Becker and J. M. Chambers, Wadsworth, 1984), but with a few paradigms adopted from the Lisp family of languages.

R began its life in 1992, when Ross Ihaka and Robert Gentleman started a project that ultimately evolved into what it is now. In the early days, their main goal was to develop something that was like S, but which had clearer underlying semantics. Around the same time, other major changes were taking place: the world wide web was quickly gaining steam, and a new open source operating system named Linux (with major components from the GNU project) was becoming a popular tool for academic researchers. With these advances, it made sense to make the software more widely available, and hence it needed a name. In part to reflect its heritage, and in part to reflect their contributions, Ross and Robert chose to call it R. At this time, R was still primitive and had restricted capabilities, but a number of other scientists realized its potential, and shortly after its release the R-core group was formed. The activities and hard work of this group of contributors was what really made the breakthrough, and due to their efforts, R quickly become more stable, reliable and forward looking.

R is now under constant development by a team of approximately 20 individuals (essentially members of R-core) and has a fairly consistent 6 month release cycle. The core language is extended through add-on packages which can be obtained and installed in a local version of R, thereby customizing it to a user's interests. These packages are perhaps one of the main advantages of R, since a wide variety of statistical, computational and visualization methods are available. These add-on packages are often written by experts in the methodology, and that has served well to ensure the high quality of the outputs. However, users should realize that the availability of packages on sites such as cran.r-project.org or www.bioconductor.org does not imply any endorsement of their scientific quality except perhaps by their authors. Textbooks, the mailing list archives, and the scientific publications that sometimes accompany a package are good places for users to derive a judgment on packages' suitability for their needs.

R has served to bring scientific computing into many peoples hands. Many statisticians world-wide use R for their teaching and research. R has become widely used in many other fields as well, physics, chemistry, sociology, and notably biology. It is used in a wide variety of industries, Google, Microsoft, various pharmaceutical companies, banks and many investment houses. Its flexibility and the ability to relatively easily code new algorithms is likely to be one of the reasons that R has seen such wide-spread adoption. While reliable estimates of the number of users are hard to obtain increases in download traffic, in frequency of posting to the email help list suggest that the user base is continuing to grow quite rapidly.

R has made scientific computing easier for many sophisticated users and it has also brought in people who are new to the field. Balancing the diverse needs of the community is a particular problem. Discussions on the mailing lists can range from the very philosophical (often surrounding variants of object oriented programming) to the somewhat simpler (but often repeated) bug report that R's numerical capabilities are questionable. For all classes of new users it may be helpful to realize that R has a long history, the actual numerical code used for most of the applications is more than 10 years old and much of it is even older. Large parts of that code have been widely tested for years and it is somewhat unlikely that it fails to perform as intended in any really obvious way (as one person put it, you may be new to R, but R is not new). The fact that R has a long history (somewhat longer than that of Java, for example) means that changes to the way that functions work (even when we know that the original version

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

was not optimal) are not likely to happen – there is simply too much code and too many users invested in the way it currently works and one must have very good reasons to modify code.

New users of R, or any other programming language, that want to do scientific programming should be conversant with the basics of computer arithmetic. To quote from "The Elements of Programming Style" by Kernighan and Plauger: 10.0 times 0.1 is hardly ever 1.0. The issue is one of representing a real number in the allocated memory of a computer, this can only be done exactly for a small subset of all numbers and for all others some rounding is needed. Interested readers should consult a good book on numerical computing and David Goldberg (1991), "What Every Computer Scientist Should Know About Floating-Point Arithmetic," ACM Computing Surveys, 23/1, 5–48, which is available online at: http://docs.sun.com/source/806-3568/ncg_goldberg.html.

As noted above the R language is largely based on the S language that was developed at Bell Laboratories by John Chambers and colleagues during the 1970s and 1980s. While a major goal was providing an interactive environment for performing statistical, and more generally scientific, computing there were other motivations for their work. One of the guiding philosophies of John's work was the notion that scientists needed to use computers to solve problems, and that if the computing environment was suitably conducive they would gradually evolve into being programmers. The main reason is that while there are many commonalities between problems, there is always some need for additional programming and the tweaking of inputs or outputs. Thus, one hopes that the language will actually help to develop the next generation of computational experts by converting some set of its users into programmers. It is worth emphasizing again, that the benefits of having a scientist conversant with, and invested in a method typically means that the method will provide the correct outputs. There is potential for the implementation to be sub-optimal, but that can generally be overcome if the method gets adopted for wide-spread use.

In his book, Algorithms + Data Structures = Programs, N. Wirth describes the fundamental notion that computer programs rely on both a set of algorithms and a set of data structures. R contains a very rich set of algorithms, but of equal importance is the ability of the user to create and use data structures that are appropriate to the problem at hand. R has a very rich and extensible collection of data structures and well designed data structures can greatly simplify many programming problems. Common examples are specific data structures do hold dates and data structures to hold time series objects. Both of these specialized containers are widely used and their use greatly simplifies many programming problems. Specialized methods can be written to deal with the specific implementations and users are then free to worry about other problems (and not converting month–day–year representations into something numeric).

In our work in computational biology we found that the complexity of most experiments was very high and the users were typically spending a great deal of their time doing very basic data management. A very typical example comes from the analysis of microarrays on some set of samples. The arrays provide us with a very large (generally 10s of thousands) of measurements one genes for each sample and at the same time we would have a separate set of data describing the characteristics of the sample. Most users would then spend some time arranging that the order of the columns in the microarray data was the same as that of the rows in the sample characteristic data (somewhat peculiarly microarray data are stored with samples as columns, while most other statistical data is stored with the samples as rows). That is fine until subsets are needed or one decides to do a permutation test (see ►Permutation Tests) for some hypothesis. At that point, depending on whether samples or genes are being permuted different operations are needed. This task is both tedious and has the potential to be done incorrectly in ways that are hard to detect. The rather simple expedient of defining a new data structure that contains both arrays, and where subsetting is defined and implemented to work appropriately greatly simplifies the analysts job and has the additional effect of making it much more likely that the right answer is obtained.

This observation leads us to another arena in which R is taking an important role: that of reproducibility in scientific computing. This issue arises often due to the fact that the analysis of any reasonably large and complex data set is error prone. The chance that mistakes are made, steps omitted increases as the number of people involved in the analysis grows and as the number of software tools increases. A dynamic document is a document that consists of both text and computer code. In greatly simplified terms the document is processed and the computer code is evaluated. An output document is created where each block of computer code is replaced with its output. Typically the computer code is used to produce the figures and tables that are needed for the final document. The final document can then be submitted as a paper to a journal or as an internal report within a group or company. The advantage of the approach is that anyone with access to the raw document and the data can reproduce the document and more importantly they can understand how every figure and table was produced.

Cross References

►Computational Statistics
►Statistical Software: An Overview

Radon–Nikodým Theorem

Takis Konstantopoulos[1], Zurab Zerakidze[2], Grigol Sokhadze[2]
[1]Professor
Heriot-Watt University, Edinburgh, UK
[2]Professor
Javakhishvili Tbilisi State University, Tbilisi, Georgia

The theorem is concerned with the existence of density (derivative) of one measure with respect to another. Let $(\Omega, \mathscr{F})$ be a measurable space, i.e., a set Ω together with a σ-algebra $\mathscr{F}$ of subsets of Ω. Suppose that ν, μ are two σ-finite positive measures on $(\Omega, \mathscr{F})$ such that ν is absolutely continuous (denoted by $\nu \ll \mu$) with respect to μ, i.e., if $\mu(A) = 0$ for some $A \in \mathscr{F}$ then $\nu(A) = 0$. The Radon–Nikodým theorem states that these exists a μ-integrable function $f : \Omega \to \mathbb{R}_+$ such that

$$\nu(A) = \int_A f(\omega)\mu(d\omega), \quad A \in \mathscr{F}.$$

Moreover, f is μ-a.e. unique, in the sense that if f' also satisfies the above then the μ-measure of the points ω such that $f(\omega) \neq f'(\omega)$ equals zero. The function f is called Radon–Nikodým derivative of μ with respect to ν and this is often denoted by

$$f(\omega) = \frac{d\nu}{d\mu}(\omega).$$

The standard proof is as follows. First, assume that $\mu(\Omega) < \infty$. Denote by G the class of all non-negative μ-integrable functions g such that

$$\int_A g(\omega)\mu(d\omega) \le \nu(A), \quad A \in \mathscr{F}.$$

Let c be the supremum of the set numbers $\{\int_\Omega g d\mu : g \in G\}$, and choose a sequence g_n of elements of G such that $\lim_{n\to\infty} \int_\Omega g_n d\mu = \int_\Omega g d\mu$. Observe that if g', g'' are elements of G then so is their maximum $\max(g', g'')$. This observation, together with the monotone convergence theorem, allows us to conclude that $f = \sup_n g_n$ is also a member of G and $\int_\Omega f d\mu = c$. This shows that $\int_A f d\mu \le \nu(A)$ for all $A \in \mathscr{F}$. To show that the difference is actually zero we need to use the Hahn decomposition of a signed measure. Details can be found in Kallenberg (2002, pp. 28–30). The general case for a σ-finite μ follows easily by taking an increasing sequence Ω_n with $\mu(\Omega_n) < \infty$ and $\cup_n \Omega_n = \Omega$, and by applying the previous construction to each Ω_n.

The theorem was proved by Johann Radon (1913) in 1913 for the case $\Omega = \mathbb{R}^n$ and generalized by Otton Nikodým (1930) in 1930 in its present form. The Radon–Nikodým derivative possesses the following properties:

1. Linearity: $\frac{d(c_1\nu_1 + c_2\nu_2)}{2\mu} = c_1\frac{d\nu_1}{d\mu} + c_2\frac{d\nu_2}{d\mu}$, $c_1, c_2 \in \mathbb{R}$.
2. Change of measure: If $\nu \ll \mu$ and g is a ν-integrable function then $\int_\Omega g d\nu = \int_\Omega g\frac{d\nu}{d\mu}d\mu$.
3. Chain rule: If $\lambda \ll \nu \ll \nu$ then $\frac{d\lambda}{d\mu} = \frac{d\lambda}{d\nu}\frac{d\nu}{d\mu}$.
4. Inverse rule: If $\nu \ll \mu$ and $\mu \ll \nu$ then $\frac{d\nu}{d\mu} = \left(\frac{d\mu}{d\nu}\right)^{-1}$.

It is worth noting that a more general statement holds, known as *Lebesgue decomposition*: Let ν, μ be σ-finite measures on $(\Omega, \mathscr{F})$. Then there exists a unique measure $\nu_a \ll \mu$ and a unique measure $\nu_s \perp \mu$ (singular with respect to μ) such that $\nu = \nu_a + \nu_s$.

Also note that the σ-finiteness condition cannot be dropped. For example, if $\Omega = \mathbb{R}$, $\mathscr{F}$ the σ-algebra of Borel sets, μ the counting measure and ν the Lebesgue measure, we certainly have $\nu \ll \mu$ but a density of ν with respect to μ does not exist.

The Radon–Nikodým theorem has numerous applications in many areas of modern mathematics. We mention a few below.

1. Conditional expectation. Let $(\Omega, \mathscr{F}, P)$ be a probability space, X a non-negative random variable with $EX = \int_\Omega X dP < \infty$, and $\mathscr{G}$ a sub-σ-algebra of $\mathscr{F}$. The notion of conditional expectation $E(X|\mathscr{G})$ of X given $\mathscr{G}$ was introduced by A.N. Kolmogorov (1933) in 1933 by means of the Radon–Nikodým derivative as follows. Consider the measure $\nu(A) = \int_A X dP$, $A \in \mathscr{G}$. Clearly, $\nu \ll P$ on $\mathscr{G}$. According to the Radon–Nikodým theorem there is a $\mathscr{G}$-measurable function $E(X|\mathscr{G})$ which satisfies the relation

$$\int_A E(X|\mathscr{G})dP = \int_A X dP, \quad A \in \mathscr{G}.$$

More generally, we define $E(X|\mathscr{G}) = E(X^+|\mathscr{G}) - E(X^-|\mathscr{G})$. For further information see Konstantopoulos (2009).

2. ►Martingales. If P, Q are two probability measures on the same measurable space $(\Omega, \mathscr{F})$ such that $Q \ll P$ then, for any sub-σ-algebra $\mathscr{G} \subset \mathscr{F}$ we have $Q \ll P$ on $(\Omega, \mathscr{G})$. If we denote by $\left(\frac{dQ}{dP}\right)_{\mathscr{F}}$ and $\left(\frac{dQ}{dP}\right)_{\mathscr{G}}$ the two Radon–Nikodým derivatives we have the consistency property $E\left[\left(\frac{dQ}{dP}\right)_{\mathscr{F}}|\mathscr{G}\right] = \left(\frac{dQ}{dP}\right)_{\mathscr{G}}$.

In fact, if $\mathscr{F}_n$ is an increasing sequence of sub-σ-algebras generating $\mathscr{F}$, then $E\left[\left(\frac{dQ}{dP}\right)_{\mathscr{F}}|\mathscr{F}_n\right]$ is a uniformly integrable martingale (Williams 1989) whose limit (a.s. and in L_1) equals $\left(\frac{dQ}{dP}\right)_{\mathscr{F}}$.

3. Kullback–Leibler divergence and Hellinger distance. In Theoretical Statistics, the notion of ▶*Kullback–Leibler divergence* was introduced by Solomon Kullback and Richard Leibler (1951) in 1951. This is a generalization of the notion of ▶entropy for two distributions. If μ and ν are two probability measures on the same space with $\nu \ll \mu$, the Kullback–Leibler divergence or distance

$$Q(\nu \parallel \mu) = -\int_\Omega \log\left(\frac{d\nu}{d\mu}\right) d\mu$$

measures the relative variability of ν with respect to μ. The quantity is always non-negative (owing to the convexity of $-\log$ and Jensen's inequality) but not symmetric. The Kulback–Leibler divergence is used in Information Theory (Cover and Thomas 1991) to define the mutual information between two random variables.

The *Hellinger distance* $H(\mu,\nu)$ between two probability measures μ and ν which are absolutely continuous to a third probability measure λ is defined by

$$H^2(\mu,\nu) = \frac{1}{2}\int_\Omega \left(\sqrt{\frac{d\mu}{d\lambda}} - \sqrt{\frac{d\nu}{d\lambda}}\right)^2 d\lambda$$

and we have $H^2(\mu,\nu) \le Q(\nu \parallel \mu)$.

4. Densities on $\mathbb{R}^n$. Let f, g be densities of two probability measures P, Q, respectively, on $\mathbb{R}^n$. Then $Q \ll P$ if and only if there are versions of f and g such that $\{x : f(x) > 0\} \subset \{x : g(x) > 0\}$. In this case, $\frac{dQ}{dP}(x) = \frac{g(x)}{f(x)}$. For example, if P is the law of n i.i.d. standard normal random variables $(\xi_1,\dots,\xi_n)$, and if Q is the law of $(\xi_1+\mu_1,\dots,\xi_n+\mu_n)$, for some constants $\mu_1,\dots,\mu_n$, then $\frac{dQ}{dP}(x_1,\dots,x_n) = \exp\sum_{j=1}^n\left(\mu_j x_j - \frac{1}{2}\mu_j^2\right)$.

5. The Radon–Nikodým derivative between two Brownian motions. This is an infinite-dimensional generalisation of the previous example. Define the probability measure $P_{T,\mu}$, on the space Ω of continuous functions $\omega : [0,T] \to \mathbb{R}$, to be the law of a Brownian motion (see ▶Brownian Motion and Diffusions) with drift μ and unit variance. We have $P_{T,\mu} \ll P_{T,0}$ and $\frac{dP_{T,\mu}}{dP_{T,0}}(\omega) = e^{\mu\omega(T)-\frac{1}{2}\mu^2 T}$. Moreover, the consistency (martingale) property $E_{T,0}\left[\frac{dP_{T,\mu}}{dP_{T,0}}|\mathscr{F}_t\right] = \frac{dP_{t,\mu}}{dP_{t,0}}$, $t \le T$, holds. Here $\mathscr{F}_t$ is the σ-algebra generated by $(\omega(s), s \le t)$. A further generalisation of this is the *Cameron–Martin–Girsanov theorem* (Cameron and Martin 1944). Let (X_t) be a measurable $(\mathscr{F}_t)$-adapted process such that $Z_t := \exp\left\{\int_0^t X_s dW_s - \frac{1}{2}\int_0^t X_s^2 ds\right\}$ is defined and is a martingale. Define Q on $(\Omega,\mathscr{F}_T)$ by $\frac{dQ}{dP_{T,0}} = Z_T$. Then the law of the process $\left(W_t - \int_0^t X_s ds, 0 \le t \le T\right)$ on $(\Omega,\mathscr{F}_T,Q)$ is again $P_{T,0}$. More general results on the absolute convergence of Gaussian measures and the calculation of a density function are studied, e.g., by Feldman (1958), Ibragimov and Rozanov (1970), Zerakidze (1969) and Yadrenko (1980). Results on diffusions and general processes appear, e.g., in Liptser and Shiryaev (1974), Gikhman and Skorokhod (1971–1975). Smooth measures were studied by Bell (1991), Daletskii and Sokhadze (1988), Bogachev (2008), Kulik and Pilipenko (2000), among others.

6. The Radon–Nikodým derivative between two ▶Poisson processes. Let P_λ be the law of a rate-λ homogeneous Poisson process on a bounded measurable set $S \subset \mathbb{R}^n$ with Lebesgue measure $|S|$. The P_λ is a probability measure on the space Ω of integer-valued random measures with no multiple points. For any $0 < \lambda, \mu < \infty$ we have that $P_\lambda \ll P_\mu$ with Radon–Nikodým derivative

$$\frac{dP_\lambda}{dP_\mu}(\omega) = \left(\frac{\lambda}{\mu}\right)^{\omega(S)} \cdot e^{-(\lambda-\mu)|S|}.$$

To see this, it is sufficient to show that for any bounded measurable $f : \Omega \to \mathbb{R}$ we have $E_\lambda[\exp\int_S f(x)\omega(dx)] = E_\mu\left[\left(\frac{\lambda}{\mu}\right)^{\omega(S)} \cdot e^{-(\lambda-\mu)|S|} \cdot \exp\int_S f(x)\omega(dx)\right]$, something that is easily verifiable by means of the Poisson characteristic functional $E_\lambda[\exp\int_S f(x)\omega(dx)] = \exp\lambda\int_S(e^{f(x)} - 1)dx$. Note that if $\widehat{P}_\lambda$ is the image of P_λ on $\mathbb{Z}_+$ under the mapping $\omega \mapsto \omega(S)$ then the formula above says that $\frac{dP_\lambda}{dP_\mu}(\omega) = \frac{d\widehat{P}_\lambda}{d\widehat{P}_\mu}(\omega(S))$. We also note that if S is not bounded, e.g., if $S = \mathbb{R}^n$, the above fails to hold because $P_\lambda \perp P_\mu$ if $\lambda \ne \mu$.

7. The Radon–Nikodým derivative between two Markov jump processes. Consider a Markov jump process in a countable state space S with transition rates $q_{x,y}$ such that $q(x) := -q(x,x) = \sum_y q_{x,y} < \infty$ for all $x \in S$, and initial distribution μ. Let Q be the matrix with entries $q_{x,y}$. In other words, Q and μ define a probability measure $P_{\mu,Q}$ on the space Ω of right-continuous piecewise-constant functions $\omega : [0,T] \to S$. We only consider finite time horizon T. We change μ, Q to $\widetilde{\mu}, \widetilde{Q}$ in a way that $\mu \ll \widetilde{\mu}$ and $q_{x,y} = 0$ whenever $\widetilde{q}_{x,y} = 0$. Then $P_{\mu,Q} \ll P_{\widetilde{\mu},\widetilde{Q}}$ and

$$\frac{dP_{\mu,Q}}{dP_{\widetilde{\mu},\widetilde{Q}}}(\omega) = \frac{\mu(\omega(0))}{\widetilde{\mu}(\omega(0))}\exp\left\{-\int_0^T (q(\omega(s)) - \widetilde{q}(\omega(s)))ds\right\} \cdot \prod_{x\ne y}\left(\frac{q_{x,y}}{\widetilde{q}_{x,y}}\right)^{N_T(\omega,x,y)},$$

where $N_T(\omega, x, y)$ is the total number of points $s \le T$ such that $\omega(s-) = x, \omega(s) = y$.

8. The Esscher transform. Let $(X_t, t \ge 0)$ be a Lévy process (see ►Lévy Processes), i.e. a stochastic process with values in $\mathbb{R}$ which is continuous in probability and has stationary-independent increments. Assume that the Laplace exponent $\psi(\beta) = \log E \exp(\beta X_1)$ is defined for β belonging to a non-trivial interval. Let $Z_t^\beta := \exp\{\beta X_t - \psi(\beta)t\}$ and define a new measure P^β via the Radon–Nikodým derivative $\frac{dP^\beta}{dP}\big|_{\mathscr{F}_t} = Z_t^\beta$, where $\mathscr{F}_t = \sigma(X_s, s \le t)$. This derivative is known as the *Esscher transform* and leads to a natural generalisation of the Cameron–Martin–Girsanov theorem: The process (X_t) is still a Lévy process under P^β. See Kyprianou (2006) for its use in Fluctuation Theory.

9. Palm probability. Let $(\Omega, \mathscr{F}, P)$ be a probability space and $M : (\Omega \times \mathbb{R}^d) \to \mathbb{R}_+$ be measurable in the first argument and a locally finite probability measure in the second. We call such an M a random measure on $\mathbb{R}^d$. Assume that $\lambda(B) = EM(B)$ is a locally finite measure. Define the Campbell measure $C(A, B) = E[\mathbf{1}_A M(B)]$, $A \in \mathscr{F}, B \in \mathscr{B}$, where $\mathscr{B}$ is the class of Borel sets on $\mathbb{R}^d$, and observe that $C(A, \cdot) \ll \lambda$ for each $A \in \mathscr{F}$. The Radon–Nikodým derivative $P^x(A) = \frac{dC(A,\cdot)}{d\lambda}(x)$ has a version which is a probability measure on $(\Omega, \mathscr{F})$ and is called *Palm probability*. If M is a simple point process (see ►Point Processes), i.e. $M(\omega, \cdot)$ takes values in $\mathbb{Z}_+$ such that $M(\omega, \{x\}) \in \{0, 1\}$, for all x and ω, then $P^x(A)$ gives the probability of A given that M places a unit mass at the point x. The concept is most useful for stationary random measures (Kallenberg 2002).

About the Authors

Takis Konstantopoulos is Professor of Probability in the School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh. He is also a member of the Maxwell Institute for Mathematical Sciences. He received a BSc from the National Technical University of Athens, Greece and a M.Sc. and Ph.D. (1989) from the University of California, Berkeley, USA. He has held a charge de recherche position in INRIA, Sophia Antipolis, France. He has served as a faculty member in the University of Texas at Austin in Electrical Engineering and in Mathematics. He has been Professor of Mathematics (2002–2005) at the University of Patras Greece where he also served as Director of the Probability and Statistics division. His research interests include stochastic processes, limit theorems, stochastic networks, applied probability and pure mathematics. He is advisor of the Centre for Education in Sciences (Patras, Greece). He is an organizer of an international program on "stochastic processes in communication sciences" at the Newton Institute for Mathematical Sciences, Cambridge, UK, January–July 2010.

Professor Zerakidze was Head of the Higher Mathematics department of Gori State University. He was also Head of a Mathematical Society of Gori region (2005–2007). He has been awarded with the Order of Honour by the President of Georgia (March 17th 2000). His work "The divisible family of measures" was included in the Big Russian Encyclopedia in the section "Probability and the mathematical statistics" page 533, Moscow, 1999.

Professor Grigol Sokhadze received his Ph.D. in Mathematics in 1992. He has (co-)authored about 100 papers in probability and statistics.

Cross References

►Brownian Motion and Diffusions
►Conditional Expectation and Probability
►Entropy
►Entropy and Cross Entropy as Diversity and Distance Measures
►Kullback-Leibler Divergence
►Martingales
►Measure Theory in Probability
►Poisson Processes
►Stochastic Processes: Applications in Finance and Insurance

References and Further Reading

Bell D (1991) Transformations of measures on an infinite-dimensional vector space. Seminar on stochastic processes (Vancouver 1990). Prog Probab 24:15–25
Bogachev V (2008) Differentiable measures and the malliavin calculus (in Russian). R & C Dynamics, Moscow
Cameron RH, Martin WT (1944) Transformation of Wiener integrals under translations. Ann Math 45:386–396
Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York
Daletskii J, Sokhadze G (1988) Absolute continuity of smooth measures (in Russian). Funct Anal Appl 22(2):77–88
Feldman I (1958) Equivalence and perpendicularity of Gaussian processes. Pac J Math 8:699–708
Gikhman I, Skorokhod A (1971–1975) Theory of stochastic processes, vol 1–3. Nauka, Moscow (in Russian)
Ibragimov I, Rozanov J (1970) Gaussian random processes (in Russian). Nauka, Moscow
Kallenberg O (2002) Foundations of modern probability, 2nd edn. Springer, New York
Kolmogorov A (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. Julius Springer, Berlin. (English translation by Chelsea, New York, 1956)
Konstantopoulos T (2009) Conditional expectation and probability. This encyclopedia

Kulik A, Pilipenko A (2000) Nonlinear transformations of smooth measures in infinite-dimensional spaces. Ukran Math J 52(9):1226–1250
Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86
Kyprianou AE (2006) Introductory lectures on fluctuations of Lévy processes with applications. Springer, Heidelberg
Liptser R, Shiryaev A (1974) Statistics of random processes (in Russian). Nauka, Moscow
Nikodým O (1930) Sur une généralisation des intégrales de M. J Radon Fundamenta Mathematicae 15:131–179
Radon J (1913) Theorie und Anwendungen der absolut additiven Mengenfunktionen. Sitzber, der Math.Naturwiss. Klasse der Kais. Akademie der Wiss. Wien, 112 Bd. Abt II a/2
Williams D (1989) Probability with martingales. Cambridge University Press, Cambridge
Yadrenko M (1980) Spectral theory of random fields (in Russian). Visha Shkola, Kiev
Zerakidze Z (1969) On the equivalence of distributions of Gaussian fields (in Russian). In: Proceedings of the Tbilisi institute of applied mathematics. Tbilisi, vol 2, pp 215–220

Random Coefficient Models

Nicholas T. Longford
Universitat Pompeu Fabra, Barcelona, Spain

Independence of the observations is a key assumption of many standard statistical methods, such as ▶analysis of variance (ANOVA) and ordinary regression, and some of its extensions. Common examples of data structures that do not fit into such a framework arise in longitudinal analysis, in which observations are made on subjects at subject-specific sequences of time points, and in studies that involve subjects (units) ocurring naturally in clusters, such as individuals within families, schoolchildren within classrooms, employees within companies, and the like. The assumption of independence of the observations is not tenable, because observations within a cluster are likely to be more similar than observations in general. Such similarity can be conveniently represented by a positive correlation (dependence).

This section describes an adaptation of the ordinary regression for clustered observations. Such observations require two indices, one for elements within clusters, $i = 1, \ldots, n_j$, and another for clusters, $j = 1, \ldots, m$. Thus, we have $n = n_1 + \cdots + n_m$ elementary units and m clusters. The ordinary regression model

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \varepsilon_{ij}, \tag{1}$$

with the usual assumptions of normality, independence and equal variance (homoscedasticity) of the deviations $\varepsilon_{ij}, \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, i.i.d., implies that the regressions within the clusters j have a common vector of coefficients $\boldsymbol{\beta}$. This restriction can be relaxed by allowing the regressions to differ in their intercepts. A practical way of defining such a model is by the equation

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \delta_j + \varepsilon_{ij}, \tag{2}$$

where δ_j is a random sample from a centred normal distribution, $\delta_j \sim \mathcal{N}\left(0, \sigma_B^2\right)$, i.i.d., independent from the ε's. With this model, the within-cluster regressions are parallel; their intercepts are $\beta_0 + \delta_j$, but the coefficients on all the other variables in $\mathbf{x}$ are common to the clusters. A more appealing interpretation of the model is that observations in a cluster are correlated,

$$\operatorname{cor}\left(y_{i_1,j}, y_{i_2,j}\right) = \frac{\sigma_B^2}{\sigma^2 + \sigma_B^2},$$

because they share the same deviation δ_j. Further relaxation of how the within-cluster regressions differ is attained by allowing some (or all) the regression slopes to be specific to the clusters. We select a set of variables in $\mathbf{x}$, denoted by $\mathbf{z}$, and assume that the regressions with respect to these variables differ across the clusters, but are constant with respect to the remaining variables;

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\boldsymbol{\delta}_j + \varepsilon_{ij}, \tag{3}$$

where $\boldsymbol{\delta}_j$ is a random sample from a multivariate normal distribution (see ▶Multivariate Normal Distributions) $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_B)$, independent from the ε's. We say that the variables in $\mathbf{z}$ are associated with (cluster-level) variation. The variance of an observation y_{ij}, without conditioning on the cluster j, is

$$\operatorname{var}\left(y_{ij}\right) = \sigma^2 + \mathbf{x}_{ij}\boldsymbol{\Sigma}_B\mathbf{x}_{ij}^{\top}.$$

We refer to σ^2 and $\mathbf{z}_{ij}\boldsymbol{\Sigma}_B\mathbf{z}_{ij}^{\top}$ as the *variance components* (at the elementary and cluster levels, respectively). The principle of invariance with respect to linear transformations of $\mathbf{z}$ implies that the intercept should always be included in $\mathbf{z}$, unless $\mathbf{z}$ is empty, as in the model in (1). The function $V(\mathbf{z}) = \mathbf{z}\boldsymbol{\Sigma}_B\mathbf{z}^{\top}$, over the feasible values of $\mathbf{z}$, defines the *pattern* of variation, and it can be described by its behaviour (local minima, points of inflection, and the like). By way of an example, suppose $\mathbf{z}$ contains the intercept and a single variable z. Denote the variances in $\boldsymbol{\Sigma}_B$ by σ_0^2 and σ_z^2, and the covariance by σ_{0z}. Then

$$V(\mathbf{z}) = \sigma_0^2 + 2z\sigma_{0z} + z^2\sigma_z^2, \tag{4}$$

and this quadratic function has a unique minimum at $z^* = -\sigma_{0z}/\sigma_z^2$, unless $\sigma_z^2 = 0$, in which case we revert to the model in (2) in which $V(\mathbf{z})$ is constant.

The model in (3) is fitted by maximum likelihood (ML) which maximizes the log-likelihood function

$$l\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}_B\right) = -\frac{1}{2}\sum_{j=1}^{m}\Big[\log\left\{\det\left(\mathbf{V}_j\right)\right\} + \left(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}\right)^{\top}\mathbf{V}_j^{-1}\left(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}\right)\Big],$$

in which $\mathbf{V}_j$ is the variance matrix of the observations in cluster j, $\mathbf{y}_j$ the vector of the outcomes for the observations in cluster j, and $\mathbf{X}_j$ the corresponding regression design matrix formed by vertical stacking of the rows $\mathbf{x}_{ij}$, $i = 1, \ldots, n_j$. The variation design matrices $\mathbf{Z}_j, j = 1, \ldots, m$, are defined similarly; with them, $\mathbf{V}_j = \sigma^2\mathbf{I}_{n_j} + \mathbf{Z}_j\boldsymbol{\Sigma}_B\mathbf{Z}_j^{\top}$, where $\mathbf{I}_{n_j}$ is the $n_j \times n_j$ identity matrix. For ML solutions, see Longford (1993) and Goldstein (2000). These and other algorithms are implemented in most standard statistical packages.

►Model selection entails two tasks, selecting a set of variables to form $\mathbf{x}$ and selecting its subset to form $\mathbf{z}$. The variables in $\mathbf{x}$ can be defined for elements or clusters; the latter can be defined as being constant within clusters. Inclusion of cluster-level variables in $\mathbf{z}$ does not have an interpretation in terms of varying regression coefficients, so associating them with variation is in most contexts not meaningful. However, the identity in (4) and its generalisations for $\boldsymbol{\Sigma}_B$ with more than two rows and columns indicate that $\mathbf{z}$ can be used for modelling variance heterogeneity. The likelihood ratio test statistic and various information criteria can be used for selecting among alternative models, so long as one is a submodel of the other; that is, the variables in both $\mathbf{x}$ and $\mathbf{z}$ of one model are subsets of (or coincide with) their counterparts in the other model.

Random coefficients can be applied to a range of models much wider than ordinary regression. In principle, we can conceive any *basis model*, characterized by a vector of parameters, which applies to every cluster. A subset of these parameters is constant across the clusters and the remainder varies according to a model for cluster-level variation. The latter model need not be a multivariate normal distribution, although suitable alternatives to it are difficult to identify. The basis model itself can be complex, such as a random coefficient model itself. This gives rise to three- or, generally, *multilevel models*, in which elements are clustered within two-level units, these units in three-level units, and so on. Generalized linear mixed models have ►generalized linear models (McCullagh and Nelder 1989) as their basis.

Random coefficient models are well suited for analysing surveys in which clusters arise naturally as a consequence of the organisation (design) of the survey and the way the studied population is structured. They can be applied also in settings in which multiple observations are made on subjects, as in longitudinal studies (Molenberghs and Verbeke 2000). In some settings it is contentious as to whether the clusters should be regarded as fixed or random. When they are assumed to be random the (random coefficient) models are often more parsimonious than their fixed-effects (ANCOVA) models, because the number of parameters involved does not depend on the number of clusters.

About the Author

Dr. Longford has been a visiting lecturer and visiting Associate Professor in Spain, Sudan, Germany, Denmark, USA, Brazil and New Zealand. He was a President of the Princeton-Trenton (NJ) Chapter of ASA.

Cross References

►Cross Classified and Multiple Membership Multilevel Models
►Linear Mixed Models
►Multilevel Analysis
►Sensometrics
►Statistical Analysis of Longitudinal and Correlated Data
►Testing Variance Components in Mixed Linear Models

References and Further Reading

Goldstein H (2000) Multilevel statistical models, 2nd edn. Edward Arnold, London
Longford NT (1993) Random coefficient models. Oxford University Press, Oxford
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London
Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer, New York

Random Field

Mikhail P. Moklyachuk
Professor
Kyiv National Taras Shevchenko University, Kyiv, Ukraine

Random field $X(t)$ on $D \subset \mathbb{R}^n$ (i.e., $t \in D \subset \mathbb{R}^n$) is a function whose values are random variables for any $t \in D$. The dimension of the coordinate is usually in the range from one to four, but any $n > 0$ is possible. A one-dimensional

random field is usually called a stochastic process. The term "random field" is used to stress that the dimension of the coordinate is higher than one. Random fields in two and three dimensions are encountered in a wide range of sciences and especially in the earth sciences, such as hydrology, agriculture, and geology. Random fields where t is a position in space-time are studied in turbulence theory and in meteorology.

Random field $X(t)$ is described by its finite-dimensional (cumulative) distributions

$$F_{t_1,\ldots,t_k}(x_1,\ldots,x_k) = P\{X(t_1) < x_1,\ldots,X(t_k) < x_k\}, k = 1,2\ldots$$

The cumulative distribution functions are by definition left-continuous and nondecreasing. Two requirements on the finite-dimensional distributions must be satisfied. The symmetry condition

$$F_{t_1,\ldots,t_k}(x_1,\ldots,x_k) = F_{t_{\pi 1},\ldots,t_{\pi k}}(x_{\pi 1},\ldots,x_{\pi k}),$$

where π is a permutation of the index set $\{1,\ldots,k\}$. The compatibility condition

$$F_{t_1,\ldots,t_{k-1}}(x_1,\ldots,x_{k-1}) = F_{t_1,\ldots,t_k}(x_1,\ldots,x_{k-1},\infty).$$

Kolmogorov Existence Theorem states: If a system of finite-dimensional distributions $F_{t_1,\ldots,t_k}(x_1,\ldots,x_k)$, $k = 1,2\ldots$, satisfies the symmetry and compatibility conditions, then there exists on some probability space a random field $X(t)$, $t \in D$, having $F_{t_1,\ldots,t_k}(x_1,\ldots,x_k)$, $k = 1,2\ldots$, as its finite-dimensional distributions.

The expectation (mean value) of a random field is by definition the Stieltjes integral

$$m(t) = EX(t) = \int_{\mathbb{R}^1} x dF_t(x).$$

The (auto-)covariance function is also expressed as the Stieltjes integral

$$B(t,s) = E(X(t)X(s)) - m(t)m(s) = \iint_{\mathbb{R}^2} xy dF_{t,s}(x,y) - m(t)m(s),$$

whereas the variance is $\sigma^2(t) = B(t,t)$.

Gaussian random fields play an important role due to several reasons: the specification of their finite-dimensional distributions is simple, they are reasonable models for many natural phenomena, and their estimation and inference are simple.

A Gaussian random field is a random field where all the finite-dimensional distributions are ▶multivariate normal distributions. Since multivariate normal distributions are completely specified by expectations and covariances, it suffices to specify $m(t)$ and $B(t,s)$ in such a way that the symmetry condition and the compatibility condition hold true. The expectation can be arbitrarily chosen, but the covariance function must be positive-definite to ensure the existence of all finite-dimensional distributions.

Wiener sheet (*Brownian sheet*) is a Gaussian random field $W(t)$, $t = (t_1,t_2) \in \mathbb{R}^2_+$ with $EW(t) = 0$ and correlation function $B(t,s) = E(X(t)X(s)) = \min\{s_1,t_1\}\min\{s_2,t_2\}$. Analogously, the n-parametric Wiener process is a Gaussian random field $W(t)$, $t \in \mathbb{R}^n_+$ with $EW(t) = 0$ and correlation function $B(t,s) = \prod_{i=1}^n \min(s_i,t_i)$. The multiparametric Wiener process $W(t)$ has independent homogeneous increments. A generalized derivative of the multiparametric Wiener process $W(t)$ is the *Gaussian white noise process* on $\mathbb{R}^n_+$ (Chung and Walsh 2005).

Poisson random fields are also reasonable models for many natural phenomena. A Poisson random field is an integer-valued (point) random field where the (random) amount of points that belong to a bounded set from the range of values of the field has a Poisson distribution and the random amounts of points that belong to nonoverlapping sets are mutually independent (Kerstan et al. 1974).

Markov random field $X(t)$, $t \in D \subset \mathbb{R}^n$, is a random function that has the Markov property with respect to a fixed system of ordered triples (S_1,Γ,S_2) of nonoverlapping subsets from the domain of definition D. The Markov property means that for any measurable set B from the range of values of the function $X(t)$ and every $t_0 \in S_2$, the following equality holds true:

$$P\{X(t_0) \in B | X(t), t \in S_1 \cup \Gamma\} = P\{X(t_0) \in B | X(t), t \in \Gamma\}.$$

This means that the future S_2 does not depend on the past S_1 when the present Γ is given. Let, for example, $D = \mathbb{R}^n$, $\{\Gamma\}$ be a family of all spheres in $\mathbb{R}^n$, S_1 be the interior of Γ, and S_2 be the exterior of Γ. A homogeneous and isotropic Gaussian random field $X(t)$, $t \in \mathbb{R}^n$, has the Markov property with respect to the ordered triples (S_1,Γ,S_2) if and only if $X(t) = \xi$, where ξ is a random variable. Nontrivial examples of homogeneous and isotropic Markov random fields can be constructed when considering the generalized random fields. Markov random fields are completely described in the class of homogeneous Gaussian random fields on $\mathbb{Z}^n$, in the class of multidimensional homogeneous generalized Gaussian random fields on the space $\mathbb{C}_0^\infty(\mathbb{R}^m)$ and the class of multidimensional homogeneous and isotropic generalized Gaussian random fields (Glimm and Jaffe 1981; Rozanov 1982; Yadrenko 1983).

Gibbs random fields form a class of random fields that have extensive applications in solutions of problems in statistical physics. The distribution functions of these

fields are determined by Gibbs distribution (Malyshev and Minlos 1985).

Homogeneous random field in the strict sense is a real-valued random function $X(t)$, $t \in \mathbb{R}^n$ (or $t \in \mathbb{Z}^n$), where all its finite-dimensional distributions are invariant under arbitrary translations, that is,

$$F_{t_1+s,\ldots,t_k+s}(x_1,\ldots,x_k) = F_{t_1,\ldots,t_k}(x_1,\ldots,x_k)\ \forall s \in \mathbb{R}^n.$$

Homogeneous random field in the wide sense is a real-valued random function $X(t)$, $t \in \mathbb{R}^n$ ($t \in \mathbb{Z}^n$), $E|X(t)|^2 < +\infty$, where $EX(t) = m = \text{const.}$ and the correlation function $EX(t)X(s) = B(t-s)$ depends on the difference $t-s$ of coordinates of points t and s.

Homogeneous random field $X(t)$, $t \in \mathbb{R}^n$, $EX(t) = 0$, $E|X(t)|^2 < +\infty$, and its correlation function $B(t) = EX(t+s)X(s)$ admit the spectral representations

$$X(t) = \int \cdots \int \exp\left\{\sum_{k=1}^{n} t_k\lambda_k\right\} Z(d\lambda),$$

$$B(t) = \int \cdots \int \exp\left\{\sum_{k=1}^{n} t_k\lambda_k\right\} F(d\lambda),$$

where $F(d\lambda)$ is a measure on the Borel σ-algebra B_n of sets from $\mathbb{R}^n$, and $Z(d\lambda)$ is an orthogonal random measure on B_n such that $EZ(S_1)Z(S_2) = F(S_1 \cap S_2)$. The integration range is $\mathbb{R}^n$ in the case of continuous time random field $X(t)$, $t \in \mathbb{R}^n$, and $[-\pi,\pi]^n$ in the case of discrete time random field $X(t)$, $t \in \mathbb{Z}^n$. In the case where the spectral representation of the correlation function is of the form

$$B(t) = \int \cdots \int \exp\left\{\sum_{k=1}^{n} t_k\lambda_k\right\} f(\lambda)d\lambda,$$

the function $f(\lambda)$ is called the spectral density of the field $X(t)$. Based on these spectral representations we can prove, for example, the *law of large numbers* for random field $X(t)$:

The mean square limit

$$\lim_{N\to\infty} \frac{1}{(2N+1)^n} \sum_{|t_i|\le N, i=1,\ldots,n} X(t) = Z\{0\}.$$

This limit is equal to $EX(t) = 0$ if and only if $E|Z\{0\}|^2 = F\{0\}$. In the case where $F\{0\} = 0$ and

$$\int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \prod_{i=1}^{n} \log\left|\log\frac{1}{|\lambda_i|}\right| F(d\lambda) < +\infty,$$

the *strong law of large numbers* holds true for the random field $X(t)$.

Isotropic random field is a real-valued random function $X(t)$, $t \in \mathbb{R}^n$, $E|X(t)|^2 < +\infty$, where the expectation and the correlation function have properties $EX(t) = EX(gt)$ and $EX(t)X(s) = EX(gt)X(gs)$ for all rotations g around the origin of coordinates. An isotropic random field $X(t)$ admits the decomposition

$$X(t) = \sum_{m=0}^{\infty} \sum_{l=1}^{h(m,n)} X_m^l(r) S_m^l(\theta_1,\theta_2,\ldots,\theta_{n-2},\varphi),$$

where $(r,\theta_1,\theta_2,\ldots,\theta_{n-2},\varphi)$ are spherical coordinates of the point $t \in \mathbb{R}^n$, $S_m^l(\theta_1,\theta_2,\ldots,\theta_{n-2},\varphi)$ are spherical harmonics of the degree m, $h(m,n)$ is the amount of such harmonics, $X_m^l(r)$ are uncorrelated stochastic processes such that $EX_m^l(r)X_{m_1}^{l_1}(s) = b_m(r,s)\delta_m^{m_1}\delta_l^{l_1}$, where δ_i^j is the Kronecker symbol, $b_m(r,s)$ is a sequence of positive definite kernels such that $\sum_{m=0}^{\infty} h(m,n)b_m(r,s) < +\infty$, $b_m(0,s) = 0, m \ne 0$.

Isotropic random field $X(t)$, $t \in \mathbb{R}^2$, on the plane admits the decomposition

$$X(r,\varphi) = \sum_{m=0}^{\infty} \left\{X_m^1(r)\cos(m\varphi) + X_m^2(r)\sin(m\varphi)\right\}.$$

The class of isotropic random fields includes homogeneous and isotropic random fields, multiparametric Brownian motion processes (see ▶Brownian Motion and Diffusions).

Homogeneous and isotropic random field is a real-valued random function $X(t)$, $t \in \mathbb{R}^n$, $E|X(t)|^2 < +\infty$, where the expectation $EX(t) = c = \text{const.}$ and the correlation function $EX(t)X(s) = B(|t-s|)$ depends on the distance $|t-s|$ between points t and s. Homogeneous and isotropic random field $X(t)$ and its correlation function $B(r)$ admit the spectral representations (Rozanov 1982; Yadrenko 1983; Yaglom 1987)

$$X(t) = c_n \sum_{m=0}^{\infty} \sum_{l=1}^{h(m,n)} S_m^l(\theta_1,\theta_2,\ldots,\theta_{n-2},\varphi) \int_0^{\infty} \frac{J_{m+(n-2)/2}(r\lambda)}{(r\lambda)^{(n-2)/2}} Z_m^l(d\lambda),$$

$$B(r) = \int_0^{\infty} Y_n(r\lambda)d\Phi(\lambda),$$

where

$$Y_n(x) = 2^{(n-2)/2}\Gamma\left(\frac{n}{2}\right)\frac{J_{(n-2)/2}(x)}{x^{(n-2)/2}}$$

is a spherical Bessel function, $\Phi(\lambda)$ is a bounded nondecreasing function called the spectral function of the field $X(t)$, $Z_m^l(d\lambda)$ are random measures with orthogonal values such that $EZ_m^l(S_1)Z_{m_1}^{l_1}(S_2) = \delta_m^{m_1}\delta_l^{l_1}\Phi(S_1 \cap S_2)$, $c_n^2 = 2^{n-1}\Gamma(n/2)\pi^{n/2}$.

Homogeneous and isotropic random field $X(t), t \in \mathbb{R}^2$, on the plane admits the spectral representation

$$X(t,\varphi) = \sum_{m=0}^{\infty} \cos(m\varphi) Y_m(r\lambda) Z_m^1(d\lambda) + \sum_{m=1}^{\infty} \sin(m\varphi) Y_m(r\lambda) Z_m^2(d\lambda).$$

These spectral decompositions of random fields form a power tool for the solution of statistical problems for random fields such as extrapolation, interpolation, filtering, and estimation of parameters of the distribution (Yadrenko 1983; Yaglom 1987a, b).

About the Author

Dr. Mikhail P. Moklyachuk is a Professor of the Department of Probability Theory, Statistics and Actuarial Mathematics, Kyiv National Taras Shevchenko University, Ukraine. He is the author and coauthor of more than 100 papers and six books, including *Robust estimates for functionals of stochastic processes* (Kyiv University Press, 2008). Professor Moklyachuk has received the Taras Shevchenko prize (Kyiv University best textbook award, 1999) for the textbook *Variational Calculus. Extremum Problems*. He is the editor of the Cooperation Unit of Zentralblatt MATH (Zentralblatt fuer Mathematik/Mathematics Abstracts), coeditor of *Current Index to Statistics*, and member of the editorial board, *Theory of Probability and Mathematical Statistics*.

Cross References

►Estimation Problems for Random Fields
►Measure Theory in Probability
►Model-Based Geostatistics
►Random Variable
►Spatial Statistics
►Stochastic Processes

References and Further Reading

Chung KL, Walsh JB (2005) Markov processes, Brownian motion, and time symmetry, 2nd ed. Springer, New York, NY
Glimm J, Jaffe A (1981) Quantum physics: a functional integral point of view. Springer, Berlin/Heidelberg/New York
Kerstan J, Matthes K, Mecke J (1974) Mathematische Lehrbücher und Monographien. II. Abt. Mathematische Monographien. Band XXVII. Akademie, Berlin
Malyshev VA, Minlos RA (1985) Stochastic Gibbs fields. The method of cluster expansions. Nauka, Moskva
Monin AS, Yaglom AM (2007a) Statistical fluid mechanics: mechanics of turbulence, volume I. Edited and with a preface by Lumley JL, Dover, Mineola, NY
Monin AS, Yaglom AM (2007b) Statistical fluid mechanics: mechanics of turbulence, volume II. Edited and with a preface by Lumley JL, Dover, Mineola, NY
Rozanov YuA (1982) Markov random fields. Springer, New York
Yadrenko MI (1983) Spectral theory of random fields. Translation Series in Mathematics and Engineering. Optimization Software, Publications Division, New York; Springer, New York
Yaglom AM (1987a) Correlation theory of stationary and related random functions. volume I. Basic results. Springer Series in Statistics. Springer, New York
Yaglom AM (1987b) Correlation theory of stationary and related random functions, volume II. Supplementary notes and references. Springer Series in Statistics. Springer, New York

Random Matrix Theory

Jack W. Silverstein
Professor
North Carolina State University, Raleigh, NC, USA

Random matrix theory (RMT) originated from the investigation of energy levels of a large number of particles in quantum mechanics. Many laws were discovered by numerical study in mathematical physics. In the late 1950s, E. P. Wigner formulated the problem in terms of the empirical distribution of a random matrix (Wigner 1955, 1958), which began the investigation into the semicircular law of Gaussian matrices. Since then, RMT has formed an active branch in modern probability theory.

Basic Concepts

Let $\mathbf{A}$ be an $n \times n$ matrix with eigenvalues $\lambda_1, \cdots, \lambda_n$. If all λ_js are real, then we can construct a 1-dimensional empirical distribution function

$$F^{\mathbf{A}}(x) = \frac{1}{n} \sum_{j=1}^{n} I(\lambda_j \le x),$$

otherwise, we may construct a 2-dimensional empirical distribution function by the real and imaginary parts of λ_j, i.e.

$$F^{\mathbf{A}}(x,y) = \frac{1}{n} \sum_{j=1}^{n} I(\Re(\lambda_j) \le x; \Im(\lambda_j) \le y).$$

Then, $F^{\mathbf{A}}$ is called the *empirical spectral distribution* (ESD) of $\mathbf{A}$. The main task of RMT is to investigate limiting properties of $F^{\mathbf{A}}$ in the case where $\mathbf{A}$ is random and the order n tends to infinity. If there is a limit distribution F, then the limit is called the *limiting spectral distribution* (LSD) of the sequence of the $\mathbf{A}$. Interesting problems include finding the explicit forms of the LSD if it exists and to investigate its properties.

There are two methods used in determining limiting properties of $F^{\mathbf{A}}$ (Bai 1999). One is the *method of moments*, using the fact that the moments of $F^{\mathbf{A}}$ are the scaled traces

of powers of **A**. The other is using *Stieltjes transforms*, defined for any distribution function F as

$$m(z) = \int \frac{1}{x - z} dF(x),$$

for $z \in \mathbb{C}$.

Contrary to the progress made on the eigenvalues of large dimensional random matrices, very few results have been obtained on the limiting properties of the eigenmatrix (i.e., the matrix of the standardized eigenvectors of **A**). Due to its importance in the application to statistics and applied areas, investigation on eigenmatrices is becoming more active.

Limiting Spectral Distributions

1. **Semicircular Law** A Wigner matrix is defined as a Hermitian (symmetric if real) matrix $\mathbf{W} = (w_{ij})_{n\times n}$ whose entries above or on the diagonal are independent. Then the ESD of $n^{-1/2}\mathbf{W}$ tends to the semicircular law with density

$$p(x) = \frac{1}{2\pi}\sqrt{4 - x^2} I(|x| < 2),$$

if $\mathrm{E}w_{ij} = 0$, $\mathrm{E}|w_{ij}|^2 = 1$ and for any $\delta > 0$,

$$\frac{1}{n^2}\sum_{ij} \mathrm{E}\left|w_{ij}^2\right| I\left(|w_{ij}| \geq \delta\sqrt{n}\right) \to 0.$$

2. **Marčenko–Pastur Law** Let $\mathbf{X} = (x_{ij})_{p\times n}$ whose entries are independent random variables with mean zero and variance 1. If $p/n \to y \in (0, \infty)$ and for any $\delta > 0$,

$$\frac{1}{np}\sum_{ij} \mathrm{E}\left|x_{ij}^2\right| I\left(|x_{ij}| \geq \delta\sqrt{n}\right) \to 0.$$

Then the ESD of $\mathbf{S}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^*$ (so-called sample covariance matrix) tends to the Marčenko–Pastur law with density

$$\frac{1}{2\pi xy}\sqrt{(b - x)(x - a)} I(a < x < b)$$

where $a = (1 - \sqrt{y})^2$ and $b = (1 + \sqrt{y})^2$. Furthermore, if $y > 1$, the LSD has a point mass $1 - 1/y$ at the origin.

3. **LSD of Products of Random Matrices** Let $\mathbf{T}$ $(p \times p)$ be a Hermitian matrix with LSD H (a probability distribution function) and $\mathbf{S}_n$, p/n satisfy the conditions in item (2). Then the ESD of $\mathbf{S}_n\mathbf{T}$ exists and the Stieltjes transform $m(z)$ is the unique solution on the upper complex plane to the equation

$$m = \int \frac{1}{t(1 - y - yzm) - z} dH(t),$$

where z is complex with positive imaginary part.

Extreme Eigenvalues and Spectrum Separation

Limits of extreme eigenvalues of large random matrices is one of the important topics. In many cases, under the assumption of finite fourth moment, the extreme eigenvalues almost surely tend to the respective boundaries of the LSD. For the product $\mathbf{S}_n\mathbf{T}$, if the support of the LSD is disconnected, then, under certain conditions, it is proved that there are no eigenvalues among the gaps and the numbers on each side are exactly the same of eigenvalues of $\mathbf{T}$, on the corresponding sides of the interval which determines the gap of the LSD (Bai and Silverstein 1999).

Further deeper investigation into extreme eigenvalues is the Tracy–Widom Law which says that $n^{2/3}$ times the difference of the extreme eigenvalues and the corresponding boundary points tends to the so-called Tracy–Widom law (Tracy and Widom 1994).

Convergence Rates of Empirical Spectral Distributions

Convergence rates of ESDs of large dimensional random matrices to their corresponding LSDs are important for application of spectral theory of large dimensional matrix. Bai inequality is the basic mathematical tool to establish the convergence rates (Bai 1993a,b). The currently known best rates are that $O(n^{-1/2})$ for the expected ESDs for Wigner matrix and for sample covariance matrix, and $O_p(n^{-2/5})$ and $O_{a.s.}(n^{-2/5+\eta})$ for their ESDs.

The exact rates are still far from known.

CLT of LSS

If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the random matrix **A** and f is a function defined on the space of the eigenvalues, then the LSS (linear spectral statistic) for the random matrix is defined by

$$\frac{1}{n}\sum_{k=1}^{n} f(\lambda_k) = \int f(x) dF^{\mathbf{A}}(x).$$

To investigate the limiting distribution of the LSS, we define $X_n(f) = n\left(\int f(x) d\left(F^{\mathbf{A}}(x) - F(x)\right)\right)$.

Under certain conditions, the normalized LSS, $X_n(f)$, is proved to tend to a normal distribution for the Wigner matrix, the product $\mathbf{S}_n\mathbf{T}$, as well as for the multivariate F-matrix, with asymptotic means and variances explicitly expressed by the Stieltjes transforms of the LSDs (Bai and Yao 2005; Bai and Silverstein 2004; Zheng 2010).

These theorems have been found to have important applications to multivariate analysis and many other areas.

Limiting Properties of Eigenvectors

Work in this area has been primarily done on the matrices in item (2) with **X** containing real entries (Silverstein

1979, 1984, 1990). Write $\mathbf{S}_n = \mathbf{O}\Lambda\mathbf{O}^*$, its spectral decomposition. When the entries of $\mathbf{X}$ are Gaussian, then $\mathbf{S}_n$ is the standard Wishart matrix, with $\mathbf{O}$ Haar-distributed in the group of $p \times p$ orthogonal matrices. The question is to compare the distribution of $\mathbf{O}$ when the entries of $\mathbf{X}$ are not Gaussian to Haar measure when p is large. This has been pursued when $\mathbf{X}$ is made up of iid random variables, by comparing the distribution of $\mathbf{y} = \mathbf{O}^*\mathbf{x}$, where $\mathbf{x}$ is a unit p-dimensional vector, to the uniform distribution on the unit sphere in $\mathbb{R}^p$. A stochastic process is defined in terms of the entries of $\mathbf{y}$, which converges weakly to Brownian bridge in the Wishart case. A necessary condition for this process to behave the same way for non Gaussian entries has been shown to be $\mathrm{E}\left(x_{11}^4\right) = 3$, matching the fourth moment of a standardized Gaussian (Silverstein 1984). For certain choices of $\mathbf{x}$ and for symmetrically distributed x_{11}, weak convergence to Brownian bridge has been shown in Silverstein (1990).

About the Author

Professor Silverstein was named IMS Fellow for "seminal contributions to the theory and application of random matrices" (2007). He has (co-)authored over 50 publications, including the book *Spectral Analysis of Large Dimensional Random Matrices* (with Z.D. Bai, 2nd edition, Springer, New York, 2009).

Cross References

►Eigenvalue, Eigenvector and Eigenspace
►Ergodic Theorem
►Limit Theorems of Probability Theory
►Multivariate Statistical Distributions
►Statistical Inference for Quantum Systems

References and Further Reading

Bai ZD (1999) Methodologies in spectral analysis of large dimensional random matrices: a review. Stat Sinica 9(3):611–677
Bai ZD (1993a) Convergence rate of expected spectral distributions of large random matrices. Part I. Wigner matrices. Ann Probab 21(2):625–648
Bai ZD (1993b) Convergence rate of expected spectral distributions of large random matrices. Part II. Sample covariance matrices. Ann Probab 21(2):649–672
Bai ZD, Silverstein JW (2004) CLT for linear spectral statistics of large-dimensional sample covariance matrices. Ann Probab 32(1):553–605
Bai ZD, Silverstein JW (1999) Exact separation of eigenvalues of large dimensional sample covariance matrices. Ann Probab 27(3):1536–1555
Bai ZD, Yao JF (2005) On the convergence of the spectral empirical process of Wigner matrices. Bernoulli 11(6):1059–1092
Silverstein JW (1979) On the randomness of eigenvectors generated from networks with random topologies. SIAM J Appl Math 37:235–245
Silverstein JW (1984) Some limit theorems on the eigenvectors of large dimensional sample covariance matrices. J Multivariate Anal 15(3):295–324
Silverstein JW (1990) Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. Ann Probab 18:1174–1194
Tracy CA, Widom H (1994) Level-spacing distributions and the Airy kernel. Commun Math Phys 159:151–174
Wigner EP (1955) Characteristic vectors bordered matrices with infinite dimensions. Ann Math 62:548–564
Wigner EP (1958) On the distributions of the roots of certain symmetric matrices. Ann Math 67:325–327
Zheng S (2010) Central limit theorem for linear spectral statistics of large dimensional F-Matrix, to appear in Ann Inst Henri Poincaré Probab Stat

Random Permutations and Partition Models

Peter McCullagh
John D. MacArthur Distinguished Service Professor
University of Chicago, Chicago, IL, USA

Set Partitions

For $n \geq 1$, a partition B of the finite set $[n] = \{1, \ldots, n\}$ is

- A collection $B = \{b_1, \ldots\}$ of disjoint non-empty subsets, called blocks, whose union is $[n]$
- An equivalence relation or Boolean function $B\colon [n] \times [n] \to \{0,1\}$ that is reflexive, symmetric and transitive
- A symmetric Boolean matrix such that $B_{ij} = 1$ if i, j belong to the same block

These equivalent representations are not distinguished in the notation, so B is a set of subsets, a matrix, a Boolean function, or a subset of $[n]\times[n]$, as the context demands. In practice, a partition is sometimes written in an abbreviated form, such as $B = 2|13$ for a partition of $[3]$. In this notation, the five partitions of $[3]$ are

$$123, \quad 12|3, \quad 13|2, \quad 23|1, \quad 1|2|3.$$

The blocks are unordered, so $2|13$ is the same partition as $13|2$ and $2|31$.

A partition B is a sub-partition of B^* if each block of B is a subset of some block of B^* or, equivalently, if $B_{ij} = 1$ implies $B^*_{ij} = 1$. This relationship is a partial order denoted by $B \leq B^*$, which can be interpreted as $B \subset B^*$ if each partition is regarded as a subset of $[n]^2$. The partition lattice $\mathcal{E}_n$ is the set of partitions of $[n]$ with this partial order. To each pair of partitions B, B' there corresponds a greatest lower bound $B \wedge B'$, which is the set intersection or Hadamard component-wise matrix product. The least upper bound

$B \vee B'$ is the least element that is greater than both, the transitive completion of $B \cup B'$. The least element of $\mathcal{E}_n$ is the partition $\mathbf{0}_n$ with n singleton blocks, and the greatest element is the single-block partition denoted by $\mathbf{1}_n$.

A permutation $\sigma\colon [n] \to [n]$ induces an action $B \mapsto B^\sigma$ by composition such that $B^\sigma(i,j) = B(\sigma(i), \sigma(j))$. In matrix notation, $B^\sigma = \sigma B \sigma^{-1}$, so the action by conjugation permutes both the rows and columns of B in the same way. The block sizes are preserved and are maximally invariant under conjugation. In this way, the 15 partitions of $[4]$ may be grouped into five orbits or equivalence classes as follows:

$$1234, \quad 123|4\,[4], \quad 12|34\,[3], \quad 12|3|4\,[6], \quad 1|2|3|4.$$

Thus, for example, 12|34 is the representative element for one orbit, which also includes 13|24 and 14|23.

The symbol $\#B$ applied to a set denotes the number of elements, so $\#B$ is the number of blocks, and $\#b$ is the size of block $b \in B$. If $\mathcal{E}_n$ is the set of equivalence relations on $[n]$, or the set of partitions of $[n]$, the first few values of $\#\mathcal{E}_n$ are 1, 2, 5, 15, 52, called Bell numbers. More generally, $\#\mathcal{E}_n$ is the nth moment of the unit Poisson distribution whose exponential generating function is $\exp(e^t - 1)$. In the discussion of explicit probability models on $\mathcal{E}_n$, it is helpful to use the ascending and descending factorial symbols

$$\alpha^{\uparrow r} = \alpha(\alpha+1)\cdots(\alpha+r-1) = \Gamma(r+\alpha)/\Gamma(\alpha)$$
$$k^{\downarrow r} = k(k-1)\cdots(k-r+1)$$

for integer $r \geq 0$. Note that $k^{\downarrow r} = 0$ for positive integers $r > k$. By convention $\alpha^{\uparrow 0} = 1$.

Partition Model

The term *partition model* refers to a probability distribution, or family of probability distributions, on the set $\mathcal{E}_n$ of partitions of $[n]$. In some cases, the probability is concentrated on the the subset $\mathcal{E}_n^k \subset \mathcal{E}_n$ of partitions having k or fewer blocks. A distribution on $\mathcal{E}_n$ such that $p_n(B) = p_n(\sigma B \sigma^{-1})$ for every permutation $\sigma\colon [n] \to [n]$ is said to be finitely exchangeable. Equivalently, p_n is exchangeable if $p_n(B)$ depends only on the block sizes of B.

Historically, the most important examples are Dirichlet-multinomial random partitions generated for fixed k in three steps as follows.

- First generate the random probability vector $\pi = (\pi_1, \ldots, \pi_k)$ from the Dirichlet distribution with parameter $(\theta_1, \ldots, \theta_k)$.
- Given π, the sequence $Y_1, \ldots, Y_n, \ldots$ is independent and identically distributed, each component taking values in $\{1, \ldots, k\}$ with probability π. Each sequence of length n in which the value r occurs $n_r \geq 0$ times has probability
$$\frac{\Gamma(\theta_\cdot)\prod_{j=1}^k \theta_j^{\uparrow n_j}}{\Gamma(n+\theta_\cdot)},$$
where $\theta_\cdot = \sum \theta_j$.
- Now forget the labels $1, \ldots, k$ and consider only the partition B generated by the sequence Y, i.e., $B_{ij} = 1$ if $Y_i = Y_j$. The distribution is exchangeable, but an explicit simple formula is available only for the uniform case $\theta_j = \lambda/k$, which is now assumed. The number of sequences generating the same partition B is $k^{\downarrow \#B}$, and these have equal probability in the uniform case. Consequently, the induced partition has probability
$$p_{nk}(B,\lambda) = k^{\downarrow \#B}\frac{\Gamma(\lambda)\prod_{b\in B}(\lambda/k)^{\uparrow \#b}}{\Gamma(n+\lambda)}, \qquad (1)$$
called the uniform Dirichlet-multinomial partition distribution. The factor $k^{\downarrow \#B}$ ensures that partitions having more than k blocks have zero probability.

In the limit as $k \to \infty$, the uniform Dirichlet-multinomial partition becomes

$$p_n(B,\lambda) = \frac{\lambda^{\#B}\prod_{b\in B}\Gamma(\#b)}{\lambda^{\uparrow n}}. \qquad (2)$$

This is the celebrated Ewens distribution, or Ewens sampling formula, which arises in population genetics as the partition generated by allele type in a population evolving according to the Fisher–Wright model by random mutation with no selective advantage of allele types (Ewens 1972). The preceding derivation, a version of which can be found in Chap. 3 of Kingman (1980), goes back to Watterson (1974). The Ewens partition is the same as the partition generated by a sequence drawn according to the Blackwell-McQueen urn scheme (Blackwell and MacQueen 1973).

Although the derivation makes sense only if k is a positve integer, the distribution (1) is well defined for negative values $-\lambda < k < 0$. For a discussion of this and the connection with GEM distributions and Poisson-Dirichlet distributions, see Pitman (2006, Sect. 3.2).

Partition Processes and Partition Structures

Deletion of element n from the set $[n]$, or deletion of the last row and column from $B \in \mathcal{E}_n$, determines a map $D_n\colon \mathcal{E}_n \to \mathcal{E}_{n-1}$, a projection from the larger to the smaller lattice. These deletion maps make the sets $\{\mathcal{E}_1, \mathcal{E}_2, \ldots\}$ into a projective system

$$\cdots \mathcal{E}_{n+1} \xrightarrow{D_{n+1}} \mathcal{E}_n \xrightarrow{D_n} \mathcal{E}_{n-1} \cdots$$

A family $p = (p_1, p_2, \ldots)$ in which p_n is a probability distribution on $\mathcal{E}_n$ is said to be mutually consistent, or

Kolmogorov-consistent, if each p_{n-1} is the marginal distribution obtained from p_n under deletion of element n from the set $[n]$. In other words, $p_{n-1}(A) = p_n\left(D_n^{-1}A\right)$ for $A \subset \mathcal{E}_{n-1}$. Kolmogorov consistency guarantees the existence of a random partition of the natural numbers whose finite restrictions are distributed as p_n. The partition is infinitely exchangeable if each p_n is finitely exchangeable. Some authors, for example Kingman (1980), refer to p as *a partition structure*.

An exchangeable partition process may be generated from an exchangeable sequence $Y_1, Y_2, \ldots$ by the transformation $B_{ij} = 1$ if $Y_i = Y_j$ and zero otherwise. The Dirichlet-multinomial and the Ewens processes are generated in this way. Kingman's (1978) paintbox construction shows that every exchangeable partition process may be generated from an exchangeable sequence in this manner.

Let B be an infinitely exchangeable partition, $B[n] \sim p_n$, let B^* be a fixed partition in $\mathcal{E}_n$, and suppose that the event $B[n] \leq B^*$ occurs. Then $B[n]$ lies in the lattice interval $[\mathbf{0}_n, B^*]$, which means that $B[n] = B[b_1]|B[b_2]|\ldots$ is the concatenation of partitions of the blocks $b \in B^*$. For each block $b \in B^*$, the restriction $B[b]$ is distributed as $p_{\#b}$, so it is natural to ask whether, and under what conditions, the blocks of B^* are partitioned independently given $B[n] \leq B^*$. Conditional independence implies that

$$p_n(B \,|\, B[n] \leq B^*) = \prod_{b \in B^*} p_{\#b}(B[b]), \tag{3}$$

which is a type of non-interference or lack-of-memory property not dissimilar to that of the exponential distribution on the real line. It is straightforward to check that the condition is satisfied by (2) but not by (1). Aldous (1996) shows that conditional independence uniquely characterizes the Ewens family.

Chinese Restaurant Process

A partition process is a random partition $B \sim p$ of a countably infinite set $\{u_1, u_2, \ldots\}$, and the restriction $B[n]$ of B to $\{u_1, \ldots, u_n\}$ is distributed as p_n. The conditional distribution of $B[n+1]$ given $B[n]$ is determined by the probabilities assigned to those events in $\mathcal{E}_{n+1}$ that are consistent with $B[n]$, i.e. the events $u_{n+1} \mapsto b$ for $b \in B$ and $b = \emptyset$. For the uniform Dirichlet-multinomial model (1), these are

$$\mathrm{pr}(u_{n+1} \mapsto b \,|\, B[n] = B) = \begin{cases} (\#b + \lambda/k)/(n+\lambda) & b \in B \\ \lambda(1 - \#B/k)/(n+\lambda) & b = \emptyset. \end{cases} \tag{4}$$

In the limit as $k \to \infty$, we obtain

$$\mathrm{pr}(u_{n+1} \mapsto b \,|\, B[n] = B) = \begin{cases} \#b/(n+\lambda) & b \in B \\ \lambda/(n+\lambda) & b = \emptyset, \end{cases} \tag{5}$$

which is the conditional probability for the Ewens process.

To each partition process p there corresponds a sequential description called the Chinese restaurant process, in which $B[n]$ is the arrangement of the first n customers at $\#B$ tables. The placement of the next customer is determined by the conditional distribution $p_{n+1}(B[n+1] \,|\, B[n])$. For the Ewens process, the customer chooses a new table with probability $\lambda/(n+\lambda)$ or one of the occupied tables with probability proportional to the number of occupants. The term, due to Pitman, Dubins and Aldous, is used primarily in connection with the Ewens and Dirichlet-multinomial models.

Exchangeable Random Permutations

Beginning with the uniform distribution on permutations of $[n]$, the exponential family with canonical parameter $\theta = \log(\lambda)$ and canonical statistic $\#\sigma$ equal to the number of cycles is

$$p_n(\sigma) = \lambda^{\#\sigma}/\lambda^{\uparrow n}.$$

The Stirling number of the first kind, $S_{n,k}$, is the number of permutations of $[n]$ having exactly k cycles, for which $\lambda^{\uparrow n} = \sum_{k=1}^{n} S_{n,k}\lambda^k$ is the generating function. The cycles of the permutation determine a partition of $[n]$ whose distribution is (2), and a partition of the integer n whose distribution is (6). From the cumulant function

$$\log(\lambda^{\uparrow n}) = \sum_{j=0}^{n-1} \log(j+\lambda)$$

it follows that $\#\sigma = X_0 + \cdots + X_{n-1}$ is the sum of independent Bernoulli variables with parameter $E(X_j) = \lambda/(\lambda+j)$, which is evident also from the Chinese restaurant representation. For large n, the number of cycles is roughly Poisson with parameter $\lambda \log(n)$, implying that $\hat{\lambda} \simeq \#\sigma/\log(n)$ is a consistent estimate as $n \to \infty$, but practically inconsistent.

A minor modification of the Chinese restaurant process also generates a random permutation by keeping track of the cyclic arrangement of customers at tables. After n customers are seated, the next customer chooses a table with probability (4) or (5), as determined by the partition process. If the table is occupied, the new arrival sits to the left of one customer selected uniformly at random from the table occupants. The random permutation thus generated is $j \mapsto \sigma(j)$ from j to the left neighbour $\sigma(j)$.

Provided that the partition process is consistent and exchangeable, the distributions p_n on permutations of $[n]$ are exchangeable and mutually consistent under the projection $\Pi_n \to \Pi_{n-1}$ on permutations in which element n is deleted from the cyclic representation (Pitman 2006, Sect. 3.1). In this way, every infinitely exchangeable random partition also determines an infinitely exchangeable random permutation $\sigma: \mathbb{N} \to \mathbb{N}$ of the natural numbers. Distributional exchangeability in this context is not to be confused with uniformity on Π_n.

On the Number of Unseen Species

A partition of the set $[n]$ is a set of blocks, and the block sizes determine a partition of the integer n. For example, the partition 15|23|4 of the set $[5]$ is associated with the integer partition 2 + 2 + 1, one singleton and two doubletons. An integer partition $m = (m_1, \ldots, m_n)$ is a list of multiplicities, also written as $m = 1^{m_1} 2^{m_2} \cdots n^{m_n}$, such that $\sum j m_j = n$. The number of blocks, usually called the number of parts of the integer partition, is the sum of the multiplicities $m_\cdot = \sum m_j$.

Each integer partition is a group orbit in $\mathcal{E}_n$ induced by the action of the symmetric group on the set $[n]$. The multiplicity vector m contains all the information about block sizes, but there is a subtle transfer of emphasis from block sizes to the multiplicities of the parts.

By definition, an exchangeable distribution on set partitions is a function only of the block sizes, so $p_n(B) = q_n(m)$, where m is the integer partition corresponding to B. Since there are

$$\frac{n!}{\prod_{j=1}^{n} (j!)^{m_j} m_j!}$$

set partitions B corresponding to a given integer partition m, to each exchangeable distribution p_n on set partitions there corresponds a marginal distribution

$$q_n(m) \frac{n!}{\prod_{j=1}^{n} (j!)^{m_j} m_j!}$$

on integer partitions. For example, the Ewens distribution on integer partitions is

$$\frac{\lambda^{m_\cdot} \Gamma(\lambda) \prod \Gamma(j)^{m_j}}{\Gamma(n+\lambda)} \times \frac{n!}{\prod_{j=1}^{n} (j!)^{m_j} m_j!} = \frac{\lambda^{m_\cdot} n! \, \Gamma(\lambda)}{\Gamma(n+\lambda) \prod_j j^{m_j} m_j!}. \tag{6}$$

This version leads naturally to an alternative description as follows. Let $M = M_1, \ldots, M_n$ be independent Poisson random variables with mean $E(M_j) = \lambda \theta^j / j$ for some positive number θ. Then $\sum j M_j$ is sufficient for θ, and the conditional distribution $\mathrm{pr}\left(M = m \,\middle|\, \sum_{j=1}^{n} j M_j = n\right)$ is the Ewens integer-partition distribution with parameter λ. This representation leads naturally to a simple method of estimation and testing, using Poisson log-linear models with model formula $1 + j$ and offset $-\log(j)$ for response vectors that are integer partitions.

The problem of estimating the number of unseen species was first tackled in a paper by Fisher (1943), using an approach that appears to be entirely unrelated to partition processes. Specimens from species i occur as a Poisson process (see ▶Poisson Processes) with rate ρ_i, the rates for distinct species being independent and identically distributed gamma random variables. The number $N_i \geq 0$ of occurrences of species i in an interval of length t is a negative binomial random variable

$$\mathrm{pr}(N_i = x) = (1-\theta)^{\nu} \theta^x \frac{\Gamma(\nu + x)}{x! \, \Gamma(\nu)}. \tag{7}$$

In this setting, $\theta = t/(1+t)$ is a monotone function of the sampling time, whereas $\nu > 0$ is a fixed number independent of t. Specimen counts for distinct species are independent and identically distributed random variables with parameters $\nu > 0$ and $0 < \theta < 1$.

The probability that no specimens from species i occur in the sample is $(1-\theta)^{\nu}$, the same for every species. Most species are unlikely to be observed if either θ is small, i.e., the time interval is short, or ν is small.

Let M_x be the number of species occurring $x \geq 0$ times, so that $M_\cdot$ is the unknown total number of species of which $M_\cdot - M_0$ are observed. The approach followed by Fisher is to estimate the parameters θ, ν by conditioning on the number of species observed and regarding the observed multiplicities M_x for $x \geq 1$ as multinomial with parameter vector proportional to the negative binomial frequencies (7). For Fisher's entomological examples, this approach pointed to $\nu = 0$, consistent with the Ewens distribution (6), and indicating that the data are consistent with the number of species being infinite. Fisher's approach using a model indexed by species is less direct for ecological purposes than a process indexed by specimens. Nonetheless, subsequent analyses by Good and Toulmin (1956), Holgate (1969) and Efron and Thisted (1976) showed how Fisher's model can be used to make predictions about the likely number of new species in a subsequent temporal extension of the original sample. This amounts to a version of the Chinese restaurant process.

At this point, it is worth clarifying the connection between Fisher's negative binomial formulation and the Ewens partition formulation. The relation between them is the same as the relation between binomial and negative binomial sampling schemes for a Bernoulli process: they are not equivalent, but they are complementary. The partition formulation is an exchangeable process indexed by *specimens*: it gives the distribution of species numbers in a

sample consisting of a fixed number of *specimens*. Fisher's version is also an exchangeable process, in fact an iid process, but this process is indexed by *species*: it gives the distribution of the sample composition for a fixed set of *species* observed over a finite period. In either case, the conditional distribution given a sample containing k species and n specimens is the distribution induced from the uniform distribution on the set of $S_{n,k}$ permutations having k cycles. For the sorts of ecological or literary applications considered by Good and Toulmin (1956) or Efron and Thisted (1976), the partition process indexed by specimens is much more direct than one indexed by species.

Fisher's finding that the multiplicities decay as $E(M_j) \propto \theta^j/j$, proportional to the frequencies in the log-series distribution, is a property of many processes describing population structure, either social structure or genetic structure. It occurs in Kendall's (1975) model for family sizes as measured by surname frequencies. One explanation for universality lies in the nature of the transition rates for Kendall's process, a discussion of which can be found in Sect. 2.4 of Kelly (1978).

Equivariant Partition Models

A family $p_n(\sigma;\theta)$ of distributions on permutations indexed by a parameter matrix θ, is said to be equivariant under the induced action of the symmetric group if $p_n(\sigma;\theta) = p_n(g\sigma g^{-1}; g\theta g^{-1})$ for all σ, θ, and for each group element $g\colon [n] \to [n]$. By definition, the parameter space is closed under conjugation: $\theta \in \Theta$ implies $g\theta g^{-1} \in \Theta$. The same definition applies to partition models. Unlike exchangeability, equivariance is not a property of a distribution, but a property of the family. In this setting, the family associated with $[n]$ is not necessarily the same as the family of marginal distributions induced by deletion from $[n+1]$.

Exponential family models play a major role in both theoretical and applied work, so it is natural to begin with such a family of distributions on permutations of the matrix-exponential type

$$p_n(\sigma;\theta) = \alpha^{\#\sigma} \exp(\mathrm{tr}(\sigma\theta))/M_\alpha(\theta),$$

where $\alpha > 0$ and $\mathrm{tr}(\sigma\theta) = \sum_{j=1}^n \theta_{\sigma(j),j}$ is the trace of the ordinary matrix product. The normalizing constant is the α-permanent

$$M_\alpha(\theta) = \mathrm{per}_\alpha(K) = \sum_\sigma \alpha^{\#\sigma} \prod_{j=1}^n K_{\sigma(j),j}$$

where $K_{ij} = \exp(\theta_{ij})$ is the component-wise exponential matrix. This family of distributions on permutations is equivariant.

The limit of the α-permanent as $\alpha \to 0$ gives the sum of cyclic permutations

$$\mathrm{cyp}(K) = \lim_{\alpha\to 0} \alpha^{-1} \mathrm{per}_\alpha(K) = \sum_{\sigma:\#\sigma=1} \prod_{j=1}^n K_{\sigma(j),j},$$

giving an alternative expression for the α-permanent

$$\mathrm{per}_\alpha(K) = \sum_{B\in\mathcal{E}_n} \alpha^{\#B} \prod_{b\in B} \mathrm{cyp}(K[b])$$

as a sum over partitions. The induced marginal distribution (10) on partitions is of the product-partition type recommended by Hartigan (1990), and is also equivariant. Note that the matrix θ and its transpose determine the same distribution on partitions, but they do not usually determine the same distribution on permutations.

The α-permanent has a less obvious convolution property that helps to explain why this function might be expected to occur in partition models:

$$\sum_{b\subset[n]} \mathrm{per}_\alpha(K[b])\,\mathrm{per}_{\alpha'}(K[\bar{b}]) = \mathrm{per}_{\alpha+\alpha'}(K). \qquad (8)$$

The sum extends over all 2^n subsets of $[n]$, and $\bar{b}$ is the complement of b in $[n]$. A derivation can be found in section 2.4 of McCullagh and Møller (2006). If B is a partition of $[n]$, the symbol $K \cdot B = B \cdot K$ denotes the Hadamard component-wise matrix product for which

$$\mathrm{per}_\alpha(K \cdot B) = \prod_{b\in B} \mathrm{per}_\alpha(K[b])$$

is the product over the blocks of B of α-permanents restricted to the blocks. Thus the function $B \mapsto \mathrm{per}_\alpha(K\cdot B)$ is of the product-partition type.

With α, K as parameters, we may define a family of probability distributions on $\mathcal{E}_n^k$, i.e., partitions of $[n]$ having k or fewer blocks, as follows:

$$p_{nk}(B) = k^{\downarrow\#B}\,\mathrm{per}_{\alpha/k}(K\cdot B)/\mathrm{per}_\alpha(K). \qquad (9)$$

The fact that (9) is a probability distribution on $\mathcal{E}_n$ follows from the convolution property of permanents. The limit as $k \to \infty$

$$p_n(B) = \alpha^{\#B} \prod_{b\in B} \mathrm{cyp}(K[b])/\mathrm{per}_\alpha(K), \qquad (10)$$

is a product-partition model satisfying the conditional independence property (3). For $K = \mathbf{1}_n$, the $n \times n$ matrix whose elements are all one, $\mathrm{per}_\alpha(\mathbf{1}_n) = \alpha^{\uparrow n}$ is the ascending factorial function. Thus the uniform Dirichlet-multinomial model (1) and the Ewens model (2) are both obtained by setting $\theta = 0$.

Leaf-Labelled Trees

Kingman's $[n]$-coalescent is a non-decreasing, $\mathcal{E}_n$-valued Markov process (see ▶Markov Processes) (B_t) in continuous-time starting from the partition $B_0 = \mathbf{0}_n$ with n singleton blocks at time zero. The coalescence intensity is one for each pair of blocks regardless of size, so each coalescence event unites two blocks chosen uniformly at random from the set of pairs. Consequently, the first coalescence occurs after a random time T_n exponentially distributed with rate $\rho(n) = n(n-1)/2$ and mean $1/\rho(n)$. After k coalescences, the partition consists of $n-k$ blocks, and the waiting time T_k for the next subsequent coalescence is exponential with rate $\rho(n-k)$. The time to complete coalescence is the sum of independent exponentials $T = T_n + T_{n-1} + \cdots + T_2$, which is a random variable with mean $2-2/n$ and variance increasing from 1 at $n = 2$ to a little less than 1.16 as $n \to \infty$. In the context of the Fisher–Wright model, the coalescent describes the genealogical relationships among a sample of individuals, and T is the time to the most recent common ancestor of the sample.

The $[n]$-coalescent is exchangeable for each n, but the property that makes it interesting mathematically, statistically and genetically is its consistency under selection or sub-sampling (Kingman 1982). If we denote by p_n the distribution on $[n]$-trees implied by the specific Markovian model described above, it can be shown that the embedded tree obtained by deleting element n from the sample $[n]$ is not only Markovian but also distributed as p_{n-1}, i.e., the same coalescent rule applied to the subset $[n-1]$. This property is mathematically essential for genealogical trees because the occurrence or non-occurrence of individual n in the sample does not affect the genealogical relationships among the remainder.

A fragmentation $[n]$-tree is a non-increasing $\mathcal{E}_n$-valued Markov process starting from the trivial partition $B_0 = \mathbf{1}_n$ with one block of size n at time $t = 0$. The simplest of these are the consistent binary Gibbs fragmentation trees studied by Aldous (1996), Bertoin (2001, 2006) and McCullagh et al. (2008). The first split into two branches occurs at a random time T_n exponentially distributed with parameter $\rho(n)$. Subsequently, each branch fragments independently according to the same family of distributions with parameter $\rho(\#b)$ for branch b, which is a Markovian conditional independence property analogous to (3). Consistency and conditional independence put severe limitations on both the splitting distribution and the rate function $\rho(n)$, so the entire class is essentially one-dimensional.

A rooted leaf-labelled tree T is also a non-negative symmetric matrix. The interpretation of T_{ij} as the distance from the root to the junction at which leaves i, j occur on disjoint branches implies the inequality $T_{ij} \geq \min(T_{ik}, T_{jk})$ for all $i, j, k \in [n]$. The set of $[n]$-trees is a subset of the positive definite symmetric matrices, not a manifold, but a finite union of manifolds of dimension $2n-1$, or n if the diagonal elements are constrained to be equal. Like partitions, rooted trees form a projective system within the positive definite matrices. A fragmentation tree is an infinitely exchangeable random tree, which is also a special type of infinitely exchangeable random matrix.

Cluster Processes and Classification Models

A $\mathcal{R}^d$-valued cluster process is a pair (Y, B) in which $Y = (Y_1, \ldots)$ is an $\mathcal{R}^d$-valued random sequence and B is a random partition of $\mathbb{N}$. The process is said to be exchangeable if, for each finite sample $[n] \subset \mathbb{N}$, the restricted process $(Y[n], B[n])$ is invariant under permutation $\sigma: [n] \to [n]$ of sample elements.

The Gauss–Ewens process is the simplest non-trivial example for which the distribution for a sample $[n]$ is as follows. First fix the parameter values $\lambda > 0$, and Σ^0, Σ^1 both positive definite of order d. In the first step B has the Ewens distribution on $\mathcal{E}_n$ with parameter λ. Conditionally on B, Y is a zero-mean Gaussian matrix of order $n \times d$ with covariance matrix

$$\operatorname{cov}(Y_{ir}, Y_{js} \mid B) = \delta_{ij}\Sigma^0_{rs} + B_{ij}\Sigma^1_{rs},$$

where δ_{ij} is the Kronecker symbol. A scatterplot color-coded by blocks of the Y values in $\mathcal{R}^2$ shows that the points tend to be clustered, the degree of clustering being governed by the ratio of between to within-cluster variances.

For an equivalent construction we may proceed using a version of the Chinese restaurant process in which tables are numbered in order of occupancy, and $t(i)$ is number of the table at which customer i is seated. In addition, $\epsilon_1, \ldots$ and $\eta_1, \ldots$ are independent Gaussian sequences with independent components $\epsilon_i \sim N_d(0, \Sigma^0)$, and $\eta_i \sim N_d(0, \Sigma^1)$. The sequence t determines B, and the value for individual i is a vector $Y_i = \eta_{t(i)} + \epsilon_i$ in $\mathcal{R}^d$, or $Y_i = \mu + \eta_{t(i)} + \epsilon_i$ if a constant non-zero mean vector is included.

Despite the lack of class labels, cluster processes lend themselves naturally to prediction and classification, also called supervised learning. The description that follows is taken from McCullagh and Yang (2006) but, with minor modifications, the same description applies equally to more complicated non-linear versions associated with generalized linear mixed models (Blei et al. 2003). Given the observation $(Y[n], B[n])$ for the 'training sample' $[n]$, together with the feature vector Y_{n+1} for specimen u_{n+1}, the conditional distribution of $B[n+1]$ is determined by

those events $u_{n+1} \mapsto b$ for $b \in B$ and $b = \emptyset$ that are compatible with the observation. The assignment of a positive probability to the event that the new specimen belongs to a previously unobserved class seems highly desirable, even logically necessary, in many applications.

If the classes are tree-structured with two levels, we may generate a sub-partition $B' \leq B$ whose conditional distribution given B is Ewens restricted to the interval $[\mathbf{0}_n, B]$, with parameter λ'. This sub-partition has the effect of splitting each main clusters randomly into sub-clusters. For the sample $[n]$, let $t'(i)$ be the number of the sub-cluster in which individual i occurs. Given B, B', the Gauss-Ewens two-level tree process is a sum of three independent Gaussian processes $Y_i = \eta_{t(i)} + \eta'_{t'(i)} + \epsilon_i$ for which the conditional distributions may be computed as before. In this situation, however, events that are compatible with the observation $B[n], B'[n]$ are of three types as follows:

$$u_{n+1} \mapsto b' \in B'[n], \quad u_{n+1} \mapsto \emptyset \subset b \in B[n], \quad u_{n+1} \mapsto \emptyset.$$

In all, there are $\#B' + \#B + 1$ disjoint events for which the conditional distribution given $B[n], B'[n], Y[n+1]$ must be computed. An event of the second type is one in which the new specimen belongs to the major class $b \in B$, but not to any of the sub-types previously observed for this class.

Further Applications of Partition Models

Exchangeable partition models are used to construct nontrivial, exchangeable processes suitable for cluster analysis and density estimation. See Frayley and Raftery (2002) for a review. Here, cluster analysis means cluster detection and cluster counting in the absence of covariate or relational information about the units. In the computer-science literature, cluster detection is also called unsupervised learning. The simplest of these models is the marginal Gauss–Ewens process in which only the sequence Y is observed. The conditional distribution $p_n(B \mid Y)$ on $\mathcal{E}_n$ is the posterior distribution on clusterings or partitions of $[n]$, and $E(B \mid Y)$ is the one-dimensional marginal distribution on pairs of units. In estimating the number of clusters, it is important to distinguish between the sample number $\#B[n]$, which is necessarily finite, and the population number $\#B[\mathbb{N}]$, which could be infinite (McCullagh and Yang 2008).

Exchangeable partition models are also used to provide a Bayesian solution to the multiple comparisons problem. The key idea is to associate with each partition B of $[k]$ a subspace $V_B \subset \mathcal{R}^k$ equal to the span of the columns of B. Thus, V_B consists of vectors x such that $x_r = x_s$ if $B_{rs} = 1$. For a treatment factor having k levels $\tau_1, \ldots, \tau_k$, the Gauss–Ewens prior distribution on R^k puts positive mass on the subspaces V_B for each $B \in \mathcal{E}_k$. Likewise, the posterior distribution also puts positive probability on these subspaces, which enables us to compute in a coherent way the posterior probability $\mathrm{pr}(\tau \in V_B)$ or the marginal posterior probability $\mathrm{pr}(\tau_r = \tau_s \mid y)$. For details, see (Gopalan and Berry 1998).

Acknowledgments

Support for this research was provided in part bt NSF Grant DMS-0906592.

About the Author

Peter McCullagh is the John D. MacArthur Distinguished Service Professor in the Department of Statistics at the University of Chicago. Professor McCullagh is a past editor of *Bernoulli*, a fellow of the Royal Society of London and of the American Academy of Arts and Sciences. He is co-author with John Nelder of the book *Generalized linear Models* (Chapman and Hall, 1989).

Cross References

►Cluster Analysis: An Introduction
►Data Mining
►Multivariate Statistical Distributions
►Permanents in Probability Theory

References and Further Reading

Aldous D (1996) Probability distributions on cladograms. In: Random discrete structures. IMA Vol Appl Math 76. Springer, New York, pp 1–18

Bertoin J (2001) Homogeneous fragmentation processes. Probab Theor Relat Fields 121:301–318

Bertoin J (2006) Random fragmentation and coagulation processes. Cambridge studies in advanced mathematics, vol 102. Cambridge University Press, Cambridge

Blackwell D, MacQueen J (1973) Ferguson distributions via Pólya urn schemes. Ann Stat 1:353–355

Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. J Mach learn Res 3:993–1022

Efron B, Thisted RA (1976) Estimating the number of unknown species: how many words did Shakespeare know? Biometrika 63:435–447

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12:42–58

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis and density estimation. J Am Stat Assoc 97:611–631

Good IJ, Toulmin GH (1956) The number of new species, and the increase in population coverage when a sample is increased. Biometrika 43:45–63

Gopalan R, Berry DA (1998) Bayesian multiple comparisons using Dirichlet process priors. J Am Stat Assoc 93:1130–1139

Hartigan JA (1990) Partition models. Commun Stat Theor Meth 19:2745–2756

Holgate P (1969) Species frequency distributions. Biometrika 65:651–660
Kelly FP (1978) Reversibility and stochastic networks. Wiley, Chichester
Kendall DG (1975) Some problems in mathematical genealogy. In: Perspectives in probability and statistics: papers in honour of M.S. Bartlett. Academic, London, pp 325–345
Kingman JFC (1975) Random discrete distributions (with discussion). J R Stat Soc B 37:1–22
Kingman JFC (1977) The population structure associated with the Ewens sampling formula. Theor Popul Biol 11:274–283
Kingman JFC (1978) The representation of partition structures. J Lond Math Soc 18:374–380
Kingman JFC (1980) Mathematics of genetic diversity. CBMS-NSF conference series in applied mathematics, 34 SIAM, Philadelphia
Kingman JFC (1982) The coalescent. Stoch Proc Appl 13:235–248
McCullagh P, Møller J (2006) The permanental process. Adv Appl Prob 38:873–888
McCullagh P, Yang J (2006) Stochastic classification models. In: Proceedings of the international congress of mathematicians, 2006, vol 3, pp 669–686
McCullagh P, Yang J (2008) How many clusters? Bayesian Anal 3:1–19
McCullagh P, Pitman J, Winkel M (2008) Gibbs fragmentation trees. Bernoulli 14:988–1002
Pitman J (2006) Combinatorial stochastic processes. Springer, Berlin
Watterson GA (1974) The sampling theory of selectively neutral alleles. Adv Appl Probab 6:217–250

Random Variable

Czesław Stępniak
Professor
Maria Curie-Skłodowska University, Lublin, Poland
University of Rzeszów, Rzeszów, Poland

Random variable (r.v.) is a real function defined on the set of outcomes. It reduces the set-theoretical problems in probability to some simpler ones in real analysis. R.v. 's are indispensable in probability computing.

Motivation

Formal definition of a r.v. is a consequence of some practical and logical needs. Let us start from a measure-theoretic frame $(\Omega, \mathcal{A}, P)$, where Ω is the set of outcomes, $\mathcal{A}$ is a σ-algebra of subsets of Ω, serving as random events, and P is a normalized measure on the space $(\Omega, \mathcal{A})$, called probability. Any real function $f = f(\omega)$ transforms the initial probability system $(\Omega, \mathcal{A}, P)$ onto some induced system $(\Omega_f, \mathcal{A}_f, P_f)$, where Ω_f is the image of Ω by f, $\mathcal{A}_f$ is the σ-algebra of of subsets B on Ω_f induced by f, while P_f is a probability measure on $(\Omega_f, \mathcal{A}_f)$ defined by

$$P_f(B) = P(\{\omega : f(\omega) \in B\}). \tag{1}$$

For practical reasons we require that the family $\mathcal{A}_f$ includes all intervals $(a; b >$, i.e., that $\mathcal{A}_f \supseteq \mathcal{B}$, where $\mathcal{B}$ means the family of Borel sets in the real line. On the other hand the right side of (1) is well defined, if and only if, $\{\omega : f(\omega) \in B\} \in \mathcal{A}$. Since $\mathcal{A}_f$ is σ-algebra generated by the intervals, the last condition may be expressed in a more readable form

$$\{\omega : f(\omega) \le c\} \in \mathcal{A} \text{ for any } c \in R. \tag{2}$$

Formal Definition

Any real function defined on the (measurable) space $(\Omega, \mathcal{A})$ satisfying the condition (2) is said to be a random variable on $(\Omega, \mathcal{A})$. Traditionally, random variables are denoted by capital letters $X(\omega), Y(\omega), Z(\omega)$, or simply X, Y, Z. The following exmple shows that not every function of outcome is a random variable.

Example 1 Let us set $\Omega = \{1,2,3,4,5\}$, $\mathcal{A} = \{\varnothing, \{1,3,5\}, \{2,4\}, \Omega\}$ and

$$f(\omega) = \begin{cases} 0, & \text{if } \omega \le 2 \\ 1, & \text{if } \omega > 2. \end{cases}$$

By setting in (2) $c = 1$ we get $\{\omega : f(\omega) < 1\} = \{1,2\} \notin \mathcal{A}$. Thus f is not random variable on the space $(\Omega, \mathcal{A})$.

The probability measure $P_X(B) = P(\{\omega : X(\omega) \in B\})$ for $B \in \mathcal{B}$ is said to be distribution of the r.v. X. This expression has mainly theoretical sense, because the Borel sets are abstractive objects. More practical in use is so called *cumulative distribution function (c.d.f)* F defined by $F(\alpha) = P(\{\omega : X(\omega) \le \alpha\})$.

Example 2 (Two-Dice Game). Here the set of outcomes may be presented in the form $\Omega = \{(i,j) : i,j = 1,2,3,4,5,6\}$, the family of random events $\mathcal{A}$ may be defined as the family of all subsets of Ω, while $X(\omega)$, for any $\omega = (i,j)$ may be defined as $i + j$. Such a r.v. takes the possible values from 2 to 12 with probabilities

$$P_X(k) = \begin{cases} \frac{k-1}{36} & \text{if} \quad k \le 7 \\ \frac{12-k+1}{36} & \text{if} \quad k > 7, \end{cases}$$

while the values of c.d.f. $F_X = F_X(\alpha)$ are collected in the Table 1.

It is worth to add that if $X = X(\omega)$ is a random variable and f is a Borel function, i.e., a real function of a real variable such that $\{x : f(x) \le c\} \in \mathcal{B}$ for all $c \in R$ then the composition $f[X(\omega)]$ is also random variable.

Random Variable. Table 1 Values of c.d.f. $F_X = F_X(\alpha)$ in example 2

Interval for α	$(-\infty, 2)$	$<2,3)$	$<3,4)$	$<4,5)$	$<5,6)$	$<6,7)$
$F_X(\alpha)$	0	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$
Interval for α	$<7,8)$	$<8,9)$	$<9,10)$	$<10,11)$	$<11,12)$	$<12,+\infty)$
$F_X(\alpha)$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	1

Classification of r.v. 's

It is well known that any c.d.f. F is continuous, perhaps outside a countable set on the real line. For practical purposes we distinguish two classes of random variables. A r.v. is

1. *Discrete,* if its c.d.f is constant in all intervals designed by the points of its discontinuity
2. *Continuous,* if there exists a nonnegative integrable function f on R, called *density,* such that $F(\alpha) = \int_{-\infty}^{+\infty} f(x)dx$ for all $\alpha \in R$.

This classification is fully justified by two different analytical tools used in presentation and calculation of the r.v.'s. Distribution of a discrete r.v. X taking values x_i for some $i \in I$ is given by *probability mass function* $p_i = P(X = x_i)$ and its expectation is calculated by the formula $Ex = \sum_i x_i p_i$. Distribution of a continuous r.v. X is usually given by its *density function* f, while its expectation is calculated by the formula $Ex = \int_{-\infty}^{+\infty} xf(x)dx$.

About the Author

Professor Czesław Stępniak was formerly working at Agricultural University in Lublin, Poland (1972–2001) and heading Department of Statistics and Econometrics, Maria Curie-Skłodowska University in Lublin, Poland (2003–2009). During academic year 1987–1988 he was visiting Mathematical Sciences Institute of Cornell University, Ithaca, NY, as a senior scientist. Jointly with E. Torgersen, C. F. J. Wu and H. Heyer, Czesław Stępniak laid mathematical foundation to comparison of statistical experiments. He was also a recipient of two awards from Ministry of Science and Education in Poland "for a series of papers." Professor Stępniak authored more than 50 articles in peer-reviewed journals. He has two descendents Marek Niezgoda and Zdzisław Otachel.

Cross References

►Expected Value
►Measure Theory in Probability

References and Further Reading

Feller W (1971) An introduction to probability theory and its applications, vol 2. Wiley, New York
Kolmogorov AN (1933) Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin [English translation: Foundations of the theory of probability (2nd edn.), Chelsea, New York, 1956]
Prokhorov YuV (1985) Random variable. In: Vinogradov IM (ed) Mathematical encyclopedia, vol 5, 9–10. Soviet Encyclopedia, Moscow (in Russian)

Random Walk

Rabi Bhattacharya
Professor of Mathematics
The University of Arizona, Tucson, AZ, USA

The simple random walk $\{S_n : n = 0, 1, \ldots\}$, starting at an integer x, is a stochastic process on the integers, given by $S_0 = x$, $S_n = x + X_1 + \ldots + X_n (n \geq 1)$, where $X_n, n \geq 1$, is an independent Bernoulli sequence: $P(X_n = 1) = p$, $P(X_n = -1) = 1 - p = q$, $0 < p < 1$. In the case, $p = q = 1/2$, it is called the *simple symmetric random walk*, while if $p \neq 1/2$, it is *asymmetric*. By the binomial theorem, $P(S_n = y \mid S_0 = 0) = C^n_{\frac{(n+y)}{2}} p^{(n+y)/2} q^{(n-y)/2}$, if y and n are of the same parity, i.e., if either both are odd or both are even. Otherwise, $P(S_n = y \mid S_0 = 0) = 0$. Here $= C^n_m = n!/(m!(n-m)!)$.

For $c \leq x \leq d$ integers, the probability $\pi(x)$ that a simple random walk, starting at x, reaches c before d satisfies the equation

$$\pi(x) = p\pi(x+1) + q\pi(x-1) \quad \text{for } c < x < d,\ \pi(c) = 1,\ \pi(d) = 0, \qquad (1)$$

as shown by conditioning on the first step X_1. For the symmetric walk, the solution of this equation is $\pi(x) = (d-x)/(d-c)$. Since $\pi(x) \to 1$ as $d \to \infty$, the symmetric walk will reach the state c, starting from any state $x > c$, with probability one. By symmetry, it will reach every state with probability one. Iterating this argument one sees that, with probability one, the symmetric random walk visits every state infinitely often. That is, the walk is *recurrent*. For the asymmetric walk, the solution to (1) is $\pi(x) = (1 - (p/q)^{d-x})/(1 - (p/q)^{d-c})$. If $p < 1/2$, then the limit of this is 1 as $d \to \infty$ and, with probability one,

the random walk will visit c, starting from $x > c$. On the other hand, if $p > 1/2$, then $\pi(x) \to (q/p)^{x-c} < 1$, as $d \to \infty$. The probability of ever reaching d, starting from $x < d$ is obtained by symmetry as 1 if $p > 1/2$ and $(p/q)^{d-x}$ if $p < 1/2$. The asymmetric simple random walk is thus *transient*. Indeed, it follows from the strong law of large numbers (*SLLN*) that if $p > 1/2$, then $S_n \to \infty$ with probability one as $n \to \infty$; and $S_n \to -\infty$, with probability one, if $p < 1/2$. For these and other properties of the random walk, such as those described below, see Feller (1968, Chap. 3), Bhattacharya and Waymire (2009, Chap. 1), or Durrett (1995, Chap. 3). For additional information, refer to Billingsley (1968), and Spitzer (1964).

For computation of various probabilities associated with a simple random walk, the following result proved by D. Andre in 1887 is very useful: Consider the polygonal path of the random walk joining successive points (j, S_j), $(j+1, S_{j+1})$ $(j = 0, 1, \ldots, n-1)$ by line segments. Let $y > 0$. Then (a) the set of paths from $(0,0)$ to $(n, y-1)$ (n and $y-1$ of the same parity) which touch or cross the level y, is in one-one correspondence with (b) the set of all paths from $(0,0)$ to $(n, y+1)$ (*Reflection principle*). To prove this, let τ be the first time a path of the type (a) touches the level y prior to time n. Then replace the segment of the path from (τ, y) to $(n, y-1)$ by its mirror reflection about the level y. This gives a path of the type (b). Conversely, given any path of the type (b), reflect about y the segment of the path from (τ, y) to $(n, y+1)$. This gives a path of the type (a). Here is an application of this principle.

Example 1 (First passage time distribution of a simple random walk). Let y be a positive integer, and $F_{n,y}$ the event that the random walk, starting at zero, reaches y for the first time at time n, i.e., $F_{n,y} = \{S_j \neq y$, for $0 \le j < n, S_n = y\}$, n and y of the same parity. Altogether there are $C^n_{\frac{n+y}{2}}$ paths from $(0,0)$ to (n,y), each having probability $p^{(n+y)/2} q^{(n-y)/2}$. Of these, the number which cross or touch the level y prior to time n and for which $S_{n-1} = y-1$ is, by the reflection principle, $C^{n-1}_{\frac{n+y}{2}}$. Also the number for which $S_{n-1} = y+1$ is $C^{n-1}_{\frac{n+y}{2}}$. Subtracting these two from the number $C^n_{\frac{n+y}{2}}$ of all paths, one obtains, for all $y \neq 0$ (treating the case $y < 0$ by symmetry),

$$\begin{aligned} P(F_{n,y}) &= \left(C^n_{\frac{n+y}{2}} - 2C^{n-1}_{\frac{n+y}{2}}\right) p^{(n+y)/2} q^{(n-y)/2} \\ &= (|y|/n) C^n_{\frac{n+y}{2}} p^{(n+y)/2} q^{(n-y)/2} \qquad (2) \\ &\quad (n = |y|, |y|+2, |y|+4, \ldots). \end{aligned}$$

One may also consider the simple symmetric random walk $S_0 = x$, $S_n = x + X_1 + \ldots + X_n$ $(n \ge 1)$, in dimension $d \ge 1$, as a stochastic process on the d-dimensional lattice Z^d, with $X_n (n \ge 1)$ i.i.d. random vectors, taking values $\pm e_j (j = 1, \ldots, d)$, each with probability $1/2d$. Here e_j is the vector whose j-th coordinate is 1 and the remaining $d-1$ coordinates are zero. It was proved by G. Polya in 1921 that this walk is recurrent in dimensions 1,2, and transient in higher dimensions.

De Moivre (1756) obtained the normal approximation to the binomial probability $P(S_n = y \mid S_0 = 0)$, as a combinatorial result. The full potential of this was realized by Laplace (1812) who formulated and derived the far reaching central limit theorem (*CLT*, see ▶Central Limit Theorems). Apparently, Gauss knew about the normal distribution as early as 1794, and assuming this as the distribution of errors of measurement, he obtained his famous method of ▶least squares. Hence the name Gaussian distribution is often used for the normal distribution. The final version of the *CLT* for a *general random walk* $S_n = X_1 + \ldots + X_n$ $(n \ge 1)$, where X_n are arbitrary independent identically distributed (*i.i.d.*) random variables with mean zero and finite variance $\sigma^2 > 0$, was obtained by Lévy (1925): $n^{-1/2}(X_1 + \ldots + X_n)$ converges in distribution to the normal distribution $N(0, \sigma^2)$ with mean zero and variance σ^2, as $n \to \infty$. In physical terms, this result says the following: if time and length are rescaled so that in one unit of rescaled time there are a large number n of i.i.d. displacements of small rescaled lengths of order $1/\sqrt{n}$, then the random walk displacements over a period of time t will appear as Gaussian with mean zero and variance $t\sigma^2$, the increments over disjoint intervals being independent. That such a Gaussian process exists with continuous sample paths was proved rigorously by N.Wiener in 1923. This process is called the *Brownian motion*, following its implicit use by A. Einstein in 1905–1906 to describe the kinetic motion of colloidal molecules in a liquid, experimentally observed earlier by the botanist R. Brown. Interestingly, even before Einstein, Bachelier (1900) described the random movements of stocks by this Gaussian process. The statement that the rescaled random walk $S_n(n = 0, 1, 2, \ldots)$ converges in distribution to Brownian motion (see ▶Brownian Motion and Diffusions) was proved rigorously by M. Donsker in 1951, and this result is known as the functional central limit theorem (*FCLT*). Both the *CLT* and the *FCLT* extend to arbitrary dimensions d.

As consequences of the *FCLT*, one can derive many asymptotic results for the simple symmetric random walk given by the corresponding result for the limiting Brownian motion. Conversely, by evaluating combinatorially some probability associated with the random walk, one may derive the corresponding probability for the Brownian motion. A Brownian motion with variance parameter $\sigma^2 = 1$ is called a standard Brownian motion, and denoted $\{B_t : t \ge 0\}$ below.

Example 2 (Boundary hitting probability of Brownian motion). Let $c \le x \le d$ be arbitrary reals. Then, using the corresponding result for the scaled ▶random walk, one obtains

$$P(\{B_t : t \ge 0\} \text{ reaches c before } d \mid B_0 = x) = (d-x)/(d-c). \quad (3)$$

Example 3 (Arcsine law). Let U denote the amount of time in $[0,1]$ the Brownian motion spends above zero, i.e., U = Lebesgue measure of the set $\{t : 0 \le t \le 1 : B_t > 0\}$, given $B_0 = 0$. Consider the polygonal path of the simple symmetric random walk $S_j(j = 0,1,\ldots n)$, starting at zero. By combinatorial arguments, such as the reflection principle, one can calculate exactly the proportion of times the polygonal path lies above zero and, by the *FCLT*, this yields

$$P(U \le x) = (2/\pi) sin^{-1}\sqrt{x}\ (0 \le x \le 1). \quad (4)$$

Acknowledgments

The author acknowledges support from the NSF grant DMS 0806011.

About the Author

Rabindranath Bhattacharya received his Ph.D. from the University of Chicago in 1967. He has held regular faculty positions at the University of California at Berkeley, the University of Arizona, and Indiana University, and is currently a Professor of Mathematics at the University of Arizona. He has co-authored a number of graduate texts and monographs: *Normal Approximation and Asymptotic Expansions* (with R. Ranga Rao, John Wiley, 1976), *Stochastic Processes with Applications* (with Edward Waymire, Wiley, 1990), *Asymptotic Statistics* (with M. Denker, Birkhauser, 1990) and, more recently, *A Basic Course in Probability Theory* (with Ed Waymire, Springer, 2007), and *Random Dynamical Systems* (with M. Majumdar, Cambridge University Press, 2007). Among his more than 80 research articles in statistics, probability and mathematics, are Special Invited Papers in the *Annals of Probability* (1977), and the *Annals of Applied Probability* (1999). Professor Bhattacharya was Associate Editor for the following journals: *Annals of Probability* (1976–1981 and 2000–2002), *Annals of Applied Probability* (2006–2009), *Econometric Theory* (1989–1999), *Journal of Multivariate Analysis* (1986–1992), *Journal of Statistical Planning and Inference* (1984–1988, and 2000–2002), and *Statistica Sinica* (2002–2008). Currently he is Associate Editor for *Sankhya* (2009–present). Among his PhD students are several distinguished international scholars such as Ed Waymire (Oregon State University), Vic Patrangenaru (Florida State University), Gopal Basak (Indian Statistical Institute), Oesook Lee (Ewha Woman's University, Seoul, Korea). Rabi Bhattacharya is a Fellow of the IMS, and is a recipient of an Alexander Von Humboldt Forschungspreis, and a Guggenheim Fellowship.

Cross References

▶Brownian Motion and Diffusions
▶Central Limit Theorems
▶Ergodic Theorem
▶Limit Theorems of Probability Theory
▶Markov Processes
▶Monte Carlo Methods in Statistics
▶Statistical Modeling of Financial Markets
▶Statistical Quality Control
▶Stochastic Modeling Analysis and Applications
▶Stochastic Processes
▶Stochastic Processes: Classification

References and Further Reading

Bachelier L (1900) Théorie de la speculation. Ann sci école norm sup 17:21–86
Bhattacharya RN, Waymire EC (2009) Stochastic processes with applications. SIAM Classics in Applied Mathematics, vol 61. SIAM, Philadelphia
Billingsley P (1968) Convergence of probability measures. Wiley, New York
De Moivre A (1756) Doctrine of chance. London
Durrett R (1995) Probability: theory and examples. 2nd edn. Duxbury, Belmont
Feller W (1968) An introduction to probability theory and its applications. vol 1. 3rd edn. Wiley, New York
Laplace PS (1812) The theéorie analitique des probabilités. Veuve Courcier, Paris
Lévy P (1925) Calcul des probabilités. Gauthier-Villars, Paris
Spitzer F (1964) Principles of random walk. Van Nostrand. Princeton

Randomization

Cristiano Ferraz
Associate Professor
Federal University of Pernambuco, Recife, Brazil

Randomization is prescribed in several statistical procedures for reasons related not only to the assurance of scientific objectivity. Randomization, in essence, may be defined as a physical mechanism to assign probabilities to events. In probability sampling, such events are related to the selection of samples from finite populations. Samples are selected according to randomization processes that

guarantee selection probabilities for any specific sample. As a consequence, it also guarantees inclusion probabilities (of first and higher orders) for any element of the population. When a simple random sampling design (without replacement) is employed, for instance, any sample A, of size n from a population $\mathcal{U}$, of size N ($n < N$) have the same probability of been selected, and its inclusion probability of first order corresponds to the sample fraction, n/N. Restrictions imposed on the randomization lead to different sample designs (e.g., systematic sampling, Bernoulli sampling, Poisson sampling, and stratified sampling) and are responsible for their statistical properties. Similarly, in comparative experiments, the events are related to the assignment of treatments to experimental units. In experiments following the randomization principle, treatments are randomly assigned to available experimental units. This means such an assignment follows a specific randomization protocol. If, for instance, the protocol implies each group of size r from a total $n = tr$ available experimental units have the same probability of receiving a given treatment, the experiment is been conducted to compare t treatments according to a completely randomized design (with r genuine replicates). Once again, restrictions in the randomization lead to different designs (e.g., randomized complete block designs, Latin square designs, and split-plot designs).

Statistical methods of sampling and design of experiments rely on randomization to make valid design-based inferences. In both cases, inferences are supported by real reference distributions, induced by randomization. Its major role may be evident from appropriately derived linear models. A linear model for data from a ▶simple random sample, for instance, may be derived as follows. Let y_i be defined as the ith observation of a variable of interest Y under a simple random sample selection scheme (such as the traditional drawing of n balls, one at a time, without replacement, from an urn with N balls labeled from 1 to N). Hence, y_i can assume any value Y_k (the value of Y associated to element $k \in \mathcal{U}$). Let also be the following indicator variable defined:

$$\delta_{ik} = \begin{cases} 1, & \text{if } y_i = Y_k \\ 0, & \text{if } y_i \neq Y_k \end{cases}$$

Now, it is possible to write

$$y_i = \sum_{k \in \mathcal{U}} \delta_{ik} Y_k. \tag{1}$$

Let Y_k be rewritten as $\mu + (Y_k - \mu)$, with $\mu = \frac{\sum_{k \in \mathcal{U}} Y_k}{N}$. Then, (1) may be rewritten as

$$y_i = \sum_{k \in \mathcal{U}} \delta_{ik} [\mu + (Y_k - \mu)] = \mu + \sum_{k \in \mathcal{U}} \delta_{ik} (Y_k - \mu). \tag{2}$$

Define $\omega_i = \sum_{k \in \mathcal{U}} \delta_{ik} (Y_k - \mu)$ and (2) may be written as

$$y_i = \mu + \omega_i. \tag{3}$$

Expression (3) is the simplest linear model. According to this model, the ith observation of a variable of interest Y, observed on a simple random sample, may be regarded as the population mean (μ) plus a random term (ω_i) with statistical properties implied by the randomization scheme. For example, the description of "balls withdrawn from an urn" scheme allows one to write

$$P(\delta_{ik} = 1) = \frac{1}{N}, \text{ for any } k \in \mathcal{U};$$
$$P(\delta_{ik} = 1, \delta_{i'k'} = 1) = 0 \text{ for } k \neq k'; \text{ and,}$$
$$P(\delta_{ik} = 1, \delta_{i'k'} = 1) = \frac{1}{N(N-1)}, \text{ for } i \neq i' \text{ and } k \neq k'.$$

Therefore, the following properties hold:

$$E(\omega_i) = E\left(\sum_{k \in \mathcal{U}} \delta_{ik} (Y_k - \mu)\right) = \frac{1}{N} \sum_{k \in \mathcal{U}} (Y_k - \mu) = 0; \tag{4}$$

$$V(\omega_i) = E(\omega_i^2) = \sum_{k \in \mathcal{U}} \sum_{k' \in \mathcal{U}} (Y_k - \mu)(Y_{k'} - \mu) E(\delta_{ik} \delta_{ik'})$$
$$= \frac{1}{N} \sum_{k \in \mathcal{U}} (Y_k - \mu)^2 = \frac{(N-1)}{N} \sigma^2 \tag{5}$$

where $\sigma^2 = \frac{\sum_{k \in \mathcal{U}} (Y_k - \mu)^2}{N-1}$. It can also be shown that

$$Cov(\omega_i, \omega_{i'}) = -\frac{\sigma^2}{N} \tag{6}$$

Clearly, properties (4), (5), and (6) are consequences of the randomization process. They are not assumptions. Based on them, estimators such as the sample mean can be evaluated and proved unbiased with variances given as stated in many sampling books.

The ideas related to the role of randomization in scientific investigation were originally proposed by Fisher (1925, 1937). Since then, the relevance of the subject motivated works by several authors. Only few of them are referenced here as suggestions for further reading by limitation of space. Hinkelmann and Kempthorne (1994), for instance, explore the role of randomization in designed experiments by deriving linear models and examining in

depth the properties of the induced reference distributions. Särndal et al. (1992) emphasize the fundamental ideas of probability sampling giving attention to unbiased estimation. Finally, Tillé (2006) describes a series of computational algorithms (randomization protocols) to select samples according to the probabilistic method.

About the Author

Dr. Cristiano Ferraz is the author of the book *Sample design for surveys quality evaluation*, 2008, VDM Verlag Dr Mueller. He has been the director of undergraduate studies in statistics at Federal University of Pernambuco-UFPE, Brazil (2005–2009). Dr Ferraz is currently a faculty member of the UFPE graduate program in statistics, and has been working on sampling and design of experiments.

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Causation and Causal Inference
►Clinical Trials, History of
►Confounding and Confounder Control
►Design of Experiments: A Pattern of Progress
►Experimental Design: An Introduction
►Medical Research, Statistics in
►Medical Statistics
►Misuse of Statistics
►Permutation Tests
►Philosophical Foundations of Statistics
►Principles Underlying Econometric Estimators for Identifying Causal Effects
►Randomization Tests
►Research Designs
►Statistical Fallacies: Misconceptions, and Myths
►Statistics: An Overview
►Superpopulation Models in Survey Sampling

References and Further Reading

Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh
Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh
Hinkelmann K, Kempthorne O (1994) Design and Analysis of Experiments, vol 1. Wiley-Interscience, New York
Särndal C-E, Swensson B, Wretmann J (1992) Model assisted survey sampling. Springer, New York
Tillé Y (2006) Sampling algorithms. Springer, New York

Randomization Tests

Eugene S. Edgington
Professor Emeritus
University of Calgary, Calgary, AB, Canada

A randomization test is a permutation test (see ►Permutation Tests) that is based on randomization (random assignment), where the test is carried out in the following way. A test statistic (such as a difference between means) is computed for the experimental data (measurements or observations). Then the data are repeatedly divided or rearranged in a manner consistent with what the random assignment procedure would have produced if the treatments had no differential effect. The test statistic is computed for each of the resulting data permutations. Those data permutations, including the one for the experimental results, constitute the reference set for determining significance. The proportion of data permutations in the reference set having test statistic values greater than or equal to (or for certain test statistics, less than or equal to) the value for the experimental results is the *p*-value (significance or probability value). Determining significance on the basis of a distribution of test statistics generated by permuting the data is characteristic of all permutation tests; it is when the basis for permuting the data is random assignment (not random sampling) that a permutation test is called a randomization test.

The null hypothesis for a randomization test is that the measurement for each experimental unit (e.g., a subject or a plot of land) is the same under one assignment to treatments as under any alternative assignment. Thus, under the null hypothesis, assignment of experimental units to treatments randomly divides the measurements among the treatments. Each data permutation in the reference set represents the results that, if the null hypothesis is true, would have been obtained for a particular assignment. A randomization test is valid for any kind of sample, no matter how the sample is selected. This is an extremely important property because the use of non-random samples is common in experimentation, and parametric statistical tables (e.g., *t* and *F* tables) are not valid for such samples.

The validity of parametric statistical tables depends on random samples, and the invalidity of application to non-random samples is widely recognized. The random sampling assumption underlying the parametric significance tables is that of a sampling procedure that gives all possible samples of *n* individuals within a specified population

the same probability of being drawn. Arguments regarding the "representativeness" of a non-randomly selected sample are irrelevant to the question of its randomness: a random sample is random because of the sampling procedure used to select it, not because of the composition of the sample. Thus random selection is necessary to ensure that samples are random.

It must be stressed that violation of the random sampling assumption invalidates parametric statistical tables not just for the occasional experiment but for virtually all experiments involving statistical tests. A person conducting a poll may be able to enumerate the population to be sampled and select a random sample by a lottery procedure, but an experimenter would not have enough time, money, or information to take a random sample of the population of the world in order to make statistical inferences about people in general. Not many experiments in biology, education, medicine, psychology, or any other field use randomly selected subjects, and those that do usually concern populations so specific as to be of little interest. For instance, when human subjects for psychological experiments are selected randomly, often they are drawn from a population of students who attend a certain university, are enrolled in a particular class, and are willing to serve as subjects. Biologists and others performing experiments on animals generally do not even pretend to take random samples although they commonly use standard hypothesis testing procedures designed to test null hypotheses about populations. These well-known facts are mentioned here as a reminder of the rareness of random samples in experimentation and of the specificity of the populations on those occasions when random samples are taken.

In most experimentation the concept of population comes into the statistical analysis because it is traditional to discuss the results of a statistical test in terms of inferences about populations, not because the experimenter has sampled randomly some population to which he wishes to generalize. The population of interest to the experiment is likely to be one that cannot be sampled randomly. Random sampling by a lottery procedure, a table of random numbers, or any other device requires sampling a finite population, but experiments of a basic nature are not designed to find out something about a particular finite existing population. For example, with either animals or human subjects the intention is to draw inferences applicable to individuals already dead and individuals not yet born, as well as those who are alive at the present time. If we were concerned only with an existing population, we would have extremely biological laws because every minute some individuals are born and some die, producing a continual change in the existing population. Thus the population of interest in most experiments is not one about which statistical inferences can be made because it cannot be sampled randomly.

A number of desirable properties of randomization tests are a function of their intelligibility. A knowledge of calculus or other aspects of "advanced mathematics" is unnecessary for an experimenter to develop a new randomization test, using only his statistical knowledge of finite statistics involving combinations and permutations. The way in which random assignment is carried out in an experiment permits an experimenter to see whether the method of permuting the data is valid for that experiment for either simple or complex randomization tests. Neither the producer nor the consumer of randomization tests needs to rely on unknown authorities to justify their decision regarding the validity of a randomization test – or of its invalidity. For professors who enjoy making their students think instead of memorize, teaching randomization tests is enjoyable, and the pleasure of reasoning at a level that permits a student to develop new statistical tests that are custom-made to fit a new type of experimental design can appeal to ingenuity of many students.

About the Author

Dr. Eugene Sinclair Edgington is Emeritus Professor of Psychology at the University of Calgary in Calgary, Alberta, Canada. He received the B.S. (1950) and M.S. (1951) degrees in psychology from Kansas State University, and the Ph.D. (1955) degree in psychology from Michigan State University. He is a member of American Statistical Association and American Psychological Association. Professor Edgington has written numerous papers and is the author of the well known book *Randomization tests* (Marcel Dekker, Inc., 1980; 4th edition with Patrick Onghena, Chapman and Hall/CRC 2007).

Cross References

►Nonparametric Statistical Inference
►Permutation Tests
►Randomization
►Robust Inference

References and Further Reading

Edgington ES (1980) Randomization tests. Marcel Dekker, New York
Edgington ES (1987) Randomization tests, 2nd edn. Marcel Dekker, New York
Edgington ES (1995) Randomization tests, 3rd edn. Marcel Dekker, New York
Edgington ES, Onghena P (2007) Randomization tests, 4th edn. Chapman & Hall/CRC, New York

Rank Transformations

W. J. Conover
Horn Professor of Statistics
Texas Tech University, Lubbock, TX, USA

Statistics

The Science of Statistics is concerned with the analysis of data. This analysis may be as simple as presenting a graph, or finding an average. In more complex analyzes a statistical model may be assumed, and inferences may be made concerning the more general characteristics of the population of data from which a sample of data was obtained.

Parametric Versus Nonparametric

If the statistical model involves assumptions regarding the distribution of probabilities that govern the population of data then the resulting statistical methods are usually called "parametric." If the statistical model involves assumptions, but not assumptions regarding the distribution of probabilities governing the population, then the resulting statistical methods are usually called "nonparametric" or "distribution-free." Many of the best nonparametric methods involve the ranks of the data rather than the data itself. By ranks of the data it is meant that the smallest observation in the data set is given rank 1, the second smallest is given rank 2, and so forth.

Rank Transformation

Parametric methods usually have some optimum property for the parametric model, but are often inferior to the nonparametric method when the parametric model is not appropriate. It is convenient in those cases to "transform" the data to ranks, and to use the parametric method on the ranks instead of the data. These are called "rank transformation methods."

Example

In some cases the rank transformation method results in a nonparametric method. An early example involves Spearman's rho, published in 1904, which is simply the Pearson product-moment correlation coefficient r calculated on the ranks of the data rather than the data itself. Thus one can test the hypothesis of independence of two variables paired as in (X, Y), without any assumptions regarding the nature of the bivariate distribution from which they came, while the parametric model assumes a bivariate normal distribution. The observations on X are replaced by their ranks from 1 to n, the observations on Y are replaced by their ranks from 1 to n, and the ranks are placed in the original n pairs where the original data were. The usual correlation coefficient r is calculated on the ranks instead of the data, and the usual hypothesis test is conducted. In the case of independence of X and Y the distribution of Spearman's rho is asymptotically (for large n) the same as the distribution of Pearson's r. The exact distribution of Spearman's rho can be found, and is given in tables (see Conover 1999, for example), which is useful when n is small.

RT-1

There are several classifications of the transformation to ranks, as outlined by Conover and Iman (1981). The first type of rank transformation involves ranking all of the data together as one group, from smallest to largest, and then replacing the data in each of the original groups with their ranks. For example, in the case of two independent samples the observations in each sample are replaced with their ranks in the combined sample. Then for a test of equal means the two-sample t-test is computed on the ranks instead of the original data, and the test statistic is compared with the t-distribution in the usual way as an approximate test. With small sample sizes exact distributions of the test statistic can be obtained. This is equivalent to the *Mann-Whitney Test*, also known as the *two-sample Wilcoxon Test* (see ▶Wilcoxon–Mann–Whitney Test). An extension to the case of several independent samples is obvious, with the one-way ▶analysis of variance being computed on the ranks instead of the original data, and the F-tables being consulted for significance. This is equivalent to the *Kruskal-Wallis Test*. Details of these tests may be found in Conover (1999).

RT-2

A second type of rank transformation involves subsets of the data being ranked separately from other subsets, as in the correlation case mentioned earlier where the observations on X were ranked among themselves, and the observations on Y were ranked among themselves. The original data are replaced by their resulting ranks, and the statistic computed on the ranks. If independence is of interest, r is calculated, resulting in Spearman's rho, as mentioned earlier.

Another example of this second type of rank transformation is in the two-way layout, where the observations in each row are ranked among themselves only, and a two-way analysis of variance is computed on these ranks to see if there is a significant difference in the column means.

Thus we obtain a form of the *Friedman Test*. In the case of only two columns we obtain a form of the ►*Sign Test*.

RT-3

This brings us to a third type of rank transformation, where the ranks are determined after an appropriate re-expression of the data. Again consider pairs of data, n observations on a bivariate (X, Y) random variable. If the null hypothesis is equal means rather than independence, as was the case in the previous paragraph, the differences $X - Y$ are first computed, and then these differences are ranked on the basis of their absolute values $|X - Y|$ with the smallest absolute difference getting rank 1, and so on. Then the signs of the difference are applied to the ranks, and the one-sample t-test is computed on these signed ranks.

RT-4

The fourth type of rank transformation is an extension of the second type and third type combined. That is, subgroups of data are re-expressed, such as by subtracting a covariate or dividing by the consumer price index. Then each re-expressed subgroup is ranked by itself, and the standard parametric test is applied to the ranks. This could lead to an ►analysis of covariance by testing equality of means on the ranks of the re-expressed groups.

Another example of this fourth type of rank transformation is a nonparametric test for equal slopes presented by Hogg and Randles (1975). Several groups of paired data (X, Y) are first combined to find the least squares regression estimate $y = a+bx$. Then the residuals $(Y-y)$ from this model are ranked overall, and compared with the ranks of the $X's$ as described in their paper in a rank version of the parametric test for the same hypothesis.

Discussion

The rank transformation may result in a nonparametric test as indicated in the examples above, or the result may be a robust test such as when the first type of rank transform is applied to a two-way layout, or it may result in a test that is not even always robust such as applying the first type of rank transformation to a two-way layout with several observations per cell and trying to test for interaction.

About the Author

Dr. William Jay Conover received his Bachelor of Science in Mathematics from Iowa State University and his Master of Arts and Ph.D. in Mathematical Statistics from Catholic University of America. He is Paul Whitfield Horn Professor of Statistics, area of Information Systems and Quantitative Sciences, College of Business Administration, Texas Tech University. He was Elected Fellow of the American Statistical Association "for significant contributions to nonparametric statistics, for wide ranging and effective statistical consulting, and for excellence as a teacher and administrator" (1979). Among many his awards, Professor Conover was awarded the 1986 Don Owen Award and 1999 Wilks Medal. He has (co-)authored about 40 refereed papers and several books, including highly-regarded text *Practical Nonparametric Statistics* (3rd edition, John Wiley & Sons, 1999). His publications have been cited approximately 400 times each year for the last two decades, as reported by the Science Citation Index and the Social Sciences Citation Index. He is listed in in Who's Who in America (since 1987), and in Who's Who in the World, (since 1990–1991).

Cross References

►Measures of Dependence
►Nonparametric Models for ANOVA and ANCOVA Designs
►Parametric Versus Nonparametric Tests
►Scales of Measurement and Choice of Statistical Methods
►Statistical Fallacies: Misconceptions, and Myths
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Conover WJ (1999) Practical nonparametric statistics, 3rd edn. Wiley, New York
Conover WJ, Iman RL (1981) Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 33:124–129
Hogg RV, Randles RH (1975) Adaptive distribution-free regression methods and their application. Technometrics 17:399–407

Ranked Set Sampling

Douglas A. Wolfe
Professor and Chair
The Ohio State University, Columbus, OH, USA

Introduction

In experimental settings where data are collected with the goal of making inferences about some aspects of an underlying population it is always important to design the study in such a way as to obtain as much useful information as possible while minimizing the overall cost of the experiment. This is particularly true when the initial step in collecting these data is to select the particular units from the finite or infinite population on which measurements

are to be taken. In this context, the goal of minimizing experimental cost is most often equivalent to minimizing the sample size while still achieving the desired accuracy of the inferences that follow.

The most commonly used approach for collecting data from a population is that of a simple random sample (SRS). If the population is infinite, the observations in a SRS are independent and identically distributed random variables. Even if the population is finite and sampling is done without replacement so that the sample observations are no longer independent, there is still a probabilistic guarantee that each measurement in the SRS can be considered as representative of the population. Despite this assurance, there is a distinct possibility that a specific SRS might not provide a truly representative picture of the complete population and larger sample sizes might be required to guard against such atypical samples.

Statisticians have, of course, developed a number of ways to guard against such unrepresentative samples without resorting to unduly large sample sizes. Sampling designs such as stratified sampling, probability sampling, and ►cluster sampling all provide additional structure on the sampling process that improves the likelihood that a collected sample provides a good representation of the population while trying to control the sampling costs involved in both the selection of the units to include in the sample and the cost of making the actual measurements on the selected units.

A novel sampling approach with this goal in mind was introduced by McIntyre (1952, reprinted in 2005) for situations where taking the actual measurements for sample units is difficult (e.g., costly, destructive, time-consuming, etc.), but there are inexpensive mechanisms readily available for either informally or formally ranking a set of sample units. Sample data collected via such a preliminary ranking scheme are known in the literature as ranked set sample (RSS) data.

Balanced Ranked Set Samples

To obtain a RSS of k observations from a population, we first select a SRS of k^2 units from the population and randomly divide them into k subsets of k units each. Within each of these subsets, the k units are rank ordered (least to greatest) by some informative mechanism (such as visual comparisons, expert opinion, or through the use of auxiliary variables) that does not involve actual measurements on the attribute of interest for the sample units. The unit that is judged to be the smallest in the first of these rank ordered subsets is then included in the RSS and the attribute of interest is formally measured for this unit. This measurement is denoted by $X_{[1]}$, where [1] is used instead of the usual round bracket (1) for the smallest order statistic because $X_{[1]}$ is only judgment ranked to be the smallest among the k units in the first subset; it may or may not actually have the smallest measurement among the k units.

The same ranking process is used to judgment rank the second subset of k units and the item ranked as the second smallest of the k units is selected and its attribute measurement, $X_{[2]}$, is obtained and added to the RSS. From the third subset of size k we select the unit judgment ranked to be the third smallest and add its attribute measurement, $X_{[3]}$, to the RSS. This process continues until we add the attribute measurement for the unit ranked to the largest of the k units in the final subset of size k, denoted by $X_{[k]}$, to the RSS.

The resulting collection of k measurements $X_{[1]}, \ldots, X_{[k]}$ is called a *balanced ranked set sample of size k*, where the term balanced refers to the fact that we have collected one judgment order statistics for each of the ranks $1, 2, \ldots, k$. This entire process is called a *cycle* and k is the *set size*. To obtain a balanced RSS with a desired total number of measured observations (i.e., total sample size) $n = kq$, we repeat the entire process for q independent cycles, yielding the balanced RSS of size n : $X_{[i]j}$, $i = 1, \ldots, k$; $j = 1, \ldots, q$.

To illustrate the advantages of RSS over SRS, we consider the problem of estimation of a population mean. Let $X_1, \ldots, X_n$ be a SRS of size n from a distribution with mean μ and finite variance σ^2. Let $X_{[1]}, \ldots, X_{[n]}$ be the judgment order statistics for a balanced RSS from this distribution based on a single cycle with set size n. Let

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \overline{X}^* = \frac{1}{n}\sum_{j=1}^{n} X_{[j]}$$

be the corresponding SRS and RSS sample means, respectively. It is well known that the SRS mean $\overline{X}$ is unbiased for μ and that $Var(\overline{X}) = \frac{\sigma^2}{n}$. Dell and Clutter (1972) and Takahasi and Wakimoto (1968) showed that the RSS sample mean $\overline{X}^*$ is also an unbiased estimator of μ, and this is true even when there are errors in the ranking mechanism used to obtain the RSS data. Moreover, they provided an explicit formula for the variance of $\overline{X}^*$, namely,

$$\begin{aligned} Var(\overline{X}^*) &= \frac{\sigma^2}{n} - \frac{1}{n^2}\sum_{j=1}^{n}\left(\mu^*_{[j]} - \mu\right)^2 \\ &= Var(\overline{X}) - \frac{1}{n^2}\sum_{j=1}^{n}\left(\mu^*_{[j]} - \mu\right)^2, \end{aligned} \tag{1}$$

where $\mu^*_{[j]} = E(X_{[j]})$, for $j = 1, \ldots, n$.

Since $\sum_{j=1}^{n}\left(\mu^*_{[j]} - \mu\right)^2 \geq 0$, it follows from Eq. (1) that the variance of the SRS mean $\overline{X}$ is always at least as large as the

variance of the RSS mean $\overline{X}^*$, regardless of the accuracy of the ranking process. Thus the RSS mean $\overline{X}^*$ is a more precise estimator of the population mean μ than the SRS mean $\overline{X}$ based on the same number of measured observations. The gain in precision is a monotonically increasing function of the quantity $\sum_{j=1}^{n}\left(\mu^*_{[j]}-\mu\right)^2$, which is itself an increasing function of the accuracy of the judgment rankings. The more reliable the judgment ranking process, the more separated will be the judgment order statistic expectations, $\mu^*_{[r]}, r = 1,\ldots,n$, and the more improvement we can expect from using RSS instead of SRS. The worst-case scenario where there is no gain from using RSS occurs when $\mu^*_{[1]} = \mu^*_{[2]} = \cdots = \mu^*_{[n]} = \mu$, which corresponds to no information in our ranking process and thus completely random rankings.

Unbalanced Ranked Set Samples

For most settings, balanced RSS is the natural and preferred approach. There are, however, settings where measuring different numbers of the various judgment order statistics (unbalanced RSS) can lead to improved RSS procedures. This is the case, for example, when we are estimating the location parameter θ for a unimodal, symmetric distribution. In that setting when the ranking process is reasonably accurate, the optimal RSS would be to measure the sample median from each of the k sets, resulting in an extremely unbalanced RSS, and then estimate θ by the average of these k set sample medians. Stokes (1995), Bhoj (1997), and Kaur et al. (1997) were the first to point out the optimality of unbalanced RSS under appropriate conditions.

Other Factors Affecting RSS

Properties of procedures based on RSS data are affected by a number of factors that are unique to this sampling approach. First, and probably foremost, is the accuracy of the ranking process. While the balanced RSS sample mean is always as efficient as the SRS sample mean based on the same number of measured observations, this gain in efficiency can be minimal if the ranking process is not reasonably accurate. Moreover, the SRS sample mean can even be more efficient than estimators based on unbalanced RSS data when the ranking process is not reliable. There have been a number of approaches in the literature to modeling this degree of imperfection in the rankings. Frey (2007) provides a general discussion of these approaches and presents a broad class of imperfect ranking models that can be used to assess the effect of imperfect ranking on RSS procedures. A second factor that affects the properties of RSS procedures is the set size. Generally speaking, the effectiveness of RSS procedures improves with increasing set size but this is counterbalanced by the fact that the ranking accuracy generally decreases with increased set size. Finally, the relative costs of sampling, ranking, and measuring units can be an important factor to consider in evaluating RSS versus SRS competitors.

Resources

The original paper by McIntyre (1952, 2005) is a good place to start with understanding the motivation behind the RSS sampling approach. Kaur et al. (1995) and Patil (1995) provide general reviews of the research and applications involving RSS data and Wolfe (2004) gives a general introduction to RSS methodology with a special emphasis on nonparametric procedures. Cheng et al. (2004) have the only monograph/textbook on the subject.

About the Author

Douglas Wolfe is Professor and Chair, Department of Statistics at The Ohio State University in Columbus. He is a co-author of two well known texts: *Nonparametric Statistical Methods* (with Myles Hollander, Wiley-Interscience, 2nd edition 1999) and *Introduction to the Theory of Nonparametric Statistics* (with Ronald Randles, Wiley, 1979). Professor Wolfe is a two-time recipient of the Ohio State University Alumni Distinguished Teaching Award (1973–1974) and (1988–89). He was a member of the Noether Award Selection Committee, (2007–2010) and was Chair, ASA Nonparametric Statistics Section (2009).

Cross References

►Handling with Missing Observations in Simple Random Sampling and Ranked Set Sampling
►Order Statistics
►Ordered Statistical Data: Recent Developments
►Ranks
►Simple Random Sample

References and Further Reading

Bhoj DS (1997) New parametric ranked set sampling. J Appl Stat Sci 6:275–289
Cheng Z, Bai ZD, Sinha BK (2004) Ranked set sampling: theory and applications. Springer, New York
Dell TR, Clutter JL (1972) Ranked set sampling theory with order statistics background. Biometrics 28:545–555
Frey J (2007) New imperfect rankings models for ranked set sampling. J Stat Planning Infer 137:1433–1445
Kaur A, Patil GP, Sinha AK, Taillie C (1995) Ranked set sampling: an annotated bibliography. Environ Ecol Stat 2:25–54
Kaur A, Patil GP, Taillie C (1997) Unequal allocation models for ranked set sampling with skew distributions. Biometrics 53: 123–130

McIntyre GA (1952, 2005) A method for unbiased sampling, using ranked sets. Aust J Agri Res 3:385–390. Reprinted in The American Statistician 59(3):230–232
Patil GP (1995) Editorial: ranked set sampling. Environ Ecol Stat 2:271–285
Stokes SL (1995) Parametric ranked set sampling. Ann Inst Stat Math 47:465–482
Takahasi K, Wakimoto K (1968) On unbiased estimates of the population mean based on the sample stratified by means of ordering. Ann Inst Stat Math 20:1–31
Wolfe DA (2004) Ranked set sampling: an approach to more efficient data collection. Stat Sci 19(4):636–643

Ranking and Selection Procedures and Related Inference Problems

S. Panchapakesan
Professor Emeritus
Southern Illinois University, Carbondale, IL, USA

Introduction

A statistical ranking or selection procedure is typically called for when the experimenter (the decision-maker) is faced with the problem of comparing a certain number k of populations in order to make a decision about preferences among them.

Consider k populations, each characterized by the value of a parameter θ. In an agricultural experiment, for example, the different populations may represent different varieties of wheat and the parameter θ may be the average yield of a variety. The classical approach in this situation is to test the so-called homogeneity hypothesis H_0 that $\theta_1 = \cdots = \theta_k$, where the θ_i are the unknown values of the parameter for the k populations. In the case of the familiar one-way classification model, the populations are assumed to be normal with unknown means $\theta_1, \ldots, \theta_k$, and a common unknown variance σ^2. The homogeneity hypothesis H_0 is tested using Fisher's ▶analysis of variance (ANOVA) technique. However, this usually does not solve the real problem of the experimenter, which is not simply to accept or reject the homogeneity hypothesis. The real goal is often to choose the best population (the variety with the largest average yield). The inadequacy of the ANOVA is not in the design aspects of the procedure; it rather lies in the types of decisions that are made on the basis of the data. The attempts to formulate the decision problem in order to achieve this realistic goal of selecting the best treatment mark the beginnings of ranking and selection theory.

The formulation of a k-sample problem as a multiple decision problem enables one to answer the natural questions regarding the best populations. The formulation of multiple decision procedures in the framework of what has now come to be known as ranking and selection procedures began with the now-classic paper by Bechhofer (1954).

Basic Formulations of the Ranking and Selection Problem

We have k populations, $\Pi_1, \ldots, \Pi_k$, each indexed by a parameter θ, where the cumulative distribution function (cdf) of Π_i is $F(x;\theta_i)$ for $i = 1, 2, \ldots, k$. We assume that the family $\{F(x;\theta)\}$ is stochastically increasing in θ, i.e., $F(x;\theta_1) \geq F(x;\theta_2)$ for $\theta_1 \leq \theta_2$ for all x, and that the parameters can be ordered from the smallest to the largest. Denote the true ordered θ-values by $\theta_{[1]} \leq \theta_{[2]} \leq \ldots \leq \theta_{[k]}$. To fix ideas, we assume that larger the value of θ, more preferable is the population. Hence, the population associated with $\theta_{[k]}$ is called the *best* population. We assume that there is no prior information as to the correspondence between the ordered and the unordered θ_i. Ranking and selection problems have generally been formulated adopting one of two main approaches known as the *indifference-zone formulation* and the *subset selection formulation.*

In the indifference-zone formulation due to Bechhofer (1954), the goal is to select a fixed number of populations. Consider the basic goal of selecting the one best population. Based on samples of size n taken from each population, we seek a procedure to select one of the populations as the best. The natural procedure would be to compute estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ from each sample and claim that the population that yielded the largest $\hat{\theta}_i$ is the best population. Here, a correct selection occurs when the selected population is the best. We require a guaranteed minimum probability of a correct selection (PCS), denoted by P^*, whenever $\theta_{[k]}$ is sufficiently larger than $\theta_{[k-1]}$. Let $\delta = \delta(\theta_{[k]}, \theta_{[k-1]})$ denote a suitably defined measure of the separation between the populations associated with $\theta_{[k]}$ and $\theta_{[k-1]}$. Let $\Omega = \{\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)\}$. Define $\Omega(\delta^*) = \{\vec{\theta} \,|\delta(\theta_{[k]}, \theta_{[k-1]}) \geq \delta^* > 0\}$. For specified δ^* and $P^*(1/k < P^* < 1)$, it is required that

$$PCS \geq P^* \text{ whenever } \vec{\theta} \in \Omega(\delta^*). \tag{1}$$

To be meaningful, we choose $P^* > 1/k$; otherwise, the requirement (1) can be met by randomly choosing one of the populations as the best. The region $\Omega(\delta^*)$ of the parameter space Ω is called the *preference-zone* (PZ) as this is where we have strong preference for a correct selection.

The complement of the PZ is known as the *indifference-zone* (IZ), a region where we do not require a guaranteed PCS. The PCS in the PZ depends, in general, on the configuration of $\vec{\theta}$. In many cases, there is a *least favorable configuration* (LFC) of $\vec{\theta}$ for which the PCS attains a minimum over the PZ for any sample size. If we can make the PCS at the LFC equal to P^*, then the probability requirement (1) will be satisfied. The usual choices for $\delta = \delta(\theta_{[k]}, \theta_{[k-1]})$ are $\delta = \theta_{[k]} - \theta_{[k-1]}$ in the case of a location parameter and $\delta = \theta_{[k]}/\theta_{[k-1]}$ in the case of a scale parameter. In the case of nonnegative θ which is not a scale parameter, one may choose either of these two special forms depending on other aspects of the problem.

Bechhofer (1954) introduced the IZ formulation by considering k normal populations with means $\theta_1, \ldots, \theta_k$, and a common known variance σ^2. Here, $\delta = \theta_{[k]} - \theta_{[k-1]}$. Based on samples of size n from these normal populations, he proposed the natural selection procedure, say R_1, which selects the population that yielded the largest sample mean. The LFC for R_1 is $\theta_{[1]} = \cdots = \theta_{[k-1]} = \theta_{[k]} - \delta^*$. For a specified (δ^*, P^*), the minimum sample size needed to meet the probability requirement (1) is given by

$$n = \left\langle 2\left(\sigma H/\delta^*\right)^2 \right\rangle, \tag{2}$$

where $\langle b \rangle$ stands for the smallest integer greater than or equal to b, H satisfies

$$\Pr\{Z_1 \le H, \ldots, Z_{k-1} \le H\} = P^*, \tag{3}$$

and the Z_i are standard normal variables with equal correlation $\rho = 0.5$.

Some generalized goals that have been considered are: (I) Selecting the t best populations for $t \ge 2$, (a) in an ordered manner or (b) in an unordered manner, and (II) Selecting a fixed subset of size m that will contain at least s of the t best populations.

The first of these itself is a special case of the general ranking goal of Bechhofer (1954), which is to partition the set of k populations into s nonempty subsets $I_1, I_2, \ldots, I_s$ consisting of $k_1, k_2, \ldots, k_s$ $(k_1 + k_2 + \cdots + k_s = k)$ populations, respectively, such that for $\Pi_i \in I_\alpha$, $\Pi_j \in I_\beta$, $1 \le \alpha < \beta \le s$, we have $\theta_i < \theta_j$.

In the above general ranking problem, Fabian (1962) introduced the idea of Δ–correct ranking. Roughly speaking, a ranking decision is Δ–correct if wrongly classified populations are not too much apart. The special case of $s = 2$ and $k_1 = k - 1$ for a location parameter family is of interest. In this case, a Δ-correct ranking is equivalent to selecting one population Π_i for which $\theta_i > \theta_{[k]} - \Delta$, $\Delta > 0$; such a population is called a *good* population. The goal of selecting a good population has been considered by several subsequent authors.

In the normal means selection problem of Bechhofer (1954) mentioned previously, if the common variance σ^2 is unknown, a single-sample procedure does not exist. It can be seen from (2) that the minimum sample size n needed in order to satisfy the probability requirement (1) cannot be determined without the knowledge of the variance. In this case, a two-stage selection procedure is necessary to control the PCS. The first two-stage procedure for this problem was studied by Bechhofer et al. (1954). This procedure uses the first stage samples to obtain an estimate of σ^2.

In the subset selection formulation for selecting the best (i.e., the population associated with $\theta_{[k]}$), we seek a rule which will select a nonempty subset of random size that includes the best population. Here no assertion is made about which population in the selected subset is the best. The size S of the selected subset is determined by the sample data. In contrast with the IZ formulation, there is no specification of a PZ (or an IZ). The experimenter specifies P^*, the minimum PCS to be guaranteed no matter what the unknown values of the θ_i are. The selection rule is based on the estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$. The expected size of the selected subset is a performance characteristic of a procedure.

In the case of normal means problem, assuming a common known variance σ^2, Gupta (1954, 1965) proposed a procedure based on a sample of size n from each population. This rule, say R_2, selects population Π_i if the sample mean $\overline{X}_i$ from it satisfies:

$$\overline{X}_i \ge \overline{X}_{[k]} - d\sigma/\sqrt{n}, \tag{4}$$

where d is a positive constant to be chosen so that the minimum PCS is guaranteed. The LFC in this case is given by $\theta_1 = \theta_2 = \cdots = \theta_k$. By equating the PCS at the LFC to P^*, we get $d = \sqrt{2H}$, where H is given by (3).

When σ^2 is unknown, Gupta (1954) proposed the rule R_3 which is R_2 with σ^2 replaced by the pooled sample variance s^2 based on $\nu = k(n-1)$ degrees of freedom and a different constant d', which turns out to be the one-sided upper $(1 - P^*)$ equicoordinate point of the equicorrelated $(k-1)$–variate central t distribution with equal correlation $\rho = 0.5$ and the associated degrees of freedom ν.

Seal (1955) proposed a class of procedures that included Gupta's maximum-type procedure and an alternative (average-type) procedure that Seal advocated using. The superiority of Gupta's procedure under certain slippage configurations and with regard to certain optimality properties and its comparative ease in handling theoretical details accelerated the growth of the subset selection literature.

Subset selection can be thought of as a screening procedure towards selecting one population as the best. The IZ approach has no requirements regarding correct selection when the true parametric configuration lies in the IZ, whereas the (random size) subset selection formulation does not control the size of the selected subset. A modified formulation, called the *restricted subset selection*, puts an upper bound for the (random) size of the selected subset (see Santner 1975). Using the restricted subset selection formulation for the goal of selecting a subset of the k populations whose size does not exceed $m(1 \le m \le k)$ so that the selected subset includes at least one of the $t(1 \le t \le k-1)$ best with a guaranteed probability, Panchapakesan (2005) has provided a fresh look at the salient features of the IZ and subset selection approaches.

Over the last almost six decades, several aspects of selection and ranking have been investigated. Substantial accomplishments have been made concerning procedures for specific univariate and multivariate parametric families, conditional procedures, nonparametric procedures, sequential and multistage procedures, procedures for restricted families such as the IFR (increasing failure rate) and IFRA (increasing failure rate on the average) distributions, decision-theoretic developments, and Bayes and empirical Bayes procedures. For detailed accounts of these, see Gupta and Panchapakesan (1979, 1985, 1996), Panchapakesan (2006) and the references contained therein.

Inference Problems Associated with Ranking and Selection

One related inference problem is the point and interval estimation of the ordered parameters, $\theta_{[1]}, \ldots, \theta_{[k]}$. Some attempts have been made to combine selecting the population associated with $\theta_{[k]}$ and estimating $\theta_{[k]}$ with simultaneous probability control; see, for example, Rizvi and Lal Saxena (1974). Another related inference problem is the estimation of the PCS for a selection procedure; see, for example, Gupta et al. (1990). Another interesting problem is known as *estimation after selection* in which the interest is to estimate the parameter of the selected population in the case of a procedure for selecting one population, or to estimate a known function such as the average of the parameters of the selected populations in the case of subset selection. Here the object of inference depends on the sample data that are to be used in the procedure. Such a statistical procedure has been called a *selective inference procedure*. This is different from a nonselective inference procedure in which the identity of the object of inference is fixed and is determined before the data were obtained. For references to several papers dealing with this, see Gupta and Panchapakesan (1996) and Panchapakesan (2006).

In a given situation, we may use a natural rule to select the best population and may want to simultaneously test if the selected population is uniquely the best. Such a problem was first considered by Gutmann and Maymin (1987). Recently, a few papers have appeared dealing with location and scale parameter cases and selecting the best multinomial cell using inverse sampling. For a discussion of these, see Cheng and Panchapakesan (2009) and the references given therein.

Concluding Remarks

Our aim here is to give a brief introduction to ranking and selection procedures. As such, we have given only a few important references. Gupta and Panchapakesan (1979) provide a comprehensive survey of the literature up to the date with a bibliography of some 600 main references. For references to later developments and other books on the subject see Gupta and Panchapakesan (1985, 1996) and Panchapakesan (2006).

About the Author

Subrahmanian Panchapakesan is Professor Emeritus of Mathematics at the Southern Illinois University at Carbondale. He has published close to 90 journal articles, book chapters and reports, mostly on ranking and selection. He is a member of the ASA, IMS, and the International Statistical Institute (elected). Professor Panchapakesan is a co-author (with Shanti S. Gupta) of the well known text *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations* (John Wiley and Sons, 1979) and a co-editor (with N. Balakrishnan) of *Advances in Statistical Decision Theory and Applications* (Boston: Birkhäuser, 1997). He received the 2003 Thomas L. Saaty Prize for Applied Advances in the Mathematical and Management Sciences (awarded by the *American Journal of Mathematical and Management Sciences*). He is currently Associate Editor of *Communications in Statistics: Theory and Methods* and *Communications in Statistics: Simulation and Computation* (2002–), and an Editorial Board Member of *American Journal of Mathematical and Management Sciences* (1993–).

Cross References

►Analysis of Variance
►Explaining Paradoxes in Nonparametric Statistics
►Multiple Statistical Decision Theory
►Sequential Sampling

References and Further Reading

Bechhofer RE (1954) A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann Math Stat 25:16–39

Bechhofer RE, Dunnett CW, Sobel M (1954) A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. Biometrika 41:170–176
Cheng S-R, Panchapakesan S (2009) Is the selected population the best?: location and scale parameter cases. Comm Stat Theor Meth 38:1553–1560
Fabian V (1962) On multiple decision methods for ranking population means. Ann Math Stat 33:248–254
Gupta SS (1956) On a decision rule for a problem in ranking means. PhD thesis (Mimeo Ser 150), University of North Carolina, Chapel Hill
Gupta SS (1965) On some multiple decision (selection and ranking) rules. Technometrics 7:225–245
Gupta SS, Leu L-Y, Liang T (1990) On lower confidence bounds for PCS in truncated location parameter models. Comm Stat Theor Meth 19:527–546
Gupta SS, Panchapakesan S (1979) Multiple decision procedures: theory and methodology of selecting and ranking populations. Wiley, New York (Reprinted by Society for Industrial and Applied Mathematics, Philadelphia, 2002)
Gupta SS, Panchapakesan S (1985) Subset selection procedures: review and assessment. Am J Math Manage Sci 5: 235–311
Gupta SS, Panchapakesan S (1996) Design of experiments with selection and ranking goals. In: Ghosh S, Rao CR (eds) Design and analysis of experiments. Elsevier, Amsterdam, pp 555–585
Gutmann S, Maymin Z (1987) Is the selected population the best? Ann Stat 15:456–461
Panchapakesan S (2005) Restricted subset selection procedures for normal means: a brief review with a fresh look at the classical formulations of Bechhofer and Gupta. Comm Stat Theor Meth 34:1265–1273
Panchapakesan S (2006) Ranking and selection procedures. In: Balakrishnan N, Read C, Vidakovic B (eds) Encyclopedia of statistical sciences, vol 10, 2nd edn. pp 6907–6915
Rizvi MH, Lal Saxena KM (1974) On interval estimation and simultaneous selection of ordered location or scale parameters. Ann Stat 2:1340–1345
Santner TJ (1975) A restricted subset selection approach to ranking and selection problems. Ann Stat 3:334–349
Seal KC (1955) On a class of decision procedures for ranking means of normal populations. Ann Math Stat 26:387–398

Ranks

Elisabeth Svensson
Professor Emerita
Swedish Business School at Örebro University, Örebro, Sweden

Uni-variate Ranking

A common approach in nonparametric statistical method is to transform data to ranks. A ranking of n ordered observations $\{x_1 < x_2 <, \ldots, < x_n\}$ will be a set of n ranks $\{r_1 <, \ldots, < r_n\}$ where x_i is represented by the rank r_i in calculations. The rank sum is $\frac{1}{2}n(n+1)$, and $\sum r^2 = \frac{n^3 - n}{12}$. Hence the mean rank is $\frac{1}{2}(n+1)$ and the variance $\frac{1}{12}(n^2-1)$ assuming uniform distribution of all possible rankings. For untied observations the rank r_i equals the number of observations less than x_{i+1}, $i = 1, \ldots, n$.

Assessments on scales with a limited number of categories will produce groups of observations that are tied to the same category, which means that these observations will share the same rank value. The midrank of an observation in the i^{th} category, $i = 1, \ldots, m$, is then

$$\bar{r}_i = \sum_{v=1}^{i-1} x_v + \tfrac{1}{2}(x_i + 1),$$

where x_v denotes the vth category frequency, $v = 1, \ldots, m$. Then $\sum r^2 < \frac{1}{12}(n^3 - n)$ and the variance is decreased, the correction term being

$$t^{(X)} = \sum_{v=1}^{m} \left(x_v^3 - x_v\right).$$

The mid-ranks of the marginal distribution X of the Fig. 1 are (2, 9, 23, 41).

The calculations of the Wilcoxon–Mann–Whitney test statistics (see ►Wilcoxon–Mann–Whitney Test) of difference between two independent groups of data and of the Spearman rank-order correlation coefficient are based on this type of rank transformations.

Augmented Ranking

In the evaluation of paired assessments made on rating scales regarding reliability of inter- or intra-rater agreement but also regarding change in outcome, the pairs of data can be transformed to ranks taking account of the information given by the pairs of data. In this *augmented ranking approach*, (aug-rank), by Svensson, the ranks are tied to the pairs of data (X, Y), i.e., to the observations in the cells of a square contingency table alternatively to

X / Y	C_1	C_2	C_3	C_4	Total
C_4			1 (31; 49)	1 (50; 50)	2
C_3		2 (13.5; 31.5)	2 (29.5; 33.5)	14 (42.5; 41.5)	18
C_2	1 (3; 15)	1 (12; 16)	11 (23; 22)	3 (34; 29)	16
C_1	2 (1.5; 1.5)	8 (7.5; 6.5)	3 (16; 12)	1 (32; 14)	14
Total	3	11	17	19	50

Ranks. Fig. 1 Example of a frequency distribution of paired assessments of a four-point rating scale, and the pairs of augmented ranks $\left(\bar{R}_{ij}^{(X)}; \bar{R}_{ij}^{(Y)}\right)$

X / Y	C_1	C_2	C_3	C_4	Total
C_4				2	2
C_3			1	17	18
C_2			16		16
C_1	3	11			14
Total	3	11	17	19	50

Ranks. Fig. 2 The rank-transformable pattern of agreement, RTPA, uniquely defined by the two sets of marginal distributions

the points of a scatter plot of data from visual analogue scale assessments. This means that the augmented rank of the assessments X depends on the pairing with Y.

The mean augmented rank according to the assessments X is

$$\overline{R}_{ij}^{(X)} = \sum_{k=1}^{i-1}\sum_{l=1}^{m} x_{kl} + \sum_{l=1}^{j-1} x_{il} + \frac{1}{2}(1 + x_{ij})$$

for $1 \le i,j \le m$, where x_{ij} is the *ijth* cell frequency, i and $j = 1,\ldots,m$ and m is the number of categories. The augmented mean rank of the observations in the *ijth* cell according to assessments Y, $\overline{R}_{ij}^{(Y)}$, is defined correspondingly, see Fig. 1. This aug-rank approach makes it possible to identify and separately analyse a possible systematic component of observed disagreement from the occasional, noise, variability, (see ►Measures of Agreement). A complete agreement in all pairs of aug-ranks, $\overline{R}_{ij}^{(X)} = \overline{R}_{ij}^{(Y)}$, for all i and $j = 1,\ldots,m$ defines *the rank-transformable pattern of agreement* (RTPA), which is uniquely related to the two marginal distributions, see Fig. 2.

About the Author

Professor Svensson is Past President Swedish Society for Medical Statistics (2000–2002). She is an Elected member of the International Statistical Institute (2002), member of the International Association for Statistical Education (2000), International Society for Clinical Biostatistics; (Member of Executive Committee 2001–2004), and Elected member of the scientific board of Statistics Sweden (2003–).

Cross References

►Measures of Agreement
►Measures of Dependence
►Nonparametric Rank Tests
►Rank Transformations
►Ranking and Selection Procedures and Related Inference Problems
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Gibbons JD, Chakraborty S (2003) Nonparametric statistical inference, 4th edn (revised and expanded). Marcel Dekker, New York

Siegel S, Castellan NJ (1988) Nonparametric statistics for the behavioral sciences, 2nd edn. McGraw Hill, New York

Svensson E (1997) A coefficient of agreement adjusted for bias in paired ordered categorical data. Biometrical J 39:643–657

Rao–Blackwell Theorem

Arthur Cohen
Professor
Rutgers University, Piscataway, NJ, USA

The Rao–Blackwell Theorem (RB Theorem) attributed to C.R. Rao and David Blackwell links the notions of sufficient statistics and unbiased estimation. Let $\mathbf{X}$, a random vector represent the data. Assume the distribution of $\mathbf{X}$ depends on a parameter $\boldsymbol{\theta}$. A statistic $S(\mathbf{X})$ is said to be sufficient if the conditional distribution of $\mathbf{X}$ given S does not depend on $\boldsymbol{\theta}$. A statistic $T(\mathbf{X})$ is said to be an unbiased estimator of $g(\boldsymbol{\theta})$, a function of $\boldsymbol{\theta}$, if $E_{\boldsymbol{\theta}}T(\mathbf{X}) = g(\boldsymbol{\theta})$ where E stands for expected value. The RB Theorem, which is constructive says the following:

Let $U(\mathbf{X})$ be any unbiased estimator of $g(\boldsymbol{\theta})$ and let σ_U^2 be the variance of U. Let

$$W(\mathbf{X}) = E[U(\mathbf{X})|S(\mathbf{X})].$$

That is, $W(\mathbf{X})$ is the conditional expected value of $U(\mathbf{X})$ given $S(\mathbf{X})$. Then $W(\mathbf{X})$ is unbiased and $\sigma_U^2 \ge \sigma_W^2$, where σ_W^2 is the variance of W.

Evaluating unbiased estimators by their variance clearly corresponds to evaluating estimators using a squared error loss function. A well known extension of the RB Theorem is achieved by replacing a squared error loss function with any convex loss function.

The utility of the theorem is highlighted in situations where it is easy to find a simple unbiased estimator of $g(\boldsymbol{\theta})$. Sometimes this can be done using only a subset of the data and then the construction typically yields an excellent unbiased estimator. We proceed with some applications and an extension of the theorem.

One issue in quality control is to estimate the proportion of items produced whose measurements do not

meet specifications i.e. fall outside a given interval (L, U). Assuming measurements are normal with mean μ and variance σ^2 the quantity to estimate is

$$p = 1 - \int_L^U \varphi(z; \mu, \sigma^2) dz,$$

where $\varphi(z; \mu, \sigma^2)$ is the normal density. Based on a sample of size n, labeled $x_1, \ldots, x_n$ sufficient statistics are $\bar{x} = \sum x_i/n$, $s^2 = \sum(x - \bar{x})^2/(n-1)$. A simple unbiased estimator of p is

$$\begin{aligned} \hat{p} &= 0 \quad \text{if } L \le x_1 \le U. \\ &= 1 \quad \text{otherwise} \end{aligned}$$

Lieberman and Resnikoff (1955) Rao–Blackwellize $\hat{p}$ by deriving $E(\hat{p}|\bar{x}, s^2)$ resulting in what turns out to be the minimum variance unbiased estimator of p.

Cohen and Sackrowitz (1974) consider a common mean model. That is, consider two independent random samples, $x_1, \ldots, x_m$ from $N(\mu, \sigma_x^2)$ and $y_1, \ldots, y_n$ from $N(\mu, \sigma_y^2)$. In the course of estimating the common mean μ it was desired to seek an unbiased estimator of $\gamma = \sigma_x^2/(\sigma_x^2 + \sigma_y^2)$. For both m and n greater than or equal to 5, the sample could be split up in such a way to quickly find an unbiased estimator of γ. Then the simple estimator could be Rao–Blackwellized. This type of application is also suitable to find a good unbiased estimator of a correlation coefficient or intraclass correlation coefficient as was done in Olkin and Pratt (1958).

Cohen et al. (1985) consider the problem of estimating a quantile of a symmetric distribution. The cases of known and unknown centers of symmetry are studied. Convex combinations of a pair of ►order statistics from the sample are intuitive simple estimators of a quantile that exceeds 0.5. That is, suppose $X_{(1)} \le \ldots \le X_{(n)}$ are the order statistics from a population whose center of symmetry is known to be θ_0. Then the ordered values of $Y_j = |X_j - \theta_0|$ are sufficient statistics. The Rao–Blackwellized version of the convex combination then is a superior estimator of the quantile in terms of mean squared error.

Given two independent samples from populations with distributions characterized by parameters θ_1 and θ_2 respectively, suppose population i, $i = 1$ *or* 2 is selected if $\bar{X}_i$ is the larger sample mean. Suppose we wish to estimate the mean of the selected population. Note such a mean is a random variable. An estimator of a selected mean is said to be unbiased if its expected value equals the expected value of the selected mean. By taking an additional single observation from the selected population, and Rao–Blackwellizing it, using all the data Cohen and Sackrowitz (1989) display an estimator of the selected mean that is minimum variance conditionally unbiased under some assumptions regarding the underlying distributions.

An extension of the RB Theorem and the construction aspect of it appears in Brown et al. (1976). The extension gives a construction based on a conditional expectation of a decision procedure given the sufficient statistic that leads to a better procedure for all bowl shaped loss functions simultaneously even those that are not convex. Furthermore the construction preserves the property of median unbiasedness of any estimator.

About the Author

Arthur Cohen is a Professor of Statistics at Rutgers University, New Jersey, USA. He served as chairperson from 1968–1977. He is a Fellow of the Institute of Mathematical Statistics and of the American Statistical Association. He served as Editor of the *Annals of Statistics* from 1989–1991. He served as one of five editors of the *Journal of Multivariate Analysis* from 1978–1989. He also was an Associate editor of the *Journal of the American Statistical Association* and the *Journal of Statistical Planning and Inference*. He is the author of over 130 papers appearing in statistical journals.

Cross References

►Adaptive Sampling
►Bivariate Distributions
►Estimation
►Loss Function
►Minimum Variance Unbiased
►Properties of Estimators
►Statistical Quality Control
►Sufficient Statistics
►Unbiased Estimators and Their Applications

References and Further Reading

Brown LD, Cohen A, Strawderman WE (1976) A complete class theorem for Strict monotone likelihood ratio with applications. Ann Stat 4:712–722

Cohen A, Sackrowitz HB (1989) Two stage conditionally unbiased estimators of the selected mean. Stat Probab Lett 8: 273–278

Cohen A, Sackrowitz HB (1974) On estimating the common mean of two normal distributions. Ann Stat 2:1274–1282

Cohen A, Lo SH, Singh K (1985) Estimating a quantile of a symmetric distribution. Ann Stat 13:1114–1128

Lieberman GJ, Resnikoff GJ (1955) Sampling plans for inspection by variables. J Am Stat Assoc 50:457–516

Olkin I, Pratt J (1958) Unbiased estimation of certain correlation coefficients. Ann Math Stat 29:201–211

Rating Scales

Elisabeth Svensson
Professor Emerita
Swedish Business School at Örebro University, Örebro, Sweden

Rating scales have been used in psychology and psychophysics for over 100 years, and the use of rating scales and other kinds of ordered classifications is nowadays inter-disciplinary and unlimited. As there are no standardised rules for the operational definitions of qualitative variables, a considerable variety in types of rating scales for the same variable, in different applications, is common. A *rating scale* consists of a number of ordered categorical recordings of an item.

The verbal descriptive scale (VDS), also called the verbal rating scale, consists of a discrete number of verbally described ordered response categories, or description of criteria, grading the level of responses. The set of categories can also refer to a time scaling, also called a frequency-of-use scale, like "often, seldom, never," Fig. 1.

A *Likert scale* is a type of VDS, the descriptive categories being agreement levels to statements. Figure 2 shows two different operational definitions of perceived health, a VDS-5 scale and a Likert scale from the same ▶questionnaire, The Short-Form 36 (SF-36). The categories of the VDS represent five levels of perceived health, from excellent to poor. The Likert scale has one level of the variable, in this case "excellent health," and five levels of agreement with the statement of excellent health. The response categories except for a complete agreement with the statement (definitely true) contain no information about other levels of health. Hence the Likert scale assessments are just comparable with the binary responses: yes my health is excellent, no my health is not excellent.

A numerical rating scale (NRS) consists of a range of numerals indicating the ordered response levels without any description of the categories, except from the end points. Figure 3 shows a seven-point NRS of pain.

A visual analog scale (VAS) consists of a straight line anchored by the extremes of the variable being measured. The variable can be measured by a bipolar construct of the VAS, the anchors being opposing adjectives, or by a mono-polar scale, the anchors being "no sign at all" to "the most extreme alternative." A rating method that combines the verbal descriptor scale and the VAS, called the *graphic rating scale (GRS)* consists of a line with no breaks or divisions. There should be three to five discrete categories beneath the horizontal line, and the extreme categories should not be worded such that they are never employed, see Fig. 4.

A *pictogram* is a visual scale, the categories being faces or other pictures with different expressions illustrating the variable of interest.

VDS-6 intensity scale	VDS-4 grading of symptom	VDS-5 time scale
How much. . . ?	[ ] no evidence	**How often. . . ?**
[ ] Extremely high	[ ] slight signs	[ ] None of the time
[ ] Very high	[ ] moderate signs	[ ] A little of the time
[ ] Moderate	[ ] considerable signs	[ ] Some of the time
[ ] Slight		[ ] Most of the time
[ ] Very low		[ ] All of the time
[ ] non-existing		

Rating Scales. Fig. 1 Examples of verbal descriptive scale categories

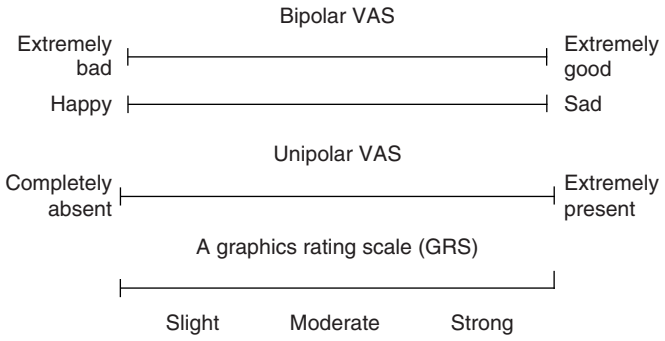
VDS-5 scale of health	Likert scale of excellent health
In general, would you say your health is	My health is excellent
[ ] Excellent	[ ] Definitely true
[ ] Very good	[ ] Mostly true
[ ] Good	[ ] Don't know
[ ] Fair	[ ] Mostly false
[ ] Poor	[ ] Definitely false

Rating Scales. Fig. 2 Examples of two different types of scales for assessment of health from the Short-Form-36 (SF-36), item 1 and item 11d, respectively

"How is your pain now?"

0	1	2	3	4	5	6
no pain at all						unbearable pain

Rating Scales. Fig. 3 Example of a numeric rating scale (NRS-7) of pain



Rating Scales. Fig. 4 Examples of Visual analogue scales (VAS) and a Graphic rating scale (GRS)

A *transitional scale*, the categories being *completely disappeared*, *much better*, *somewhat better, unchanged, somewhat worse, much worse* is useful when patient's perceived change after treatment is evaluated.

Assessments on rating scales, of any kind, produce ordinal data, the responses indicating only an ordering, although the use of numerical labelling could give a false impression of mathematical values. These so-called rank-invariant properties of ordinal data are well recognized, and several authors have stressed the fact that arithmetic operations are not appropriate for such data, therefore rank based statistical methods are recommended for analysis of data from rating scales.

About the Author

For biography *see* the entry ►Ranks.

Cross References

►Measures of Agreement
►Scales of Measurement
►Validity of Scales
►Variables

References and Further Reading

Teeling Smith G (ed) (1988) Measuring health: a practical approach. Wiley, Chichester

Svensson E (2000a) Concordance between ratings using different scales for the same variable. Stat Med 19(24):3483–3496

Svensson E (2000b) Comparison of the quality of assessments using continuous and discrete ordinal rating scales. Biomet J 42: 417–434

Record Statistics

Mohammad Ahsanullah[1], Valery B. Nevzorov[2]
[1]Professor
Rider University, Lawrenceville, NJ, USA
[2]Professor
St. Petersburg State University, St. Petersburg, Russia

In 1952, Chandler defined the so-called record times and record values and gave groundwork for a mathematical theory of records. For six decades beginning his pioneering work, about 500 papers and some monographs devoted to different aspects of the theory of records appeared. This theory relies largely on the theory of ►order statistics and is especially closely connected to extreme order statistics. Records are very popular because they arise naturally in many fields of studies such as climatology, sports, medicine, traffic, industry and so on. Such records are memorials of their time. The annals of records reflect the progress in science and technology and enable us to study the evaluation of mankind on the basic of record achievements in various areas of its activity. A large number of record data saved for a long time inspirited the appearance of different mathematical models reflecting the corresponding record processes and forecasting the future record results.

Definitions of Records

Let $X_1, X_2, \ldots$ be a sequence of random variables and $X_{1,n} \le X_{2,n} \le \ldots \le X_{n,n}$, $n = 1, 2, \ldots$, be the corresponding

order statistics. For any $n = 1, 2, \ldots$ denote also $M(n) = X_{n,n} = \max\{X_1, X_2, \ldots, X_n\}$ and $m(n) = X_{1,n} = \min\{X_1, X_2, \ldots, X_n\}$. Now one can define the classical upper record times $L(n)$ and upper record values $X(n)$ as follows:

$$L(1) = 1,\ X(1) = X_1 \text{ and then}$$

$$L(n+1) = \min\{j : X_j > X(n)\}, \quad X(n+1) = X_{L(n+1)},$$

$$n = 1, 2, \ldots \tag{1}$$

One can use the following alternative definitions:

$$L(1) = 1, \quad L(n+1) = \min\{j : X_j > M(L(n))\}, \ldots$$
$$n = 1, 2, \ldots, \tag{2}$$

and

$$X(n) = M(L(n)), \quad n = 1, 2, \ldots$$

If we replace the sign $>$ and $M(L(n))$ in (2) by $<$ and $m(L(n))$, respectively, we obtain the definitions of the lower record times and the lower record values. Indeed, the theories of upper and lower records coincide practically in all their details, since the lower records for the sequence $X_1, X_2, \ldots$ correspond to the upper records for the sequence $-X_1, -X_2, \ldots$ Using the sign $\geq$ in (1) instead of $>$ we introduce the so-called weak upper records, when any repetition of the previous record value is also considered as a new record. Analogically, the sign $\leq$ in (1) gives the opportunity to define the weak lower record times and the weak lower record values. Note that the theory of weak records has practical meaning only for sequences of the initial X's, which have discrete distributions.

The so-called kth records are a natural extension of records. The kth record times $L(n,k)$ and the kth record values $X(n,k)$ for any $k = 1, 2, \ldots$ are defined as follows:

$$L(1,k) = k, \quad L(n+1,k) = \min\{j > L(n,k) : X_j$$
$$> X_{j-k,j-1}\},\ n = 1, 2, \ldots, \tag{3}$$

and

$$X(n,k) = X_{L(n,k)-k+1,L(n,k)}, \quad n = 1, 2, \ldots \tag{4}$$

To be precise, (3) and (4) define the kth upper record times and the kth upper record values respectively. One can also introduce the kth lower record times and the kth lower record values changing the event $X_j > X_{j-k,j-1}$ in (3) by $X_j < X_{k,j-1}$ and replacing $X_{L(n,k)-k+1,L(n,k)}$ in (4) by $X_{k,L(n)}$. If $k = 1$ then definitions of kth record values $X(n,k)$ and kth record times $L(n,k)$ coincide with the definitions of $X(n)$ and $L(n)$ given in (1).

We will use $N(n)$ to denote the number of records among random variables $X_1, X_2, \ldots, X_n$, $n = 1, 2, \ldots$, and $N(n,k)$ will denote the number of kth records in a sequence $X_1, X_2, \ldots, X_n$ respectively.

Sequential Ranks and Record Indicators

The essential role in the theory of records play the sequential ranks $R(n)$, the record indicators ξ_j, $j = 1, 2, \ldots$, and the indicators of the *kth* records $\xi_j(k)$, $j = 1, 2, \ldots$, $k = 1, 2, \ldots$. The sequential rank $R(n)$ denotes the rank of X_n among $X_1, X_2, \ldots, X_n$, i.e.,

$$X_n = X_{R(n),n},\ n = 1, 2, \ldots$$

The record indicator ξ_j is defined as follows: $\xi_j = 1$ if X_j is a record value and $\xi_j = 0$ otherwise. Analogically, the indicator $\xi_j(k)$ can be defined for any $k = 1, 2, \ldots$ and $j \geq k$:

$\xi_j(k) = 1$, if X_j is the kth record value and $\xi_j(k) = 0$ otherwise. There are some useful relations between record indicators and sequential ranks. We will formulate some simple equalities for indicators of the upper records. Indeed, analogical results are also valid for the lower records. Note, that

$$\{\xi_j = 1\} = \{R(j) = j\} = \{X_j = M(j)\} = \{X_j = X_{j,j}\}$$
$$= \{X_j \text{ is a record value}\}$$

and for any $k = 1, 2, \ldots$ and $j \geq k$

$$\{\xi_j(k) = 1\} = \{R(j) \geq j - k + 1\} = \{X_{j-k+1,j} \geq X_{j-k,j-1}\}$$
$$= \{X_j > X_{j-k,j-1}\} = \{X_j \text{ is a } kth \text{ record}\}.$$

The record indicators allow us to give convenient relations for the numbers of records $N(n)$ and $N(n,k)$:

$$\{N(n) = m\} = \{\xi_1 + \ldots + \xi_n = m\},$$
$$n = 1, 2, \ldots,\ m = 1, 2, \ldots, n; \tag{5}$$
$$\{N(n,k) = m\} = \{\xi_k(k) + \ldots + \xi_n(k) = m\},$$
$$n = k, k+1, \ldots,\ m = 1, 2, \ldots, n-k+1. \tag{6}$$

The classical theory of records is connected with the situation when the initial sequence of random variables $X_1, X_2, \ldots$ is a sequence of independent random variables with a common continuous distribution function. In this case, sequential ranks and record indicators have a number of useful and rather convenient properties.

Theorem 1 For independent identically distributed random variables $X_1, X_2, \ldots$ with a continuous distribution function F the sequential ranks $R(1), R(2), \ldots$ are independent and $P\{R(n) = m\} = 1/n$, $m = 1, 2, \ldots, n$, $n = 1, 2, \ldots$

Theorem 2 Under conditions of Theorem 1, for any fixed $k = 1, 2, \ldots$, indicators $\xi_k(k), \xi_{k+1}(k), \ldots$ are independent and $P\{\xi_j(k) = 1\} = k/j$, $j = k, k+1, \ldots$

As a partial case of Theorem 2 it follows that under conditions of Theorem 1 record indicators $\xi_1, \xi_2, \ldots$ are independent and $P\{\xi_j = 1\} = 1 - P\{\xi_j = 0\} = 1/j$, $j = 1, 2, \ldots$

These results together with equalities (5) and (6) allow us to find that the distributions of $N(n)$ and $N(n,k)$ are expressed as distributions of sums of independent random variables. One can also see that under conditions of Theorem 1, there are some convenient relations for the *kth* record times $L(n,k)$ and, in particular, for the record times $L(n)$:

$$P\{L(n,k) > m\} = P\{N(m,k) < n\} = P\{\xi_k(k) + \xi_{k+1}(k) + \ldots + \xi_m(k) < n\} \quad (7)$$

and

$$P\{L(n) > m\} = P\{N(m) < n\} = P\{\xi_1 + \xi_2 + \ldots + \xi_m < n\}. \quad (8)$$

It follows from relations (7) and (8) that if $X_1, X_2, \ldots$ is a sequence of independent identically distributed random variables with any continuous distribution function F, then distributions of record times and numbers of records do not depend on F.

One more important situation in the classical record theory is connected with sequences of independent identically distributed random variables having a discrete distribution. Without loss of generality, we can suppose that X's take nonnegative integer values. For discrete distributions we introduce another type of record indicators. Let $\eta_n = 1$ if n is a record value in the sequence $X_1, X_2, \ldots$, that is there exists such $m = 1, 2, \ldots$ that $X(m) = n$, and $\eta_n = 0$ otherwise (compare with indicators ξ_n!). Analogously, for any $k = 1, 2, \ldots$ we can introduce indicators $\eta_n(k)$ for kth record values : $\eta_n(k) = 1$, if n is a kth record value in the sequence $X_1, X_2, \ldots$, and $\eta_n(k) = 0$ otherwise. The following results are valid for such type of indicators.

Theorem 3 Let $X, X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables taking values $0, 1, 2, \ldots$ with probabilities $p_n = P\{X = n\} > 0$, $n = 0, 1, 2, \ldots$ Then for any fixed $k = 1, 2, \ldots$ indicators $\eta_n(k)$, $n = 0, 1, 2, \ldots$, are independent and

$$P\{\eta_n(k) = 1\} = 1 - P\{\eta_n(k) = 0\} = (p_n / P\{X \geq n\})^k, \quad n = 0, 1, 2, \ldots$$

Indeed, under $k = 1$, as a partial case of this theorem, one gets that record indicators $\eta_0, \eta_1, \eta_2, \ldots$ are independent and $P\{\eta_n = 1\} = 1 - P\{\eta_n = 0\} = p_n / P\{X \geq n\}$, $n = 0, 1, 2, \ldots$

It is easy to see that under conditions of Theorem 3, we can express distributions of *kth* record values for discrete random variables via distributions of sums of independent indicators:

$$P\{X(n,k) > m\} = P\{\eta_0(k) + \ldots + \eta_m(k) < n\}, \quad m = 0, 1, 2, \ldots, \ n = 1, 2, \ldots, \quad (9)$$

and, in particular, under $k = 1$ one has equality

$$P\{X(n) > m\} = P\{\eta_0 + \ldots + \eta_m < n\}, \ m = 0, 1, 2, \ldots, \ n = 1, 2, \ldots \quad (10)$$

As an example, we can consider the case, when X's have the geometric distribution with some parameter $0 < p < 1$, that is $P\{X_j = n\} = (1-p)p^n$, $n = 0, 1, 2, \ldots$, for $j = 1, 2, \ldots$. In this situation, $P\{\eta_n(k) = 1\} = (1-p)^k$ and $P\{\eta_n(k) = 0\} = 1 - (1-p)^k$, for any $n = 0, 1, 2, \ldots$. It means that the sum $\eta_0(k) + \ldots + \eta_m(k)$ has the binomial $B(m+1, q)$ distribution with a parameter $q = (1-p)^k$. Hence,

$$P\{X(n,k) > m\} = \sum_{j=0}^{n-1} ((m+1)!/j! \ (m+1-j)!) q^j (1-q)^{m+1-j}, \quad \text{if } 1 \leq n \leq m+1,$$

and

$$P\{X(n,k) > m\} = 1, \text{ if } n > m+1.$$

It was mentioned above that for discrete distributions it is useful to introduce weak records together with classical (strong) record values. Weak records may arise, for example, in some sports competitions where any athlete who repeats the record achievement is also declared as a record-holder. If we consider X's having a common discrete distribution, it is useful to introduce one more type of record statistics. Let conditions of Theorem 3 be valid. We define random variables $\mu_0, \mu_1, \mu_2, \ldots$, where μ_n denotes the number of those weak records in the sequence $X_1, X_2, \ldots$ that are equal to n. The following result is valid.

Theorem 4 Let $X, X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables taking values $0, 1, 2, \ldots$ with probabilities $p_n = P\{X = n\} > 0$, $n = 0, 1, 2, \ldots$ Then for any fixed $k = 1, 2, \ldots$, random variables $\mu_0, \mu_1, \mu_2, \ldots$ are independent and

$$P\{\mu_n = m\} = (1 - r(n))(r(n))^m, \ n = 0, 1, 2, \ldots; \ m = 0, 1, 2, \ldots,$$

where

$$r(n) = p_n / P\{X \geq n\}.$$

Let $X_\omega(1), X_\omega(2), \ldots$ denote the weak (upper) record values in the sequence $X_1, X_2, \ldots$ Then for any $n = 1, 2, \ldots$

and $m = 0, 1, 2, \ldots$ the following relation is valid:

$$P\{X_\omega(n) > m\} = P\{\mu_0 + \mu_1 + \ldots + \mu_m < n\}. \quad (11)$$

Thus, we see that there are some convenient representations of record values, record times, numbers of records ((5–11) among them), which allow us to impress these record statistics in terms of sums of independent random variables.

Distributions of Record Times

Let us consider the classical case, when $X_1, X_2, \ldots$ are independent and have a continuous distribution function F. Using the independence of the corresponding record indicators $\xi_1, \xi_2, \ldots$ one gets for any $n = 1, 2, \ldots$ and $1 < j(1) < j(2) < \ldots < j(n)$ that

$$\begin{aligned} &P\{L(1) = 1, L(2) = j(2), \ldots, L(n) = j(n)\} = \\ &\quad P\{\xi_1 = 1, \xi_2 = 0, \ldots, \xi_{j(2)-1} = 0, \xi_{j(2)} = 1, \\ &\quad \xi_{j(2)+1} = 0, \ldots, \xi_{j(3)-1} = 0, \\ &\xi_{j(3)} = 1, \ldots, \xi_{j(n)-1} = 0, \xi_{j(n)} = 1\} = \\ &\quad P\{\xi_1 = 1\}P\{\xi_2 = 0\} \ldots P\{\xi_{j(2)-1} = 0\} \\ &\quad P\{\xi_{j(2)} = 1\}P\{\xi_{j(2)+1} = 0\} \ldots \\ &\quad P\{\xi_{j(3)-1} = 0\}P\{\xi_{j(3)} = 1\} \ldots \\ &\quad P\{\xi_{j(n)-1} = 0\}P\{\xi_{j(n)} = 1\} = \\ &\quad 1/(j(2) - 1)(j(3) - 1) \ldots (j(n) - 1)j(n). \end{aligned} \quad (12)$$

Note that here the joint distribution of record times does not depend on F. Now one can see from (12) that

$$\begin{aligned} P\{L(n) = m\} = \sum 1/(j(2) - 1)(j(3) - 1) \\ \ldots (j(n-1) - 1)(m - 1)m, \end{aligned}$$

where the sum is taken over all $j(2), j(3), \ldots, j(n-1)$, such that $1 < j(2) < j(3) < \ldots < j(n-1) < m$.

It follows from (12) also that

$$\begin{aligned} P\{L(n) = j(n)|L(n-1) = j(n-1), L(n-2) \\ = j(n-2), \ldots, L(2) = j(2), L(1) = 1\} \\ = j(n-1)/j(n)(j(n) - 1) \end{aligned}$$

and

$$P\{L(n) = j|L(n-1) = i\} = i/j(j-1), \quad n = 2, 3, \ldots, j > i.$$

Hence, we see that the sequence of record times $L(1), L(2), \ldots$ in the announced situation forms a Markov chain (see ▶Markov Chains).

It was mentioned above that record times are closely related to the random variables $N(n)$, since for any $n = 1, 2, \ldots$ and $m = 1, 2, \ldots$ the following equalities are valid:

$$P\{L(n) > m\} = P\{N(m) < n\}$$

and

$$P\{L(n) = m\} = P(N(m-1) = n-1,\ N(m) = n\}. \quad (13)$$

Equalities (5) and (13) allow us to express the distributions of $L(n)$ in terms of independent record indicators:

$$\begin{aligned} P\{L(n) = m\} &= P\{N(m-1) = n-1, \xi_m = 1\} \\ &= P\{N(m-1) = n-1\}/m \\ &= P\{\xi_1 + \xi_2 + \ldots \xi_m - 1 \\ &= n-1\}/m. \end{aligned} \quad (14)$$

Representations (5) and (14) of $N(m)$ and $L(n)$ give a possibility to find the distributions of these record statistics via the Stirling numbers of the first kind $S_n(k)$, which are defined by equalities

$$x(x-1)\ldots(x-n+1) = \sum_{k\geq 0} S_n(k)x^k.$$

It appears that

$$P\{N(m) = k\} = (-1)^k S_m(k)/m! = |S_m(k)|/m!,$$

$$k = 1, 2, \ldots, m,$$

and

$$P\{L(n) = m\} = |S_{m-1}(n-1)|/m!,\ m = n,\ n+1, \ldots$$

Representations (5) and (14) give the following formulas for the corresponding generating functions:

$$P_m(s) = Es^{N(m)} = s(1+s)(2+s)\ldots(m-1+s)/m!$$

and

$$\begin{aligned} Q_n(s) = Es^{L(n)} &= 1 - (1-s)\sum_{k=0}^{n-1}(-\log(1-s))^k/k! \\ &= \int_0^{-\log(1-s)} v^{n-1}\exp(-v)dv/(n-1)! \end{aligned} \quad (15)$$

Distributions of Record Values

Let us again consider the case when $X_1, X_2, \ldots$ are independent and have a continuous distribution function F. The record value $X(n)$ can be presented as $X_{L(n)}$, where $L(n)$ is

the corresponding record time with a generating function (15). One can see that then

$$P\{X(n) < x\} = P\{X_{L(n)} < x\} = P\{M(L(n)) < x\}, \quad (16)$$

where $M(n) = \max\{X_1, X_2, \ldots, X_n\}$ and $P\{M(n) < x\} = F^n(x)$. It follows from (15) and (16) that

$$\begin{aligned} P\{X(n) < x\} &= P\{M(L(n)) < x\} \\ &= \sum_{m \geq 0} P\{M(m) < x\} P\{L(n) = m\} \\ &= \sum_{m \geq 0} F^m(x) P\{L(n) = m\} = Q_n(F(x)) \\ &= \int_0^{-\log(1-F(x))} v^{n-1} \exp(-v) dv/(n-1)! \end{aligned}$$

The following result is valid for distributions of record values.

Theorem 5 Let $X(1) < X(2) < \ldots$ be the record values in a sequence of independent random variables having a common continuous distribution function F, and let $U(1) < U(2) < \ldots$ be the record values related to the uniform distribution on the interval $[0,1]$. Then for any $n = 1, 2, \ldots$ the random vector $(F(X(1)), \ldots, F(X(n)))$ has the same distribution as $(U(1), \ldots, U(n))$.

Corollary 1 Let $X(1) < X(2) < \ldots$ and $Y(1) < Y(2) < \ldots$ be, respectively, the record values in a sequence of independent random variables $X_1, X_2, \ldots$ having a common continuous distribution function F_1 and in a sequence of independent identically distributed random variables $Y_1, Y_2, \ldots$ with a continuous distribution function F_2. Then for any $n = 1, 2, \ldots$ the vector $(Y(1), Y(2), \ldots, Y(n))$ has the same distribution as the vector $(H_2(X(1)), H_2(X(2)), \ldots, H_2(X(n)))$, where $H_2(x) = G_2(F_1(x))$ and G_2 is the inverse function to F_2. Analogously, the vectors $(X(1), X(2), \ldots, X(n))$ and $(H_1(Y(1)), H_1(Y(2)), \ldots, H_1(Y(n)))$, where $H_1(x) = G_1(F_2(x))$ and G_1 is the inverse function to F_1, are identically distributed.

Let us consider the partial case of record values $Z(1) < Z(2) < \ldots$ related to the standard exponential $E(1)$ distribution (the case, when $F(x) = 1 - \exp(-x)$, $x > 0$). We get

$$P\{Z(n) < x\} = \int_0^x v^{n-1} \exp(-v) dv/(n-1)!,$$

that is, in this situation $Z(n)$ has the gamma-distribution with parameter n. It means that $Z(n)$ has the same distribution as the sum $v_1 + v_2 + \ldots + v_n$ of independent $E(1)$-distributed random variables $v_1, v_2, \ldots$ Moreover, for any $n = 1, 2, \ldots$ the vector $(Z(1), Z(2), \ldots, Z(n))$ has the same distribution as the vector $(v_1, v_1 + v_2, \ldots, v_1 + v_2 + \ldots + v_n)$. It means that the vector $(Z(1), Z(2) - Z(1), \ldots, Z(n) - Z(n-1))$ consists of independent elements and each of these elements has the standard exponential $E(1)$ distribution.

Combining the previous results, we can get the following representation for record values $X(1) < X(2) < \ldots$ related to any continuous distribution function F. Let G below denote the inverse function to F.

Representation 1 For any $n = 1, 2, \ldots$

$$(X(1), X(2), \ldots, X(n)) \stackrel{d}{=} (H(v_1), H(v_1 + v_2), \ldots, H(v_1 + v_2 + \ldots + v_n)),$$

where $v_1, v_2, \ldots$ are independent random variables having the exponential $E(1)$ distribution and $H(x) = G(1 - \exp(-x))$.

Taking into account the property of the exponential records it is not difficult to obtain the joint density function $f_n(x_1, x_2, \ldots, x_n)$ of the record values $Z(1), Z(2), \ldots, Z(n)$. It appears that

$$f_n(x_1, x_2, \ldots, x_n) = \exp(-x_n), \quad \text{if } 0 < x_1 < x_2 < \ldots < x_n, \text{ and } f_n(x_1, x_2, \ldots, x_n) = 0, \quad \textit{otherwise}.$$

In the general case, when $X_1, X_2, \ldots$ have a distribution function F and a density function f, the joint density function of the record values $X(1), X(2), \ldots, X(n)$ is given by the formula

$$f_n(x_1, x_2, \ldots, x_n) = r(x_1) r(x_2) \ldots r(x_n) (1 - F(x_n)), \quad x_1 < x_2 < \ldots < x_n,$$

where $r(x) = f(x)/(1 - F(x))$.

Now we consider the conditional distributions

$$\varphi(x | x_1, x_2, \ldots, x_n) = P\{X(n+1) > x | X(1) = x_1, X(2) = x_2, \ldots, X(n) = x_n\}, \; x_1 < x_2 < \ldots < x_n < x,$$

for record values $X(1) < X(2) < \ldots < X(n) < X(n+1)$. It appears that

$$\begin{aligned} \varphi(x | x_1, x_2, \ldots, x_n) &= P\{X(n+1) > x | X(n) = x_n\} \\ &= (1 - F(x))/(1 - F(x_n)), x > x_n. \end{aligned} \quad (17)$$

It is interesting that equality (17) does not need the continuity of the distribution function F. It follows from (17) that record values $X(1), X(2), \ldots$ form a Markov chain.

If we now consider discrete X's taking values $0, 1, 2, \ldots$ then (17) can be rewritten in the form

$$P\{X(n+1) > j | X(n) = m\} = P\{X > j\}/P\{X \geq m+1\}, \quad j > m \geq n-1.$$

It follows from the latter equality that in this case

$$P\{X(1) = j_1, X(2) = j_2, \ldots, X(n) = j_n\} = P\{X = j_n\}\omega(j_1)\omega(j_2)\ldots\omega(j_{n-1}), \quad 0 \le j_1 < j_2 < \ldots < j_{n-1} < j_n,$$

where $\omega(j) = P\{X = j\}/P\{X > j\}$.

The simplest discrete case is presented by the geometric distribution. The following result is valid.

Theorem 6 Let $X, X_1, X_2, \ldots$ be independent identically distributed random variables such that

$$P\{X = j\} = (1-p)p^{j-1},\ j = 1, 2, \ldots;\ 0 < p < 1,$$

and $X(1) < X(2) < \ldots$ be the record values in a sequence $X_1, X_2, \ldots$ Then the interrecord values $X(1), X(2) - X(1), X(3) - X(2), \ldots$ are independent and have the same geometric distribution as X.

Distributions of *k*th Record Values

The kth record values $X(n,k)$ are a natural extension of records $X(n)$. It is interesting that distributions of the *kth* records can be expressed via distributions of the classical record values. Really, together with a sequence of independent random variables $X_1, X_2, \ldots$ having a common distribution function F, let us consider one more sequence

$$Y_1 = \min\{X_1, X_2, \ldots, X_k\},$$
$$Y_2 = \min\{X_{k+1}, X_{k+2}, \ldots, X_{2k}\}, \ldots$$

Now let $X(n,k)$ be the kth record value based on X's and $Y(n)$ be the classical records based on the sequence $Y_1, Y_2, \ldots$ It appears that for any fixed $k = 2, 3, \ldots$ and any $n = 1, 2, \ldots$ the vector $(X(1,k), X(2,k), \ldots, X(n,k))$ has the same distribution as the vector $(Y(1), Y(2), \ldots, Y(n))$.

Note that this result is valid for discrete X's as well. One can immediately obtain some important results for the *kth* records taking into account the analogous results for the classical record values. For example, if $Z(n,k)$, $n = 1, 2, \ldots$, denote the *kth* records for the standard exponential distribution, then the vector $(Z(1,k), Z(2,k), \ldots, Z(n,k))$ has the same distribution as the vector $(v_1/k, (v_1 + v_2)/k, \ldots, (v_1 + v_2 + \ldots + v_n)/k)$, where $v_1, v_2, \ldots$ are the independent exponentially $E(1)$ distributed random variables. Hence, the following relation is valid for kth records related to a sequence of $X_1, X_2, \ldots$ with a continuous distribution function F.

Representation 2 For any $n = 1, 2, \ldots$

$$(X(1,k), X(2,k), \ldots, X(n,k)) \stackrel{d}{=} (H(v_1/k), H((v_1 + v_2)/k), \ldots, H((v_1 + v_2 + \ldots + v_n)/k),$$

where $v_1, v_2, \ldots$ are independent random variables having the exponential $E(1)$ distribution, $H(x) = G(1 - \exp(-x))$ and G is the inverse function to F.

Some useful results for the kth records follow immediately from representation 2 and analogous results for the classical records. Say, one gets that

$$P\{X(n,k) < x\} = \int_0^{-k\log(1-F(x))} v^{n-1}\exp(-v)dv/(n-1)!$$

and this equality is valid for any $k = 1, 2, \ldots$ and any continuous distribution function F.

Theorem 7 For any $k = 1, 2, \ldots$ the sequence $X(1,k)$, $X(2,k), \ldots$ forms a Markov chain and

$$P\{X(n+1,k) > x | X(n,k) = u\} = ((1 - F(x))/(1 - F(u)))^k, x > u.$$

More complete theory of records is given in monographs (Ahsanullah 1988, 1995, 2004; Ahsanullah and Nevzorov 2001; Arnold et al. 1998; Nevzorov 2000). Different results for record values can be found in references (Adke 1993; Ahsanullah 1978, 1979, 1981, 1987, 1988, 1995, 2004; Ahsanullah and Nevzorov 2001, 2004, 2005; Akhundov and Nevzorov 2008; Akhundov et al. 2007; Andel 1990; Arnold et al. 1998; Bairamov 2000; Balakrishnan and Nevzorov 2006; Ballerini and Resnick 1985, 1987; Berred et al. 2005; Biondini and Siddiqui 1975; Chandler 1952; Deheuvels 1984; Deheuvels and Nevzorov 1994; Dziubdziela and Kopocinsky 1976; Foster and Stuart 1954; Gulati and Padgett 2003; Gupta 1984; Haiman 1987; Houchens 1984; Nagaraja 1978, 1982; Nevzorov 1984, 1987, 1990, 1992, 1995, 2000; Nevzorov and Balakrishnan 1998; Nevzorov et al. 2003; Nevzorova et al. 1997; Pfeifer 1982, 1984, 1991; Renyi A 1962; Resnick 1973; Shorrock 1972a, b; Siddiqui and Biondini 1975; Smith 1988; Smith and Miller 1986; Stepanov 1992; Tata 1969; Vervaat 1973; Williams 1973; Yang 1975).

Acknowledgments

The work was partially supported by RFBR grant 09-01-00808.

Cross References

►Markov Chains
►Order Statistics
►Ordered Statistical Data: Recent Developments
►Sequential Ranks

References and Further Reading

Adke SR (1993) Records generated by Markov sequences. Stat Probab Lett 18:257–263

Ahsanullah M (1978) Record values and the exponential distribution. Ann Inst Stat Math 30A:429–433

Ahsanullah M (1979) Characterizations of the exponential distribution by record values. Sankhya B 41:116–121

Ahsanullah M (1981) On a characterization of the exponential distribution by weak homoscedasticity of record values. Biometr J 23:715–717

Ahsanullah M (1987) Two characterizations of the exponential distribution. Commun Stat Theory Meth 16:375–381

Ahsanullah M (1988) Introduction to record statistics. Ginn, Needham Heights

Ahsanullah M (1995) Record statistics. Nova Science, Commack

Ahsanullah M (2004) Record values – theory and applications. University Press of America, Lanham

Ahsanullah M, Nevzorov VB (2001) Ordered random variables. Nova Science, New York

Ahsanullah M, Nevzorov VB (2004) Characterizations of distributions by regressional properties of records. J Appl Stat Sci 13:33–39

Ahsanullah M, Nevzorov VB (2005) Order statistics. Examples and exercises. Nova Science, New York

Akhundov I, Nevzorov VB (2008) Characterizations of distributions via bivariate regression on differences of records. In: Records and branching processes. Nova Science, New York, pp 27–35

Akhundov I, Berred A, Nevzorov VB (2007) On the influence of record terms in the addition of independent random variables. Commun Stat Theory Meth 36:1291–1303

Andel J (1990) Records in an $AR(1)$ process. Ricerche Mat 39: 327–332

Arnold BC, Balakrishnan N, Nagaraja HN (1998) Records. Wiley, New York

Bairamov IG (2000) On the characteristic properties of exponential distribution. Ann Inst Stat Math 52:448–452

Balakrishnan N, Nevzorov VB (2006) Record values and record statistics. In: Encyclopedia of statistical sciences, 2nd edn. Wiley, 10, 6995–7006

Ballerini R, Resnick S (1985) Records from improving populations. J Appl Probab 22:487–502

Ballerini R, Resnick S (1987) Records in the presence of a linear trend. Adv Appl Probab 19:801–828

Berred A, Nevzorov VB, Wey S (2005) Normalizing constants for record values in Archimedean copula processes. J Stat Plan Infer 133:159–172

Biondini R, Siddiqui MM (1975) Record values in Markov sequences. In: Statistical inferences and related topics –2 Academic, New York, pp 291–352

Chandler KN (1952) The distribution and frequency of record values. J R Stat Soc Ser B 14:220–228

Deheuvels P (1984) The characterization of distributions by order statistics and record values – a unified approach. J Appl Probab 21:326–334 (Correction, 22, 997)

Deheuvels P, Nevzorov VB (1994) Limit laws for k-record times. J Stat Plan Infer 38:279–308

Dziubdziela W, Kopocinsky B (1976) Limiting properties of the kth record values. Zastos Mat 15:187–190

Foster FG, Stuart A (1954) Distribution free tests in time-series band on the breaking of records. J R Stat Soc B16:1–22

Gulati S, Padgett WJ (2003) Parametric and nonparametric inference from record breaking data. Springer, London

Gupta RC (1984) Relationships between order statistics and record values and some characterization results. J Appl Probab 21: 425–430

Haiman G (1987) Almost sure asymptotic behavior of the record and record time sequences of a stationary Gaussian process. In: Mathematical statistics and probability theory, vol A. D. Reidel, Dordrecht, pp 105–120

Houchens RL (1984) Record value theory and inference. PhD thesis, University of California, Riverside

Nagaraja HN (1978) On the expected values of record values. Aust J Stat 20:176–182

Nagaraja HN (1982) Record values and extreme value distributions. J Appl Probab 19:233–239

Nevzorov VB (1984) Record times in the case of nonidentically distributed random variables. Theory Probab Appl 29:808–809

Nevzorov VB (1987) Records. Theory Probab Appl 32:201–228

Nevzorov VB (1990) Generating functions for the kth record values – a martingale approach. Zap Nauchn Semin LOMI 184:208–214 (in Russian). Translated version in J Soviet Math 44: 510–515

Nevzorov VB (1992) A characterization of exponential distributions by correlation between records. Math Meth Stat 1:49–54

Nevzorov VB (1995) Asymptotic distributions of records in nonstationary schemes. J Stat Plan Infer 44:261–273

Nevzorov VB (2000) Records: mathematical theory. Translations of mathematical monographs, vol 194. Am Math Soc

Nevzorov VB, Balakrishnan N (1998) Record of records. In: Handbook of statistics, vol 16. Elsevier, Amsterdam, pp 515–570

Nevzorov VB, Balakrishnan N, Ahsanullah M (2003) Simple characterization of Student's t_2-distribution. Stat 52(part 3):395–400

Nevzorova LN, Nevzorov VB, Balakrishnan N (1997) Characterizations of distributions by extremes and records in Archimedean copula process. In: Advances in the theory and practice of statistics: a volume in honor of Samuel Kotz. Wiley, New York, pp 469–478

Pfeifer D (1982) Characterizations of exponential distributions by independent nonstationary record increments. J Appl Probab 19:127–135. (Correction, 19, 906)

Pfeifer D (1984) Limit laws for inter-record times from nonhomogeneous record values. J Organ Behav Stat 1:69–74

Pfeifer D (1991) Some remarks on Nevzorov's record model. Adv Appl Probab 23:823–834

Renyi A (1962) Theorie des elements saillants d'une suite d'observations. Colloquim on combinatorial methods in probability theory. Math. Inst., Aarhus Univ., Aarhus, Denmark, 1–10 August 1962, pp 104–117. See also: Selected papers of Alfred Renyi, vol. 3 (1976), Akademiai Kiado, Budapest, pp 50–65

Resnick SI (1973) Limit laws for record values. Stoch Proc Appl 1: 67–82

Shorrock RW (1972a) A limit theorem for inter-record times. J Appl Probab 9:219–223

Shorrock RW (1972b) On record values and record times. J Appl Probab 9:316–326

Siddiqui MM, Biondini RW (1975) The joint distribution of record values and inter-record times. Ann Probab 3:1012–1013

Smith RL (1988) Forecasting records by maximum likelihood. J Am Stat Assoc 83:331–338

Smith RL, Miller JE (1986) A non-Gaussian state space model and application in prediction of records. J R Stat Soc Ser B 48:79–88

Stepanov AV (1992) Limit theorems for weak records. Theor Probab Appl 37:586–590
Tata MN (1969) On outstanding values in a sequence of random variables. Z Wahrscheinlichkeitsth verw Geb 12:9–20
Vervaat W (1973) Limit theorems for records from discrete distributions. Stochast Proc Appl 1:317–334
Williams D (1973) On Renyi's record problem and Engel's series. Bull Lond Math Soc 5:235–237
Yang MCK (1975) On the distribution of the inter-record times in an increasing population. J Appl Probab 12:148–154

Recursive Partitioning

Hugh A. Chipman
Canada Research Chair in Mathematical Modelling, Professor
Acadia University, Wolfville, NS, Canada

Introduction

Recursive partition (RP) models are a flexible method for specifying the conditional distribution of a variable y, given a vector of predictor values x. Such models use a tree structure to recursively partition the predictor space into subsets where the distribution of y is successively more homogeneous. The terminal nodes of the tree correspond to the distinct regions of the partition, and the partition is determined by splitting rules associated with each of the internal nodes. By moving from the root node through to the terminal node of the tree, each observation is then assigned to a unique terminal node where the conditional distribution of y is determined. The two most common response types are continuous and categorical, with corresponding tasks often known as regression and classification.

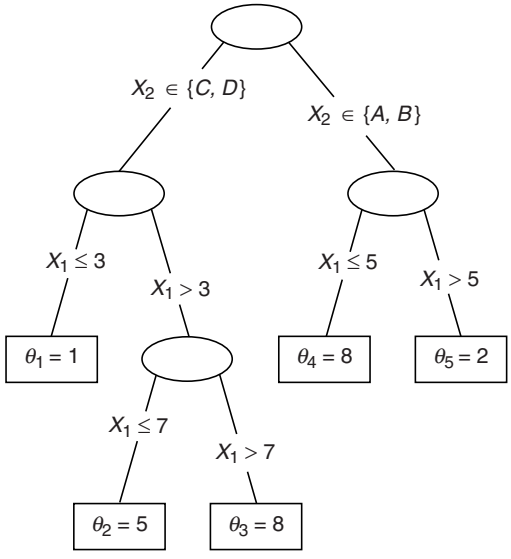
Given a data set, a common strategy for finding a good tree is to use a greedy algorithm to grow a tree and then to prune it back to avoid overfitting. Such greedy algorithms typically grow a tree by sequentially choosing splitting rules for nodes on the basis of maximizing some fitting criterion. This generates a sequence of trees each of which is an extension of the previous tree. A single tree is then selected by pruning the largest tree according to a model choice criterion such as cost-complexity pruning, cross-validation, or hypothesis tests of whether two adjoining nodes should be collapsed into a single node.

Early work in RP models includes Morgan and Sonquist (1963), who developed a recursive partitioning strategy (AID – Automatic Interaction Detection) for a continuous response. There were many offshoots of this work, including Kass (1980) and Hawkins and Kass (1982). Recursive partitioning models were popularized in the statistical community by the book "Classification and Regression Trees" by Breiman et al. (1984). RP models have also been developed in the machine learning community, with work by Quinlan on the ID3 (1986 and references therein) and C4.5 (1993) algorithms being among the most widely recognized.

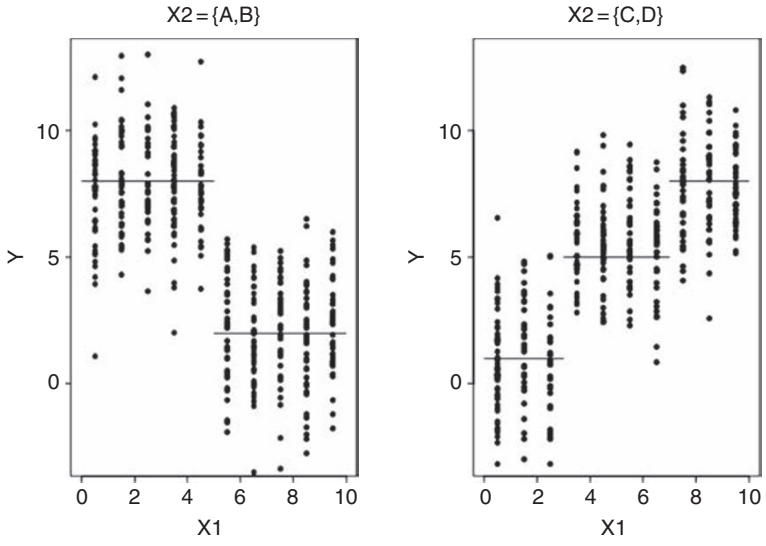
Structure of a RP model

A RP model describes the conditional distribution of y given a vector of predictors $x = (x_1, x_2, \ldots, x_p)$. This model has two main components: a tree T with b terminal nodes, and a parameter $\Theta = (\theta_1, \theta_2, \ldots, \theta_b)$ which associates the (possibly vector-valued) parameter θ_j with the jth terminal node. If x lies in the region corresponding to the jth terminal node then $y|x$ has distribution $f(y|\theta_j)$, where we use f to represent a parametric family indexed by θ_j. The model is called a regression tree or a classification tree according to whether the response y is quantitative or qualitative, respectively. An example of a RP model with binary splits is displayed in Fig. 1, and data sampled from its induced partition is displayed in Fig. 2.

Before describing the example tree, we discuss the general structure of a RP model for the case of a binary tree. A binary tree T subdivides the predictor space as follows: Each internal node has an associated splitting rule which uses a predictor to assign observations to either its left or



Recursive Partitioning. Fig. 1 A regression tree where $y \sim N(\theta, 2^2)$ and $x = (x_1, x_2)$



Recursive Partitioning. Fig. 2 A realization of 800 observations sampled from the tree model depicted in Fig. 1

right child node. The terminal nodes thus identify a partition of the predictor space according to the subdivision defined by the splitting rules. For quantitative predictors, the splitting rule is based on a split value s, and assigns observations for which $\{x_i \leq s\}$ or $\{x_i > s\}$ to the left or right child node respectively. For qualitative predictors, the splitting rule is based on a category subset C, and assigns observations for which $\{x_i \in C\}$ or $\{x_i \notin C\}$ to the left or right child node respectively.

Several assumptions have been made to simplify exposition. First, splitting rules are assumed to subdivide a region into two sub-regions, giving a binary tree. Second, only one predictor variable is assumed to be used for each splitting rule. Both these restrictions can be relaxed.

For illustration, Fig. 1 depicts a regression tree model where $y \sim N(\theta, 2^2)$ and $x = (x_1, x_2)$. x_1 is a quantitative predictor taking values in [0,10], and x_2 is a qualitative predictor with categories (A, B, C, D). The binary tree has nine nodes of which $b = 5$ are terminal nodes. The terminal nodes subdivide the x space into five nonoverlapping regions. The splitting variable and rule are displayed at each internal node. For example, the leftmost terminal node corresponds to $x_1 \leq 3.0$ and $x_2 \in \{C, D\}$. The θ_i value which identifies the mean of y given x is displayed at each terminal node. Note that θ_i decreases in x_1 when $x_2 \in \{A, B\}$, but increases in x_1 when $x_2 \in \{C, D\}$. A realization of 800 observations sampled from this model is displayed in Fig. 2.

If y were a qualitative variable, a classification tree model would be obtained by using an appropriate categorical distribution at each terminal node. For example, if y was binary with categories C_1 or C_2, one might consider the Bernoulli model $P(y \in C_1) = \theta = 1 - P(y \in C_2)$ with a possibly different value of θ at each terminal node. A standard classification rule for this model would then classify y into the category yielding the smallest expected misclassification cost. When all misclassification costs are equal, this would be the category with largest probability.

Learning the RP Model

To learn or estimate a RP model, we assume that a training sample consisting of tuples $(x_i, y_i), i = 1, \ldots, n$ is available. Both the tree T and the terminal node parameters Θ must be estimated using the training data.

For a fixed T, a common assumption is that the response values are i.i.d. within each terminal node. The data in each terminal node can be considered a separate sample, and conventional estimation techniques (e.g., maximum likelihood) yield familiar node parameter estimates $\hat{\theta}_j$ such as the sample mean for a continuous normal response and sample proportions for a categorical multinomial response.

Armed with a recipe for estimating Θ given T, we can now consider estimation of T. First, an objective function must be specified, providing a mechanism to assess the quality of a particular tree T. The log-likelihood of the training data is one such criterion. For a normal response model, the corresponding criterion would be the minimization of a residual sum of squares. For a multinomial response, the multinomial log-likelihood would be used. Ciampi (1991) was one of the first to develop a likelihood-based approach to RP models. Other criteria have been

proposed for specific response classes, such as the Gini index (Breiman et al. 1984) for a categorical response.

With an objective function quantifying the quality of a tree, the estimation problem becomes a search over all possible trees to optimize the objective. Although splitting rules for continuous x are real-valued, the objective function will only change when training points are moved among terminal nodes of the tree. Thus it is common to consider only splitting rules defined at data points, and require that each terminal node contain at least one training point. The search over the set of trees is thus a combinatorial search over a finite but very large discrete space.

The most common search algorithm is a greedy forward search, in which all training observations are initially grouped into a single node. The algorithm considers splitting into two child nodes, examining all possible splits on all possible variables. The splitting rule yielding the best value of the objective function (e.g., the smallest residual sum of squares when summed over the two child nodes) is selected. The procedure is repeated in each child node recursively until a large tree is grown.

Several strategies can be employed to decide how large a tree to grow. In the CHAID algorithm of Kass (1980), hypothesis tests were used to decide when to stop subdividing, yielding a final tree. Breiman et al. (1984) suggest growing a maximal tree, and then pruning away sibling nodes that do not significantly improve the objective function over the value assigned to their parent node. Their reasoning was that the forward greedy search might sometimes stop early, missing significant effects. For example, in the tree displayed earlier, no initial split leads to a large reduction in residual sum of squares because of the interaction pattern. Their backward pruning was facilitated by the idea of cost-complexity pruning, in which a modified objective function was minimized:

$$\text{Loss}(T;\alpha) = \text{RSS}(T) + \alpha|T|, \tag{1}$$

where $|T|$ represents the number of terminal nodes of the tree. Penalty parameter $\alpha \geq 0$ controls the trade-off between tree size and accuracy. Breiman et al. showed that (1) can be minimized as α increases from 0 to ∞ by considering a nested sequence of pruned trees, starting with the largest tree identified. The optimal α and a corresponding tree are selected so as to minimize a cross-validated estimate of the objective function.

While other methods for identifying the best tree have been proposed, the greedy forward search is quick and can be quite effective.

Strengths and Weaknesses of RP Models

The structure of RP models enables them to identify *interactions*. For instance, in Figs. 1 and 2, we see an interaction effect between X_1 and X_2: If $X_2 = \{A, B\}$ then response y decreases with increasing X_1. If $X_2 = \{C, D\}$ then response y increases with increasing X_1. This is perhaps the greatest strength of RP models, and one of the reasons they are used for exploratory data analysis.

This strength is also a weakness. If the relation between predictors and response is *additive*, very large trees will be needed to capture this relationship. For instance, if

$$y = x_1 + x_2 + x_3 + x_4 + x_5 + \text{error},$$

then a tree with 32 terminal nodes will be required to even approximate this function with a single step along each of the five predictor axes.

Trees are popular among practitioners because of their *interpretability*. It is natural to interpret the sequence of conditions leading to a terminal node of a tree. Care must be taken with such interpretations, especially if dependencies exist among predictors. In such cases, multiple trees with different splits on different variables may fit the data equally well.

In addition to dealing with mixed predictor types, RP models can handle missing values of predictors via several strategies. For missing predictor values in the training data, one could (i) treat "missing" as a new category for a categorical predictor, or (ii) identify surrogate splitting variables that produce splits similar to a missing predictor. If predictor values are missing when making predictions for new observations, either of these strategies may be employed, or one may terminate the branching process when a missing value is needed in a branch, and base predictions on the interior node.

The most common form of RP models utilize a single variable for each splitting rule. This *axis alignment* aids in interpretability, but can be a weakness if variation in the response occurs along a linear combination of predictors, rather than along the axes. The additive function of five variables mentioned above is an example of this.

By virtue of subdividing the data into smaller subgroups, an RP model can suffer from *sparsity*, especially if more complex statistical models are utilized in the terminal nodes. For instance, a significant challenge in modifying RP models for survival data with censoring (LeBlanc and Crowley 1993) is the pooled nature of Kaplan–Meier estimates (see ►Kaplan-Meier Estimator) of the survival curve. This data sparsity is one of the primary reasons for the use of simple models in terminal nodes.

A weakness of RP models is *sensitivity* of results to small data perturbations. Breiman (1996) demonstrated that when RP models were fit to bootstrap samples of the data, there could be substantial variation in tree structure. While this would seem to be a weakness, Breiman leveraged this idea to produce Ensemble methods discussed below in section ►"Ensembles of Trees".

Because of the greedy nature of the search over the space of trees, inference for the resultant model is difficult. Although confidence intervals and hypothesis tests can easily be constructed conditional on a specific tree T, the adaptive nature of the learning algorithm means that the statistical properties of estimators, intervals and tests will be seriously undermined. Methods that take account of the search include adjustments for multiple testing (Hawkins and Kass 1982) and Bayesian approaches (Chipman et al. 1998; Denison Mallick and Smith 1998).

Extensions

The popularity of RP models has lead to a number of extensions and the development of related methods.

A variety of search strategies have been proposed as alternatives to the greedy forward stepwise approach. These include the use of stochastic search optimizers such as genetic algorithms (Fan and Gray 2005) and simulated annealing (Sutton 1991; Lutsko and Kuijpers 1994) and MCMC (Chipman et al. 1998; Denison et al. 1998). Tibshirani and Knight (1999) used the bootstrap to perturb data before executing a greedy search.

Variations on the tree structure have also been considered, including splitting rules based on linear combinations of real-valued predictors (Loh and Vanichsetakul 1988). Some RP algorithms (e.g., AID) allow nodes to have more than two child nodes, complicating the search but sometimes making interpretation clearer. Quinlan's C4.5 splits categorical predictors by generating a different child node for each categorical level of the corresponding predictor.

The statistical model in terminal nodes has also been extended to richer models, such as linear regression (Alexander and Grimshaw 1996; Chipman et al. 2002), ►generalized linear models (Chipman et al. 2003), and Gaussian process models (Gramacy and Lee 2008).

Ensembles of Trees

RP models have been used as a "base learner" in a number of algorithms that seek to achieve greater predictive accuracy by combining together multiple instances of a model.

In noticing the sensitivity of trees to small perturbations, Breiman (1996) developed a strategy known as bootstrap aggregation or "Bagging" for generating multiple trees and combining them to achieve greater prediction accuracy. For instance, with a continuous response, each bootstrap tree would be used to generate predictions at a particular test point, and these predictions would be averaged to form an ensemble prediction.

A further enhancement led to Random Forests (Breiman 2001). Additional variation in the search algorithm was introduced by randomizing the choice of predictor in splitting rules. This led to a richer set of trees, and could further improve predictive accuracy.

Another form of ensemble model using RP models is boosting (Freund and Schapire 1997). In this algorithm, a sequence of RP models are learned, each depending on those already identified via data weights that depend on predictive accuracy of earlier RP models. These weights encourage the next RP model to better fit those observations that have been incorrectly classified. At the end of the boosting sequence, an ensemble prediction is generated by a weighted combination of predictions from each learner in the ensemble.

Although neither boosting or random forests require that the base learner be a RP model, these have yielded the most popular and successful form of ensemble model.

Related Work

A model closely related to RP models is the hierarchical mixture of experts model (Jordan and Jacobs 1994). In this model, a different logistic function of the predictors is used in each interior node to probabilistically assign data points to the left and right children. In doing so, the hard boundaries associated with splitting rules are replaced with soft decisions indexed by continuous parameters. In terminal nodes, predictions are given by ►logistic regression. Tree size and topology is typically fixed in advance, and the tree learning algorithm becomes a continuous optimization problem.

About the Author

Hugh A. Chipman is Professor and Canada Research Chair in Mathematical Modelling, Acadia University Department of Mathematics and Statistics. He is Editor-Elect (2010) and will be Editor (2011–2014), *Technometrics*. He was elected as a Fellow of the American Statistical Association (2008) and received the CRM-SSC award (Canada, 2009).

Cross References

►Data Mining
►Exploratory Data Analysis
►Interaction
►Kaplan-Meier Estimator
►Logistic Regression

References and Further Reading

Alexander WP, Grimshaw SD (1996) Treed regression. J Comput Graph Stat 5:156–175
Breiman L (1996) Bagging predictors. Mach Learn 24:123–140
Breiman L (2001) Random forests. Mach Learn 45:5–32
Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees, Wadsworth, Belmont
Chipman HA, George EI, McCulloch RE (1998) Bayesian CART model search. J Am Stat Assoc 93:935–948
Chipman HA, George EI, McCulloch RE (2002) Bayesian treed models. Mach Learn 48:299–320
Chipman HA, George EI, McCulloch RE (2003) Bayesian treed generalized linear models. In: Bernardo JM, Bayarri M, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) Bayesian statistics vol 7. Oxford University Press, Oxford
Ciampi A (1991) Generalized regression trees. Comput Stat Data Anal 12:57–78
Denison D, Mallick B, Smith AFM (1998) A Bayesian CART algorithm. Biometrika 85:363–377
Fan G, Gray JB (2005) Regression analysis using TARGET. J Comput Graph Stat 14:206–218
Gramacy RB, Lee HKH (2008) Bayesian treed Gaussian process models with an application to computer modeling. J Am Stat Assoc 103:1119–1130
Hawkins DM, Kass GV (1982) Automatic interaction detection. In: Hawkins DM (ed) Topics in applied multivariate analysis. Cambridge University Press, Cambridge
Jordan MI, Jacobs RA (1994) Mixtures of experts and the EM algorithm. Neural Comput 6:181–214
Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. Appl Stat 29:119–127
LeBlanc M, Crowley J (1993) Survival trees by goodness of split. J Am Stat Assoc 88:457–467
Loh W-Y, Vanichsetakul N (1988) Tree-structured classification via generalized discriminant analysis. J Am Stat Assoc 83: 715–725
Lutsko JF, Kuijpers B (1994) Simulated annealing in the construction of near-optimal decision trees. In: Cheeseman P, Oldford RW (eds) Selecting models from data: AI and statistics IV. Springer, New York, pp 453–462
Morgan JA, Sonquist JN (1963) Problems in the analysis of survey data and a proposal. J Am Stat Assoc 58:415–434
Quinlan JR (1986) Induction of decision trees. Mach Learn 1: 81–106
Quinlan JR (1993) C4.5: tools for machine learning. Morgan Kauffman, San Mateo
Sutton C (1991) Improving classification trees with simulated annealing. In: Keramidas E (ed) Proceedings of the 23rd symposium on the interface, Interface Foundation of North America
Tibshirani R, Knight K (1999) Model search by bootstrap 'bumping'. J Comput Graph Stat 8:671–686

Regression Diagnostics

Shuangzhe Liu[1], Alan H. Welsh[2]
[1]Associate Professor, Faculty of Information Sciences and Engineering
University of Canberra, Canberra, ACT, Australia
[2]E.J. Hannan Professor of Statistics
Australian National University, Canberra, ACT, Australia

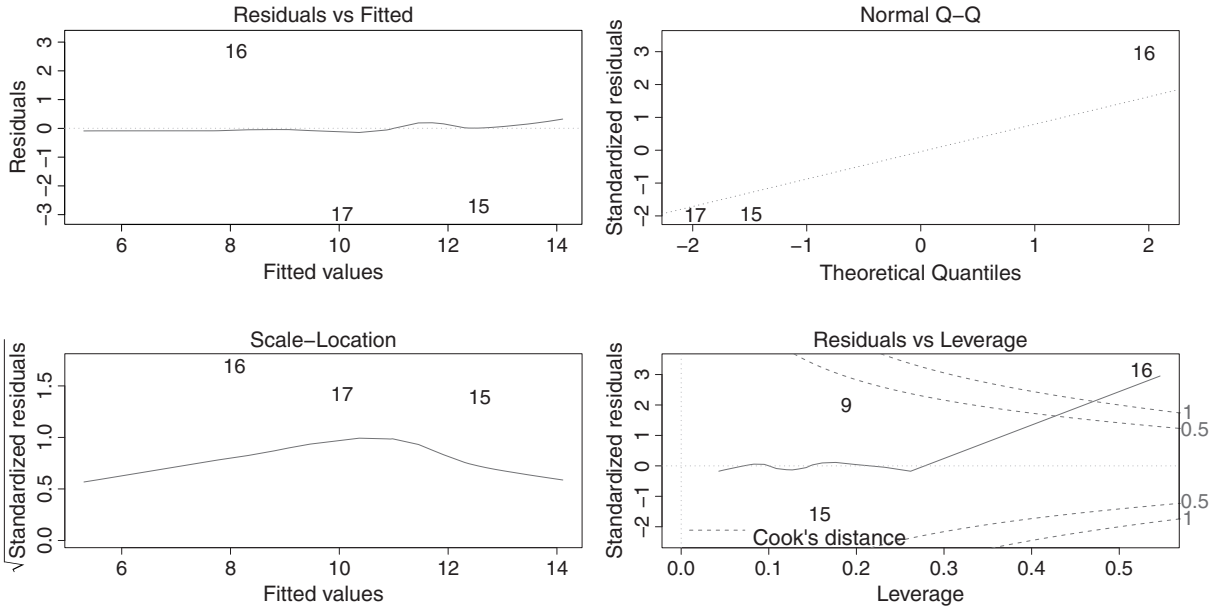
Regression diagnostics are a set of mostly graphical methods which are used to check empirically the reasonableness of the basic assumptions made in the model. These informal methods are an important part of regression modelling: many formal conclusions and inferences (including confidence intervals, statistical tests, prediction etc.) derived from a fitted model only make sense if the assumptions of the model hold. If the regression assumptions are violated, any application of results obtained from the model can be very misleading.

For a data set of n observations of a response variable y and k explanatory variables x_j $(j = 1, \ldots, k)$, the standard linear regression model (see ►Linear Regression Models) for the relationship between the response and the explanatory variables can be written in matrix notation as

$$y = X\beta + \epsilon, \tag{1}$$

where $y = (y_i)$ is an n-vector of observations, $X = (x_{ij})$ is an $n \times k$ matrix of independent variables, $\beta = (\beta_j)$ is a k-vector of unknown parameters and $\epsilon = (\epsilon_i)$ is an n-vector of unobserved random variables, often called errors. The basic assumptions of the model are that the relationship between y and X is linear, the ϵ_i are independent, have constant variance and are normally distributed.

The basic quantities on which diagnostics are based are the residuals and fitted values. For any estimator $\hat{\beta}$ of β, the fitted values are $\hat{y} = X\hat{\beta}$ and the residuals are $\hat{\epsilon} = y - \hat{y} = y - X\hat{\beta}$. The residuals provide information about the errors in the model so are fundamental in diagnostics. Various forms of standardized residuals can also be calculated. If X is of full column rank so $\hat{\beta} = (X'X)^{-1}X'y$ is the least squares estimator of β, the fitted values can be written as $\hat{y} = Hy$, where $H = X(X'X)^{-1}X' = (h_{ij})$ is the hat matrix and the ith diagonal element h_{ii} is called the leverage of the ith observation. The residuals can be standardized as $\hat{\epsilon}_i/s$, where $s^2 = (n-k)^{-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2$, as $\hat{\epsilon}_i/s(1-h_{ii})^{1/2}$ (internally Studentized) or as $\hat{\epsilon}_i/s_{(i)}(1-h_{ii})^{1/2}$ (externally Studentized), where $s^2_{(i)} =$



Regression Diagnostics. Fig. 1 Diagnostic plots based on the least squares fit of a linear regression model to the salinity data of Ruppert and Carroll (1980)

$(n - k - 1)^{-1} \sum_{j \neq i}^{n} \hat{\epsilon}_{(j)}^2$ and $\hat{\epsilon}_{(j)}$ is the residual for the jth observation calculated from the $n - 1$ observations after excluding the jth observation. Other useful quantities include ▶Cook's Distance which is a measure of influence involving the square of the Studentized residual and the potential function $h_{ii}/(1 - h_{ii})$.

The most widely used diagnostic plots are residual plots which plot the residuals against the fitted values (checking for linearity, constant variance and ▶outliers), spread plots which plot the square root of the Studentized residuals against the fitted values (checking for constant variance, outliers), QQ-plots which plot the ordered residuals against their expected values under normality (normality, outliers), and leverage plots which plot Studentized residuals against the leverage (checking for ▶influential observations). These four plots are illustrated in Fig. 1 for the salinity data (Ruppert and Carroll 1980) which have 28 observations and 3 explanatory variables. The plots are supplemented by lines and curves which aid in their interpretation. The most interesting features are the departure from normality in the upper tail (observation 16) shown in the QQ-plot and the confirmation that this observation is influential in the leverage plot. In general, outliers may be difficult to find without the use of robust method: using robust methods, Ruppert and Carroll also identified observations 15 and 17 as outliers in these data. Other useful plots include added-variable plots (examining the relationship between y and x_j after adjusting for the other explanatory variables) and partial-residual plots (checking for linearity). In addition, there are a number of specialized plots which can be used to check for dependence: these include various time series and spatial plots, correlograms (ACF, PACF), variograms and spectrum plots. These methods are extensively documented in the statistical literature.

See for example the list of references at the end of this entry.

Graphical methods are preferred in diagnostics because they are more informative than numerical ones and often suggest ways in which deficiencies in a model can be rectified. A good illustration is Anscombe's (1973) set of 4 different datasets with the same summary statistics but four distinct regression relationships between the response and explanatory variables.

Diagnostic methods are important in all statistical modelling including generalised linear models (de Jong and Heller 2008), time series analysis (Li 2003) etc.

About the Authors

Shuangzhe Liu is Associate Professor in the Discipline of Mathematics and Statistics in the Faculty of Information Sciences and Engineering at the University of Canberra.He

holds a PhD in Econometrics from the Tinbergen Institute, University of Amsterdam. He is a member, Statistical Society of Australia (2010–), and Australian Mathematical Sciences Institute (2007–). He is a Contributing Editor, *Current Index to Statistics* (2000–), and an Associate Editor, *Chilean Journal of Statistics* (2009–).

Alan Welsh is the E.J. Hannan Professor of Statistics and the Head of the Centre for Mathematics and its Applications at the Australian National University. He is a fellow of the Australian Academy of Science, the Institute for Mathematical Statistics and the American Statistical Association. He is currently Applications Editor of the *Australian and New Zealand Journal of Statistics* and an Associate Editor of the *Journal of the American Statistical Association*. He has published over 95 papers and a book on statistical inference.

Cross References

►Cook's Distance
►Influential Observations
►Linear Regression Models
►Outliers
►Residuals
►Robust Regression Estimation in Generalized Linear Models
►Simple Linear Regression

References and Further Reading

Anscombe FJ (1973) Graphs in statistical analysis. Am Stat 27: 17–21
Atkinson AC, Riani M (2000) Robust diagnostic regression analysis. Springer, New York
Belsley DA, Kuh E, Welsch RE (2004) Regression diagnostics: identifying influential data and sources of collinearity, 2nd edn. Wiley, New York
Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall/CRC, New York
Cook RD, Weisberg S (1999) Applied regression including computing and graphics. Wiley, New York
de Jong P, Heller GZ (2008) Generalized linear models for insurance data. Cambridge University Press, Cambridge
Fox J (1991) Regression diagnostics: an introduction. Sage, New York
Fox J (2008) Applied regression analysis and generalized linear models, 2nd edn. Sage, New York
Li WK (2003) Diagnostic checks in time series. Chapman & Hall/CRC, New York
Ruppert D, Carroll RJ (1980) Trimmed least squares in the linear model. J Am Stat Assoc 75:828–838
Wheeler D (2009) Spatially varying coefficient regression models: diagnostic and remedial method for collinearity. Vdm Verlag Dr. Müller, p 132. ISBN 3-63911437-X

Regression Models with Increasing Numbers of Unknown Parameters

Asaf Hajiyev
Professor, Chair
Baku State University, Baku, Azerbaijan

Introduction

Consider the regression model

$$y_i = f(x_i, \theta) + \varepsilon_i, \quad i = 1, 2, \ldots, N \tag{1}$$

where x_i is the point of observation, y_i an observable value, ε_i a random error at the point x_i, and $\theta = (\theta_1, \theta_2, \ldots, \theta_m)^T$ is the vector of unknown parameters. Let us suppose that the number of unknown parameters m depends on the number of observations N and m may increase, when N becomes larger. Such regressions are called models with increasing number of unknown parameters. The variances of observation error are unknown and may be different. At each point x_i there is only one observable value, y_i, that does not allow estimation of the variance.

Regression models with an increasing number of unknown parameters and with unknown and different variances of observation error are of interest in important applications. This is because, with an increased number of unknown parameters, the unknown function can be approximated more accurately in experiments. Moreover, in some applications, repeated tests at a single point are costly (financially and technically), which hampers the estimation of the unknown error variance, which is different at different observation points.

Regression models have been widely addressed in numerous publications (Demidenko 1989; Huet et al. 1996; Sen and Srivastava 1997), but models with an increasing number of unknown parameters have received little attention, which motivates our interest in this subject. The main aims of our investigations are

- Direct estimation (without estimation of a variance) of the elements of the covariance matrix of the vector $\sqrt{N}(\theta^* - \theta)$, where θ^* is the least square estimator (l.s.e.).
- Construction of a confidence band for the unknown function $f(x, \theta)$.

Linear Regression Models

Let us assume that

$$f(x, \theta) = \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \ldots + \theta_{m(N)} \phi_{m(N)}(x), \tag{2}$$

where $\phi_1(x), \phi_2(x), \ldots, \phi_{m(N)}(x)$ is a system of linearly independent and bounded functions. Expression (1) can be rewritten in a vector form as

$$Y = X\theta + \varepsilon, \tag{3}$$

where Y is the vector of observable values, X the design matrix, defined as $X = //x \quad ij//, x_{ij} = \phi_i(x_j)$, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$; with $\theta = (\theta_1, \theta_2, \ldots, \theta_{m(N)})^T$ being the vector of unknown parameters and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^T$ denoting the error-vector. The number of unknown parameters depends on the number of observations and moreover

$$m(N)/N \to 0, \; as \; n \to \infty. \tag{4}$$

The sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ is assumed to have uniformly bounded and independent random variables with $E\varepsilon_i = 0, E\varepsilon_i^2 = {\sigma_i}^2$ being unknown, different, and

$$0 < (\sigma_*)^2 \le {\sigma_i}^2 \le (\sigma^*)^2 < \infty.$$

Let $\theta^* = (X^TX)^{-1}X^TY^Y$ be the l.s.e and $trA = \sum_{i=1}^{n} a_{ii}$ be the trace of the matrix A with elements a_{ij}.

Definition 1 The vector $\theta = (\theta_1, \theta_2, \ldots, \theta_{m(N)})^T$ with random elements and increasing dimension converges to zero in probability $\theta \xrightarrow{P} 0$, if $\sum_{i=1}^{m(N)} \theta_i^2 \xrightarrow{P} 0$ as $N \to \infty$.

Let $0 < \lambda_1(N) \le \lambda_2(N) \le \ldots \le \lambda_m(N)$ be eigenvalues of the matrix $(X^TX)/N$.

Theorem 1 Let the conditions (2)–(4) be true. Then $(\theta^* - \theta) \xrightarrow{P} 0$, if and only if $(1/N)tr(X^TX/N) \xrightarrow{P} 0$ as $N \to \infty$.

Definition 2 The vector $\theta^P = ({\theta_1}^P, {\theta_2}^P, \ldots, {\theta_{m(N)}}^P, 0, \ldots, 0)^T$ is called m-finite and p-consistent estimator of the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_N)^T$, if

$$\forall \delta > 0 \; P\left\{\sum_{i=1}^{m(N)} \left({\theta_i}^P - \theta_i\right)^2 < \delta\right\} \ge p \text{ holds true, as } N \to \infty.$$

Example 1 Consider $f(x, \theta) = \sum_{i=1}^{\infty} \theta_i \phi_i < \infty, |\phi_i(x)| \le 1$, $\sum_{i=1}^{\infty} \theta_i^2 < \infty$. The problem is to find such $m(p, N, \delta)$ ($\forall \delta > 0$ and given $0 < p < 1, \quad N > 0$), for which $P\left\{\sum_{i=1}^{m(N)} \left(\theta_i^* - \theta_i\right)^2 < \delta\right\} \ge p$ holds true, where θ_i^* is the l.s.e. on N observations. For simplicity, we assume $E\varepsilon_i = 0, E\varepsilon_i^2 \le 1$.

Consider $y_i = \sum_{i=1}^{m(N)} \theta_i \phi_i(x) + \delta_i$, where

$$\delta_i = \sum_{j=m(N)+1}^{\infty} \theta_i \phi_i(x) + \varepsilon_i, \; \delta_i$$
$$= \sum_{j=m(N)+1}^{\infty} \theta_i \phi_i(x) \to 0 \text{ as } N \to \infty.$$

Assuming that for large values of $N \; P\left\{\sum_{i=1}^{\infty} \left({\theta_i}^* - \theta_i\right)^2 < \delta\right\} \approx P\left\{\sum_{i=1}^{m(N)} \left({\theta_i}^* - \theta_i\right)^2 < \delta\right\}$. If the conditions of the Theorem 1 hold true, then we obtain

$$P\left\{\sum_{i=1}^{m(N)} \left({\theta_i}^* - \theta_i\right)^2 < \delta\right\} \ge 1 - (m+1)/(N\lambda_1(N)\delta)$$

from Chebyshev inequality. Taking $p = 1 - (m+1)/(N\lambda_1(N)\delta)$, we get $m = (1-p)(N\lambda_1(N)\delta) - 1$. Now in the capacity of a consistent estimator of the vector $\theta = (\theta_1, \theta_2, \ldots, \theta_N)^T$, we can take the vector $\theta^P = ({\theta_1}^P, {\theta_2}^P, \ldots, {\theta_{m(N)}}^P, 0, \ldots, 0)^T$, where m was found, above. According to the Theorem 1 the vector θ^* is a consistent estimator of θ.

Estimation of Covariance Matrix

Denote

$$D_N = E(\theta^* - \theta)(\theta^* - \theta)^T$$
$$= (1/N)(X^TX/N)^{-1}[X^T(E\varepsilon\varepsilon^T)X/N](X^TX/N)^{-1}$$
$$= (1/N)(X^TX/N)^{-1}[X^TI(\sigma^2)X/N](X^TX/N)$$

where $I(\sigma^2) = //z_{ij}//$ is an unknown matrix, $//z_{ij}// = \sigma_i\sigma_j\delta_{ij}$, $\delta_{ij}(i, j = 1, 2, \ldots, N)$ is Kroneker symbol

$$C_N = X^TI(\sigma^2)X/N, \; C_N = //c_{kl}//, \; k, l = 1, 2, \ldots, m;$$
$$y^* = X\theta^*,$$

$$I_{kl}(x) = //{a_{ij}}^{kl}//, \; i, j = 1, 2, \ldots, m;$$

$$//{a_{ij}}^{kl}// = \phi_k(x_j)\phi_l(x_j)\delta_{ij}, \; {c_{kl}}^*$$
$$= (1/N)(y^* - y)I_{kl}(x)(y^* - y),$$
$$C_N{}^* = //{c_{kl}}^*//, \; k/l = 1, 2, ..., m.$$

Theorem 2 Let $E{\varepsilon_i}^4 < \infty$ and $\left(m\sqrt{m}\right)/(N\lambda_1(N)) \to 0$, as $N \to \infty$, then

$$\left({c_{ij}}^* - c_{ij}\right) \xrightarrow{P} 0, \quad E\left({c_{ij}}^* - c_{ij}\right) \to 0 \; as \; N \to \infty.$$

Remark 1 In Theorem 2 we do not need the existence of the limit c_{kl}^* and c_{kl}. This is because the difference between them converges to zero, in probability.

Theorem 3 If $\left(m\sqrt{m}\right)/(N\lambda_1(N))$ will be bounded, then $\sqrt{N}(\theta^* - \theta) \Rightarrow N(0, D_N)$ as $N \to \infty$, where $\Rightarrow N(0, D_N)$ means a convergence in probability to the normal distribution with the covariance matrix D_N.

Different approaches for estimating the elements of covariance matrix were suggested in Belyaev and Hajiyev (1979), Hajiyev (2004), and Wu (1986).

Nonlinear Regression Models

Let us assume that $f(x_i, \theta)$ in (1) is a nonlinear function and

$$f(x_i, \theta),\ \partial f(x, \theta)/\partial\theta,\ \partial^2 f(x, \theta)/\partial\theta_i\partial\theta_j, (i, j = 1, 2, \ldots, m)$$

are bounded and continuous functions of (x, θ), $\theta \in \Theta$ is a compact set. Denote

$$f_{ij}(\theta) = \partial f(x_j, \theta)/\partial\theta_i;$$

$F_N(\theta)$ is the matrix with elements $f_{ij}(\theta)$, $0 < \mu_1{}^N(\theta) \le \ldots \le \mu_m{}^N(\theta)$ eigenvalues of the matrix $\left[F_N{}^T(\theta)F_N(\theta)/N\right]$ and $B(r)$ be the sphere of the radius $r > 0$ centered at the point θ^*. A least squares estimator of θ is constructed by the iterative process

$$\theta_N(s+1) = \theta_N(s) + \left[F_N{}^T(\theta_N(s))F_N(\theta_N(s))\right]^{-1} F_N{}^T(\theta_N(s))(y - f(x, \theta_N(s))). \quad (5)$$

The question arises as to whether the iterative process (5) converges or not. Relation (5) can be represented as

$$\theta_N(s+1) = u(\theta_N(s)) = \theta_N(s) + A_N(\theta_N(s))\delta_N(\theta_N(s)),$$

where

$$A_N(\theta_N(s)) = \left[F_N{}^T(\theta_N(s))F_N(\theta_N(s))/N\right]^{-1} F_N{}^T(\theta_N(s))$$

$$\delta_N(\theta_N(s)) = y - f(x, \theta_N(s)),\ \delta_N{}^*(\theta_N(s)) = y - f(x, \theta^*).$$

Define

$$\zeta_{N,r}{}^p(\theta) = m(\partial A_N(\theta)/\partial\theta_p)\varepsilon, p = 1, 2, .., m; \theta \in B(r),$$
$$L_p = \partial u_N(\theta)/\partial\theta_p,$$
$$\tau_N(r) = \max_{p=1,2,\ldots,m}\sup_{\theta\in\Theta}//L_p//.$$

Below, the convergence of random variables is understood as convergence in probability.

Theorem 4 If there exists such N that $m(N)^5/\left[N\left(\lambda_1{}^N(\theta)\right)^4\right] \to 0, r \to 0$, then

$$m(N)\tau_N(r) \to 0 \text{ and for any } p,\ \zeta_{N,r}{}^p(\theta) \to 0, r \to 0.$$

Introduce $\rho_N(\theta) = u_N(\theta) - \theta, \rho^* = \rho(\theta^*)$.

Theorem 5 Let $\theta(0) \in B(r)$ and $\tau_N(r) + (//\rho^*//)/r < 1$. Then under the conditions of Theorem 4, there exists a random variable θ_N such that

$$\sqrt{N}(\theta_N - \theta^*) \Rightarrow N[0, \sum(\theta^*)] \text{ as } N \to \infty,$$

where

$$\sum(\theta^*) = \left[F_N{}^T(\theta^*)F_N(\theta^*)/N\right]^{-1} \left[F_N{}^T(\theta^*)I(\sigma^2)F_N(\theta^*)/N\right]\left[F_N{}^T(\theta^*)F_N(\theta^*)/N\right]^{-1},$$

that is, θ_N is a $\sqrt{N}$ consistent estimator and θ_N can be used as l.s.e. on N observations. Using the approach suggested in Hajiyev and Hajiyev (2009), (similarly as for linear models) the elements of a covariance matrix can be estimated.

The Construction of Asymptotic Confidence Bands

Consider the quadratic form

$$(\theta^* - \theta)^T(D_N)^{-1}(\theta^* - \theta) \le \chi_\gamma{}^2(m)/N. \quad (6)$$

According to Theorem 5, the left side of (6) has asymptotically chi-square distribution random with degrees of freedom m. In (6) instead of D_N^{-1} (according to the Theorem 2) can be used estimates (Hajiyev and Hajiyev 2009) the matrix D_N^{-1} elements. For the construction of a confidence band for $f(x, \theta)$, it is necessary to find $\inf f(x, \theta)$ and $\sup f(x, \theta)$, $\theta \in \varepsilon_\gamma(\theta)$, which are lower and upper boundaries of a confidence band and

$$\varepsilon_\gamma(\theta) = \left[\theta : (\theta^* - \theta)^T\left(D_N{}^{-1}\right)(\theta^* - \theta) \le \chi_\gamma^2(m)/N\right]$$

is the confidence ellipsoid, $\chi_\gamma{}^2(m)$ is the $\gamma > 0$ level quantile of the ▸chi-square distribution with m degrees of freedom.

About the Author

Dr. Asaf Hajiyev is a Professor and Chair, Department of Queuing Systems, Institute of Cybernetics, the Azerbaijan National Academy of Sciences (ANAS) and Head, Department of Probability and Mathematical Statistics, Baku State University. In 1980, he was awarded the Azerbaijan Lenin Komsomol Prize in Science and Technology. He is a Member of Bernoulli Society for Mathematical Statistics and Probability (1986), Elected member of the International Statistical Institute (2000), Elected correspondent-member Azerbaian National Academy of Science (2001), Elected member of the Third World Academy of Sciences (TWAS), Italy (2004). In 2008, he was elected a Member of the Mongolian National Academy of Sciences. Professor Hajiyev is Head of the Coordinating Council of the Azerbaian National Academy of Science in Mathematics. He

was a Member of the Azerbaijan Parliament: first (1995–2000), second (2000–2005), and third (2005–2010) convocations. During 2004–2006, he was elected Vice President of the Parliamentary Assembly of the Black Sea Economy Cooperation. He has authored more than 100 scientific papers and two books, including *Encyclopedia in Theory of Probability and Mathematical Statistics*.

Cross References

►Convergence of Random Variables
►Eigenvalue, Eigenvector and Eigenspace
►Least Squares
►Linear Regression Models
►Nonlinear Regression

References and Further Reading

Belyaev YuK, Hajiyev AH (1979) Sov J Comput Syst Sci 4:79–83
Demidenko EZ (1989) Optimization and regression. Nauka, Moscow. In Russian
Hajiyev AH (2004) Linear regression models with increasing numbers of unknown parameters. Doklady Mathematics 70(3): 887–891
Hajiyev AH, Hajiyev VG (2009) Nonlinear regression models with increasing numbers of unknown parameters. Doklady, Math 426(2):166–169
Huet S, Bouvier A, Gruey MA, Jolivet E (1996) Statistical tools for nonlinear regression. Springer, New-York
Sen A, Srivastava M (1997) Regression analysis: theory, methods and applications. Springer, Berlin
Wu CFJ (1986) Jackknife, bootstrap, and other resampling methods in regression analysis. Ann Stat 14:1261–1350

Regression Models with Symmetrical Errors

Francisco José A. Cysneiros
Associate Professor
CCEN-UFPE - Cidade Universitária - Recife, Recife, Brazil

The normality assumption is a very attractive option for the errors of regression models with continuous response variables. However, when it is not satisfied, some transformation can be adopted for the response variable to obtain, at least, the symmetry property. It is known that the estimates of the coefficients in normal regression models are sensitive to extreme observations. Alternatives to the assumption of normal errors have been proposed in the literature. One of those alternatives is to consider that the errors have distributions with heavier tails than the normal distribution, in order to reduce the influence of outlier observations. In this context, Lange et al. (1989) proposed the Student t model with unknown ν degrees of freedom. In the last decade, several results appeared as alternatives to modeling distributions other than the normal errors as, for instance, the symmetrical (or elliptical) distributions. Some of these results can be found in Fang et al. (1990), and Fang and Anderson (1990).

Symmetrical Nonlinear Models

Consider the symmetrical nonlinear model

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where $y_1, \ldots, y_n$ are the observed responses, $\mu_i = \mu_i(\boldsymbol{\beta}; \mathbf{x})$ is an injective and at least twice differentiable function with respect to $\boldsymbol{\beta}$. In addition, we suppose that the derivative matrix $\mathbf{D}_\beta = \partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$ has rank $p(p < n)$ for all $\boldsymbol{\beta} \in \Omega_\beta \subset I\!R^p$, where Ω_β is a compact set with interior points, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the parameter vector of interest, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is a vector of explanatory variable values and $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with the symmetrical density function $f_{\epsilon_i}(\epsilon) = g(\epsilon^2/\phi)/\sqrt{\phi}$, $y \in I\!R$, where $g : I\!R \to [0, \infty)$ is such that $\int_0^\infty g(u)du < \infty$. The function $g(\cdot)$ is typically known as the density generator. We will denote $\epsilon_i \sim S(0, \phi, g)$. The symmetrical class includes all symmetrical continuous distributions with heavier and lighter tails than the normal ones. When they exist, $E(Y_i) = \mu_i$ and $Var(Y_i) = \xi\phi$, where $\xi > 0$ is a constant that may be obtained from the expected value of the radial variable or from the derivative of the characteristic function (see, for instance, Fang et al. 1990). The log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$ is given by $L(\boldsymbol{\theta}) = -n/2 \log\phi + \sum_{i=1}^n log\{g(u_i)\}$, where $u_i = \phi^{-1}\{y_i - \mu_i\}^2$. The score functions for $\boldsymbol{\beta}$ and ϕ take, respectively, the forms

$$\mathbf{U}_\beta(\boldsymbol{\theta}) = \frac{1}{\phi}\mathbf{D}_\beta^T\mathbf{V}(\mathbf{y} - \boldsymbol{\mu}) \ \ and$$

$$\mathbf{U}_\phi(\boldsymbol{\theta}) = 2\phi^{-1}\{Q_V(\boldsymbol{\beta}, \phi)/\phi - n\},$$

where $\mathbf{V} = diag\{v_1, \ldots, v_n\}$ with $v_i = -2W_g(u_i)$, $W_g(u) = \frac{g'(u)}{g(u)}$, $g'(u) = \frac{dg(u)}{du}$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, $Q_V(\boldsymbol{\beta}, \phi) = \{(\mathbf{y} - \boldsymbol{\mu})^t\mathbf{V}(\mathbf{y} - \boldsymbol{\mu})\}$. The Fisher information matrix for $\boldsymbol{\theta}$ can be expressed as $\mathbf{K}_{\theta\theta} = diag\{\mathbf{K}_{\beta\beta}, K_{\phi\phi}\}$, where $\mathbf{K}_{\beta\beta} = 4d_g\phi^{-1}\mathbf{D}_\beta^T\mathbf{D}_\beta$ and $K_{\phi\phi} = n(4\phi^2)^{-1}(4f_g - 1)$ with $d_g = E\left\{W_g^2(U^2)U^2\right\}$, $f_g = E\left\{W_g^2(U^2)U^4\right\}$ and $U \sim S(0, 1, g)$. Thus, $\boldsymbol{\beta}$ and ϕ are orthogonal. Due to the similarity between the inference for elliptical and normal models, it is reasonable to expect that for large n and under suitable regularity conditions, the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are approximately normal of means $\boldsymbol{\beta}$ and ϕ and variance–covariance matrices $\mathbf{K}_{\beta\beta}^{-1}$ and $\mathbf{K}_{\phi\phi}^{-1}$, respectively. General expressions for $W_g(u)$, $W_g'(u)$, d_g, f_g, and ξ may be found, for instance, in Cysneiros and Paula (2005).

Parameter Estimation

Some iterative procedures such as Newton–Raphson, BFGS, and Fisher scoring method can be used. Fisher scoring method can be easily applied to obtain $\hat{\boldsymbol{\theta}}$, where the iterative process can be interpreted as a modified least squares. The iterative process for $\hat{\boldsymbol{\theta}}$ take the form

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} &= \left\{\mathbf{D}_{\beta}^{T(m)}\mathbf{D}_{\beta}^{(m)}\right\}^{-1}\mathbf{D}_{\beta}^{T(m)}\mathbf{Z}^{(m)},\\ \phi^{(m+1)} &= \frac{1}{n}Q_V(\boldsymbol{\beta}^{(m+1)},\phi^{(m)}), \quad m=0,1,2,\ldots, \quad (2)\end{aligned}$$

where $\mathbf{Z} = \mathbf{D}_{\beta}\boldsymbol{\beta} + \frac{1}{4d_g}\mathbf{V}(\mathbf{y}-\boldsymbol{\mu})$. In linear case, we have $\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T\mathbf{V}^{(m)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{(m)}\mathbf{y}$. Starting values should be given for $\boldsymbol{\beta}$ and ϕ, for example, least square estimates. As we can see from the iterative process (2), the observations with small value for v_i are down weighted for estimating $\boldsymbol{\beta}$. In particular, for the normal model, we have $v_i = 1$, $\forall i$. For the Student t model with ν degrees of freedom, power exponential with shape parameter k, and logistic type II distributions, the values of v_i are given in the Table 1.

It may be showed for the Student t and logistic type II distributions, v_i is inversely proportional to u_i. This property also follows for the power exponential distribution when $0 < k \leq 1$. Then, robustness aspects of $\hat{\beta}$ against outlying observations appear in these three heavy-tailed error distributions. In general, when the errors of the model have distribution with heavier tails than normal, the values of the weights v_i have small values for u_i large. Thus, models where the distribution of error have heavy tails can reduce the influence of extreme observations, while in the normal nonlinear regression model the weights are equal for all observations. In consequence, estimates in symmetrical regression models are less sensitive to the extreme observations than normal regression models. Extensions in the area of heteroscedastic symmetrical regression models can be found in Cysneiros et al. (2007, 2010) and codes in S-Plus and R to fit symmetrical regression models can be obtained in the Web page www.de.ufpe.br/~cysneiros/elliptical/elliptical.html.

Regression Models with Symmetrical Errors. Table 1
Expression of v_i for some symmetrical distributions

Distribution	v_i
Normal	1
Student-t	$\frac{\nu+1}{(\nu+u_i)}$
Logistic-II	$\frac{2exp(-\sqrt{u_i})-1}{(-2\sqrt{u_i})[1+exp(-\sqrt{u_i})]}$
Power exponential	$\frac{1}{(1+k)u_i^{k/(k+1)}}$

About the Author

Francisco Cysneiros is Associate Professor and Vice-Director of graduate studies (statistics graduate program) of Department of Statistics at Federal University of Pernambuco, Brazil. He is also Vice-Head of Department of Statistics (2005–2009). Professor Cysneiros is currently a member of the Advisory Board of Biometric Brazilian Journal (2008–2010) and he is an Associate Editor of the *Chilean Journal of Statistics*. He has served as a member of the Exact Science Committee – FACEPE (Research Foundation of Pernambuco, Brazil) and he is a fellowship researcher of the CNPq/Brazil since 2006.

Cross References

►Heavy-Tailed Distributions
►Logistic Regression
►Nonlinear Models
►Nonlinear Regression
►Student's t-Distribution

References and Further Reading

Cysneiros FJA, Paula GA (2005) Restricted methods in symmetrical linear regression models. Comput Stat Data Anal 49:689–708
Cysneiros FJA, Paula GA, Galea M (2007) Heteroscedastic symmetrical linear models. Stat Probab Lett 77:1084–1090
Cysneiros FJA, Cordeiro GM, Cysneiros AHMA (2010) Corrected maximum likelihood estimators in heteroscedastic symmetric nonlinear models. J Stat Comput Sim 80:451–461
Fang KT, Anderson TW (1990) Statistical inference in elliptical contoured and related distributions. Allerton Press, New York
Fang KT, Kotz S, Ng KW (1990) Symmetric multivariate and related distributions. Chapman & Hall, London
Lange KL, Little RJ, Taylor J (1989) Robust statistical modelling using the t-distribution. J Am Stat Assoc 84:881–896

Relationship Between Statistical and Engineering Process Control

Alberto Luceño
Professor
University of Cantabria, Santander, Spain

Introduction

Many industrial processes must be adjusted from time to time to continuously maintain their outputs close to target. The reason for this is that such processes may be affected by disturbances produced, for example, by machines loosing their adjustment, components wearing out, and varying feed stock characteristics. Industrial control is a continual endeavor to keep measures of quality as close as possible to their target values for indefinite periods of time.

This may be attained using process monitoring and process adjustment tools. Monitoring implies continually checking the desired state of the process to detect and eliminate assignable causes of variation that can send the process out of control. Adjustment implies forecasting future deviations and taking corrective actions by feedback and/or feedforward. Process control can potentially benefit by using complementary tools of process monitoring and process adjustment within the same application.

Process Monitoring Techniques

Process monitoring, or process surveillance, is a part of Statistical Process Control (SPC) that is used when the process can be brought to a satisfactory state of statistical control by systematically applying standardization of criteria, materials, methods, practices and processes.

Process monitoring is usually implemented by means of ►control charts, such as the Shewhart charts, the CUmulative SUM (CUSUM) charts, or the Exponentially Weighted Moving Average (EWMA) charts, among others.

The purpose of such methods is to continually check, or supervise, the state of the process in order to detect any conceivable out of control situation as soon as possible while simultaneously minimizing the rate of false alarms (i.e., alarms that eventually turn out to have no special cause).

When an alarm is triggered, a search for the special and potentially assignable cause of variation that presumably produced the alarm should be started. This search should end with the detection of such assignable cause and its permanent removal from the system. If the search fails, so that no special cause is eventually found, the alarm should be counted as a false alarm.

Process Adjustment Techniques

Process adjustment is often considered as a part of Engineering Process Control (EPC) and is used when the process cannot be brought to a satisfactory state of statistical control, even after systematic application of standardization techniques. Much efforts have recently been dedicated, however, to bring some important features of process adjustment to the attention of the statistical community (e.g., see Box and Kramer 1992; Box and Luceño 1997a,b, 2002; Box et al. 2009; Luceño 2003, or Montgomery et al. 1994).

Process adjustment is often implemented by first using forecasting tools to estimate future deviations from target and subsequently modifying, or adjusting, an input compensatory variable so as to make those predicted deviations equal to zero (or to an appropriate small value in asymmetric situations). A process adjustment scheme may use feedback adjustments, feedforward adjustments, or a combination of both. Some types of feedback adjustment schemes are repeated adjustment schemes, constrained adjustment schemes, Proportional Integral Derivative (PID) control schemes, bounded adjustment schemes, among other.

The purpose of these methods is to indicate when and by how much the process has to be sampled and adjusted to keep it close to target. The only actions called for are to sample and to adjust the process when and as indicated by the adjustment scheme. The objective may be to minimize the output variance (or the mean squared error at the output) without any additional constraints, or to minimize the output variance constrained by a bound on the input variance, or by a bound on the frequency of adjustment, or on the frequency of sampling, or on the amount of each adjustment, among other possibilities.

Conclusion

One can tentatively conclude that declarations of alarms and searches for special and potentially assignable causes of variation are not called for in the context of process adjustment techniques, but in the context of process monitoring techniques. By the same token, process adjustments are not called for in the context of process monitoring techniques, but much more appropriately in the context of process adjustment techniques.

Nevertheless, the appropriate combination of process monitoring and process adjustment tools, and their complementary use in SPC, is the subject of controversy within the statistical community. Further information can be found in the bibliography that follows, as well as in many documents produced by the International Organization for Standardization (ISO) and related organizations (e.g., ANSI, DIN, BSI, CEN).

About the Author

For biography *see* the entry ►Control Charts.

Cross References

►Control Charts
►Industrial Statistics
►Statistical Quality Control
►Statistical Quality Control: Recent Advances

References and Further Reading

Box GEP, Kramer T (1992) Statistical process monitoring and feedback adjustment. A discussion. Technometrics 34:251–267
Box GEP, Luceño A (1997a) Statistical control by monitoring and feedback adjustment. Wiley, New York
Box GEP, Luceño A (1997b) Discrete proportional-integral adjustment and statistical process control. J Qual Technol 29:248–260

Box GEP, Luceño A (2002) Feedforward as a supplement to feedback adjustment in allowing for feedstock changes. J Appl Stat 29:1241–1254
Box GEP, Luceño A, Paniagua-Quiñones MA (2009) Statistical control by monitoring and adjustment, 2nd edn. Wiley, New York
Luceño A (2003) Dead-band adjustment schemes for on-line feedback quality control. In: Khattree R, Rao CR (eds) Handbook of statistics: statistics in industry, vol 22. Elsevier, Amsterdam, pp 695–727
Montgomery DC (2005) Introduction to statistical quality control, 5th edn. Wiley, New York
Montgomery DC, Keats BJ, Runger GC, Messina WS (1994) Integrating statistical process control and engineering process control. J Qual Technol 26:79–87
NIST/SEMATECH (2009) e-Handbook of statistical methods. http://www.itl.nist.gov/div898/handbook/
Ruggery F, Kenetts RS, Faltin FW (eds) (2007) Encyclopedia of statistics in quality and reliability. Wiley, New York

Relationships Among Univariate Statistical Distributions

Lawrence M. Leemis
Professor
The College of William & Mary, Williamsburg, VA, USA

Certain statistical distributions occur so often in applications that they are named. Examples include the binomial, exponential, normal, and uniform distributions. These distributions typically have *parameters* that allow for a certain degree of flexibility for modeling. Two important applications of these common statistical distributions are: (a) to provide a probability model the outcome of a random experiment, and (b) to provide a reasonable approximation to a data set.

Statistical distributions are traditionally introduced in separate sections in introductory probability texts, which obscures the fact that there are relationships between these distributions. The purpose of this section is to overview the various types of relationships between these common univariate distributions.

Common distributions and their relationships are presented in the encyclopedic work of Johnson et al. (1994, 1995) and Johnson et al. (2005). More concise treatments are given in Balakrishnan and Nevzorov (2003), Evans et al. (2000), Patel et al. (1976), Patil et al. (1985a, b), and Shapiro and Gross (1981). Figures that highlight the relationships between distributions are given in Casella and Berger (2002), Leemis and McQueston (2008), Marshall and Olkin (1985), Morris and Lock (2009), Nakagawa and Yoda (1977), Song (2005), and Taha (1982).

Since there are well over 100 named distributions used by probabilists and statisticians, the next three sections simply classify and illustrate some of the relationships.

Special Cases

The first type of relationship between statistical distributions is known as a *special case*, which occurs when one distribution collapses to a second distribution for certain settings of its parameters. Two well-known examples are:

- A ▶gamma distribution collapses to the exponential distribution when its shape parameter equals 1.
- A normal distribution with mean μ and variance σ^2 collapses to a standard normal distribution when $\mu = 0$ and $\sigma = 1$.

There are also certain special cases in which two statistical distributions overlap for a single setting of their parameters. Examples include (a) the exponential distribution with a mean of two and the ▶chi-square distribution with two degrees of freedom, (b) the chi-square distribution with an even number of degrees of freedom and the Erlang distribution with scale parameter two, and (c) the Kolmogorov–Smirnov distribution (all parameters known case) for a sample of size $n = 1$ and the $U(1/2,1)$ distribution, where U denotes the uniform distribution (see ▶Uniform Distribution in Statistics).

Transformations

The second type of relationship between statistical distributions is known as a *transformation*. The term "transformation" is used rather loosely here, to include the distribution of an order statistic, truncating a random variable, or taking a mixture of random variables. Some well-known examples include:

- The random variable $(X - \mu)/\sigma \sim N(0,1)$ when $X \sim N(\mu, \sigma^2)$, where N denotes the normal distribution.
- An Erlang random variable is the sum of mutually independent and identically distributed exponential random variables.
- The natural logarithm of a log normal random variable has the normal distribution.
- A hyperexponential random variable is the mixture of mutually independent exponential random variables.
- An order statistic taken from a sample of mutually independent $U(0,1)$ random variables has the ▶beta distribution.

- A geometric random variable is the floor of an exponential random variable.
- If X has the F distribution with parameters n_1 and n_2, then $\left(1 + (n_1/n_2)X\right)^{-1}$ has the beta distribution.
- If $X \sim U(0,1)$ then the floor of 10^X has the Benford distribution (Benford 1938).

It is also the case that two random variables from different statistical families can be combined via a transformation to form another common distribution, for example,

$$\frac{Z}{\sqrt{Y/n}} \sim t(n)$$

where $t(n)$ is the t distribution with n degrees of freedom, Z is a standard normal random variable, and Y is a chi-square random variable with n degrees of freedom that is independent of Z.

Limiting Relationships

The third type of relationship between statistical distributions is known as a *limiting* or *asymptotic* relationship, which is typically formulated in the limit as one or more parameters approach the boundary of the parameter space. Three well-known examples are:

- A standard normal distribution is the limit of a t distribution as its degrees of freedom parameter approaches infinity.
- If $X_1, X_2, \ldots, X_n$ are mutually independent $U(0,1)$ random variables, then

$$n(1 - \max\{X_1, X_2, \ldots, X_n\})$$

  approaches an exponential random variable in the limit as $n \to \infty$.
- The gamma distribution approaches the normal distribution as its shape parameter approaches infinity.

Bayesian Relationships

The fourth type of relationship between statistical distributions is known as a *Bayesian* or *stochastic parameters* relationship, in which one or more of the parameters of a distribution are considered to be random variables rather than fixed constants. Two well-known examples are:

- If a random variable has a ▶binomial distribution with fixed parameter n and random parameter p which has the beta distribution, then the resulting random variable has the beta–binomial distribution.
- If a random variable has a negative binomial distribution with fixed parameter n and random parameter p which has the beta distribution, then the resulting random variable has the beta-negative binomial distribution.

Internal Properties

The fifth and last type of relationship between statistical distributions is actually a relationship between a statistical distribution and itself. There are occasions when a particular operation on one or more random variables from a certain statistical family result in a new random variable that remains in that family. These are best thought of as properties of a statistical distribution rather than relationships between statistical distributions. Some well-known examples include:

- The *linear combination* property indicates that linear combinations of mutually independent random variables having this particular distribution come from the same distribution family. For example, if $X_i \sim N\left(\mu_i, \sigma_i^2\right)$ for $i = 1, 2, \ldots, n$; $a_1, a_2, \ldots, a_n$ are real constants, and $X_1, X_2, \ldots, X_n$ are mutually independent, then

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

- The *convolution* property indicates that sums of mutually independent random variables having this particular distribution come from the same distribution family. For example, if $X_i \sim \chi^2(n_i)$ for $i = 1, 2, \ldots, n$, and $X_1, X_2, \ldots, X_n$ are mutually independent, then

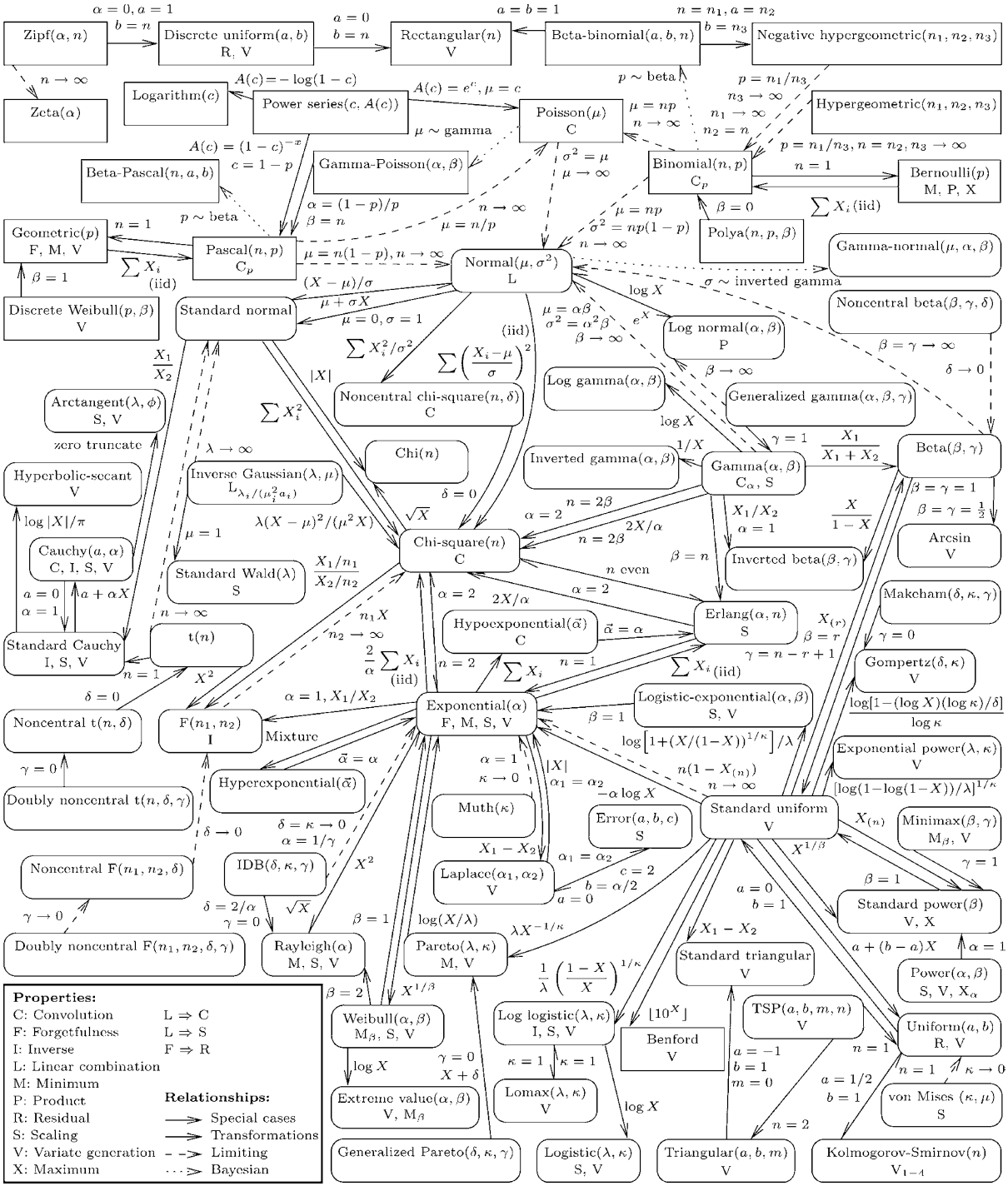
$$\sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} n_i\right),$$

  where χ^2 denotes the chi-square distribution. The convolution property is a special case of the linear combination property.
- The *scaling* property implies that any positive real constant times a random variable having this distribution comes from the same distribution family. For example, if $X \sim Weibull(\alpha, \beta)$ and k is a positive, real constant, then

$$kX \sim Weibull(\alpha k^{\beta}, \beta).$$

- The *product* property indicates that products of mutually independent random variables having this particular distribution come from the same distribution family. For example, if $X_i \sim \log\ normal\left(\mu_i, \sigma_i^2\right)$ for $i = 1, 2, \ldots, n$, and $X_1, X_2, \ldots, X_n$ are mutually independent, then

$$\prod_{i=1}^{n} X_i \sim log\ normal\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$



Relationships Among Univariate Statistical Distributions. Fig. 1 Univariate distribution relationships

- The *inverse* property indicates that the reciprocal of a random variable of this type comes from the same distribution family. For example, if $X \sim F(n_1, n_2)$ then

$$\frac{1}{X} \sim F(n_2, n_1),$$

where F denotes the F distribution.

- The *minimum* property indicates that the smallest of mutually independent and identically distributed random variables from a distribution comes from the same distribution family. For example, if $X_i \sim exponential(\lambda_i)$ for $i = 1, 2, \ldots, n$, and $X_1, X_2, \ldots, X_n$ are mutually independent, then

$$\min\{X_1, X_2, \ldots, X_n\} \sim exponential\left(\sum_{i=1}^{n} \lambda_i\right),$$

  where the exponential parameter is a rate.
- The *residual* property indicates that the conditional distribution of a random variable left-truncated at a value in its support belongs to the same distribution family as the unconditional distribution. For example, if $X \sim U(a, b)$, and k is a real constant satisfying $a < k < b$, then the conditional distribution of X given $X > k$ belongs to the uniform family.

Many of the relationships described here are contained in Fig. 1 from Leemis and McQueston (2008), which is reprinted with permission from *The American Statistician*.

About the Author

Dr. Lawrence Leemis is a Professor and former Department Chair, Department of Mathematics, The College of William & Mary in Virginia, U.S.A. He has authored or co-authored more than 100 articles, book chapters, and reviews. He has published three books: *Reliability: Probabilistic Models and Statistical Methods* (Prentice-Hall, 1995), *Discrete-Event Simulation: A First Course*, with Steve Park, (Prentice-Hall, 2006), and *Computational Probability: Algorithms and Applications in the Mathematical Sciences*, with John Drew, Diane Evans, and Andy Glen (Springer, 2008). He has won more than ten research and teaching awards, including the INFORMS Computing Society Prize in 2006. He is currently an associate editor for *Naval Research Logistics*, and has previously been an associate Editor for *IEEE Transactions on Reliability* and a Book Reviews Editor for the *Journal of Quality Technology*.

Cross References

►Beta Distribution
►Binomial Distribution
►Chi-Square Distribution
►F Distribution
►Gamma Distribution
►Geometric and Negative Binomial Distributions
►Normal Distribution, Univariate
►Statistical Distributions: An Overview
►Testing Exponentiality of Distribution
►Uniform Distribution in Statistics
►Univariate Discrete Distributions: An Overview

References and Further Reading

Balakrishnan N, Nevzorov VB (2003) A primer on statistical distributions. Wiley, Hoboken
Benford F (1938) The law of anomalous numbers. Proc Am Phil Soc 78:551–572
Casella G, Berger R (2002) Statistical inference, 2nd edn. Duxbury, Belmont
Evans M, Hastings N, Peacock B (2000) Statistical distributions, 3rd edn. Wiley, New York
Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, New York
Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol I, 2nd edn. Wiley, New York
Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distributions, vol II, 2nd edn. Wiley, New York
Leemis LM, McQueston JT (2008) Univariate distribution relationships. Am Stat 62(1):43–53
Marshall AW, Olkin I (1985) A family of bivariate distributions generated by the bivariate Bernoulli distribution. J Am Stat Assoc 80:332–338
Morris CN, Lock KF (2009) Unifying the named natural exponential families and their relatives. Am Stat 63(3):247–253
Nakagawa T, Yoda H (1977) Relationships among distributions. IEEE Trans Reliab 26(5):352–353
Patel JK, Kapadia CH, Owen DB (1976) Handbook of statistical distributions. Marcel Dekker, New York
Patil GP, Boswell MT, Joshi SW, Ratnaparkhi MV (1985a) Discrete models. International Co-operative Publishing House, Burtonsville
Patil GP, Boswell MT, Ratnaparkhi MV (1985b) Univariate continuous models. International Co-operative Publishing House, Burtonsville
Shapiro SS, Gross AJ (1981) Statistical modeling techniques. Marcel Dekker, New York
Song WT (2005) Relationships among some univariate distributions. IIE Trans 37:651–656
Taha HA (1982) Operations research: an introduction, 3rd edn. Macmillan, New York

Renewal Processes

Kosto V. Mitov[1], Michael A. Zazanis[2]
[1]Professor, Faculty of Aviation
National Military University, Pleven, D. Mitropolia, Bulgaria
[2]Professor
Athens University of Economics and Business, Athens, Greece

Let $\{X_n; n \in \mathbb{N}\}$ be a sequence of independent, identically distributed random variables with values in $\mathbb{R}^+$ and distribution function F. The process $\{S_n; n \in \mathbb{N}_0\}$ defined

by means of $S_0 := 0$, $S_n := S_{n-1} + X_n$, $n = 1, 2, \ldots$ is called an *ordinary renewal process*. The non-negative random variables X_n are called increments or, in many applications, inter-event times. In connection with the sequence of random points in time, $\{S_n\}$, one can define the *counting process* $N_t = \sum_{n=0}^{\infty} 1(S_n \le t)$, $t \in \mathbb{R}^+$, where $1(A)$ designates the *indicator function* of the event A (which is 1 if A occurs and 0 otherwise). The *renewal function* associated with a renewal process is the increasing, right-continuous function $U(t) := EN_t = \sum_{n=0}^{\infty} F^{*n}(t)$ where $F^{*n}(t)$ denotes the n-fold *convolution* of the distribution function F with itself (hence $F^{*n}(t) = P(S_n \le t)$).

Renewal processes are intimately related to the theory of the so-called *renewal equation* which is a linear integral equation of the form

$$Z(x) = z(x) + \int_0^x Z(x-y)F(dy) \tag{1}$$

where $z : \mathbb{R}^+ \to \mathbb{R}$ is a Borel function, bounded on finite intervals, and F a probability distribution on $\mathbb{R}^+$. F and z are assumed to be given and the object is to determine the (unique) solution Z which is bounded on finite intervals, and study its asymptotic behavior as $x \to \infty$. Its solution is given, in terms of the renewal function by the convolution $Z(x) = \int_0^x z(x-y)U(dy)$.

Renewal processes are important as special cases of random ▶point processes. In this respect the Poisson process (see ▶Poisson Processes) on the real line is the simplest and most important renewal process. They occur naturally in the theory of replacement of industrial equipment, the theory of queues, in branching processes, and in many other applications. In the framework of perpetual replacement of a single item, X_n is the life of the nth such item which, as soon as it fails, is replaced by a new one with independent duration distributed according to F. Then N_t is the number of items used in the time interval $[0, t]$ and S_{N_t} is the time of the last replacement before t. We define three additional processes $\{A_t; t \ge 0\}$, $\{B_t; t \ge 0\}$, and $\{C_t; t \ge 0\}$ as follows: $A_t := t - S_{N_t - 1}$ is the *age*, $B_t := S_{N_t} - t$ is the *remaining life*, and $C_t := A_t + B_t = X_{N_t}$ is the total life duration of the item currently in use. (The age and remaining life are also known as the backward and forward recurrence times.) The statistics of these processes can be described by means of appropriate renewal equations. For instance, if $W_x(t) := P(A_t \le x)$ then conditioning on S_1 (using the so-called "renewal argument") we obtain

$$W_x(t) = (1 - F(t))1(t \le x) + \int_0^t W_x(t-s)dF(s). \tag{2}$$

If we allow the first increment to have a different distribution from all the others, i.e. if we set $S_0 = X_0$ and $S_n = S_{n-1} + X_n$, $n = 1, 2, \ldots$ where X_0 is independent of the $\{X_n\}$ and, unlike them, has distribution F_0, different from F, we obtain a *delayed renewal process*. This type of process is important because it provides additional flexibility in accommodating different initial conditions. Of course, its limiting properties are not affected by this modification. Of particular importance, assuming the mean m to be finite, is the choice $F_0 = F_I$, given by

$$F_I(x) := \frac{1}{m} \int_0^x (1 - F(y))dy. \tag{3}$$

With this choice, $\{S_n\}$ becomes a *stationary point process*. F_I is called the *integrated tail distribution* associated with the distribution F.

Of fundamental importance are the limit theorems related to renewal processes. If $m := \int_0^{\infty} xdF(x)$ denotes the mean of the increments, then the *Elementary Renewal Theorem* states that $\lim_{t\to\infty} t^{-1}U(t) = m^{-1}$. (The result holds also in the case $m = \infty$ provided that we interpret m^{-1} as 0.) A refinement is possible if the increments have finite second moment, in which case $\lim_{t\to\infty} (U(t) - t/m) = EX_1^2/(2m^2)$. An analogous bound, due to Lorden (1970), also holds for all $t \ge 0$: $U(t) \le t/m + EX_1^2/m^2$. When the second moment exists we also have a Central Limit Theorem for the number of events up to time t: As $t \to \infty$, $\frac{N_t - t/m}{\sigma\sqrt{t/m^3}} \xrightarrow{d} Z$ where Z is a standard Normal random variable and $\sigma^2 = Var(X_1)$.

Much deeper is *Blackwell's Theorem* which states that, if F in non-lattice and the mean m is finite then

$$\lim_{t\to\infty} (U(t+h) - U(t)) = h/m \qquad \text{for all } h > 0. \tag{4}$$

(A distribution F on $\mathbb{R}^+$ is *lattice* with lattice size δ if there exists $\delta > 0$ such that the support of F is a subset of $\{n\delta; n = 0, 1, 2, \ldots\}$ and δ is the largest such number.) If F is lattice (δ) then (4) still holds, provided that h is an integer multiple of δ. Also, if $m = \infty$ the theorem still holds with $m^{-1} = 0$. Blackwell's original proof (1948) of (4) depended on harmonic analysis techniques. In the 1970s with the widespread use of coupling techniques simpler probabilistic proofs of the renewal theorem became available. (See Lindvall [1992] for a complete account.) An integral version of Blackwell's theorem, the *Key Renewal Theorem*, states that, if z is directly Riemann integrable then the limit $\lim_{x\to\infty} \int_0^x z(x-y)dU(y)$ exists and equals $m^{-1} \int_0^{\infty} z(x)dx$. This then gives the limiting behavior of any function which satisfies a renewal equation (1). (Direct Riemann integrability is a direct extension of the Riemann

integral from bounded intervals to unbounded ones: Fix $h > 0$ and let $\overline{\gamma}_n(h) = \sup_{nh \le x < (n+1)h} z(x)$, $\underline{\gamma}_n(h) = \inf_{nh \le x < (n+1)h} z(x)$. Set $\overline{I}(h) := \sum_{n=0}^{\infty} h\overline{\gamma}_n(h)$ and $\underline{I}(h) := \sum_{n=0}^{\infty} h\underline{\gamma}_n(h)$. Clearly, if $h_1 > h_2 > 0$ then $\underline{I}(h_1) \le \underline{I}(h_2) \le \overline{I}(h_2) \le \overline{I}(h_1)$, though these quantities may not necessarily be finite. If $\lim_{h\to 0} \underline{I}(h)$ and $\lim_{h\to 0} \overline{I}(h)$ exist and are equal then z is directly Riemann integrable. It should be noted that the direct Riemann integral is more restrictive than either the improper Riemann integral or the Lebesgue integral.)

The discrete version of the renewal theorem is simpler but not elementary. Suppose we are given a probability distribution $\{f_n; n = 1, 2, \ldots\}$ which is *non–arithmetic*, i.e. g.c.d.$\{n : f_n > 0\} = 1$ and has mean $m = \sum_{n=1}^{\infty} nf_n$, and define the renewal sequence $\{u_n; n = 0, 1, 2, \ldots\}$ via $u_0 = 1$, $u_n = f_n + f_{n-1}u_1 + \cdots + f_1 u_{n-1}$. Then $\lim_{n\to\infty} u_n = m^{-1}$ (interpreted as 0 when $m = \infty$). This is the celebrated Erdös–Feller–Pollard (1948) renewal theorem (see Feller [1968, 1971, Vol. 1, Chap. 13]) which marks the beginning of modern renewal theory and played a central rôle in the treatment of ▶Markov chains with countable state space. Interesting behavior arises if the non-arithmetic distribution function $\{f_n\}$ has infinite mean: Suppose that $\sum_{k=n+1}^{\infty} f_k = L(n)n^{-\alpha}$ where $0 < \alpha < 1$ and $L(n)$ is a *slowly varying* function. (A real function L is said to be slowly varying if it is positive, measurable, and for every $\lambda > 0$, $L(\lambda x)/L(x) \to 1$ as $x \to \infty$.) Then (Garsia and Lamperti 1962) $\lim_{n\to\infty} n^{1-\alpha} L(n)u_n = \pi^{-1} \sin \pi\alpha$. If $1/2 < \alpha < 1$, this can be sharpened to $\lim_{n\to\infty} n^{1-\alpha} L(n)u_n = \pi^{-1} \sin \pi\alpha$. Analogous results in continuous time are also proved. Suppose that $F(.)$ is continuous, $F(0+) = 0$, $F(\infty) = 1$, $m = \infty$, and

$$1 - F(t) \sim \frac{t^{-\alpha} L(t)}{\Gamma(1-\alpha)} \Leftrightarrow m(t) := \int_0^t (1 - F(u))du \sim \frac{t^{1-\alpha} L(t)}{\Gamma(2-\alpha)}, \; t \to \infty, \qquad (5)$$

where $\alpha \in [0, 1)$ and $L(\cdot)$ is a slowly varying function at infinity. Under these conditions the growth rate of $U(t)$ is given by (see e.g. Bingham et al. [1987, Chap. 8]),

$$U(t) \sim C_\alpha t/m(t), \text{ as } t \to \infty, \text{ where } C_\alpha = [\Gamma(\alpha+1)\Gamma(2-\alpha)]^{-1}.$$

Erickson (1970) proved a version of Blackwell's theorem in the infinite mean cycle case. It states that if in (5), $\alpha \in \left(\frac{1}{2}, 1\right]$, then for any fixed $h > 0$

$$\lim_{t\to\infty} m(t)[U(t) - U(t-h)] = C_\alpha h.$$

If $\alpha \in \left(0, \frac{1}{2}\right]$, then lim has to be replaced by lim. Several versions of the Key Renewal Theorem in the infinite mean cycle case are also proved in Teugels (1968), Erickson (1970), and Anderson and Athreya (1987).

Using the Key Renewal Theorem one can obtain the asymptotic behavior of the age and the current and residual life. If Y is a random variable with distribution $P(Y \le y) = \frac{1}{m}\int_0^y x dF(x)$ and V is uniform in $[0, 1]$ and independent of Y, then

$$(A_t, B_t, C_t) \xrightarrow{d} (VY, (1-V)Y, Y) \text{ as } t \to \infty.$$

In particular the limiting marginal distribution of the age (which is the same as that of the residual life) is

$$\lim_{t\to\infty} P(A_t \le x) = F_I(x),$$

the integrated tail distribution given in (3). The limiting behavior of these processes gives rise to the so called "renewal paradox." For instance, the limiting distribution of the item currently in use is

$$\lim_{t\to\infty} P(C_t \le x) = \frac{1}{m}\int_0^x y dF(y)$$

with corresponding mean, provided that the second moment of F exists, given by $m + \sigma^2/m$. Hence if we inspect such a process a long time after it has started operating (and is therefore in equilibrium) the part we are going to see will have longer life duration than average. Of course this is simply an instance of *length-biased sampling* and its effects are more pronounced when the variability of the distribution F around its mean is large.

In the infinite mean cycle case the life time processes A_t and B_t have a linear growth to infinity, i.e. the normalized processes A_t/t and B_t/t have non-degenerate limit laws, jointly or separately. This result is usually called the Dynkin–Lamperti theorem (Dynkin 1955; Lamperti 1962). (See also Bingham et al. [1987, Chap. 8]). The theorem states that the condition (5) with $\alpha \in (0, 1)$ is necessary and sufficient for the existence of non-degenerate limit laws for A_t/t, B_t/t,

$$\lim_{t\to\infty} P(A_t/t \le x) = \pi^{-1} \sin \pi\alpha \int_0^x u^{-\alpha}(1-u)^{\alpha-1}, \quad 0 < x < 1,$$

$$\lim_{t\to\infty} P(B_t/t \le x) = \pi^{-1} \sin \pi\alpha \int_0^x u^{-\alpha}(1+u)^{-1} du, \quad x > 0.$$

An important and immediate generalization of the renewal equation (1) is to allow F to be a general positive finite measure on $\mathbb{R}^+$. Setting $\|F\| := F(\mathbb{R}^+)$ one distinguishes the *excessive* case where $\|F\| > 1$, the *defective* case where $\|F\| < 1$, and the *proper* case we have already discussed, where $\|F\| = 1$. In the excessive case one can always find

a (unique) $\beta > 0$ such that $\int_0^\infty e^{-\beta x} dF(x) = 1$. One can define then a probability distribution function $F^\#$ via the relationship $dF^\#(x) = e^{-\beta x} dF(x)$, $x \geq 0$. Multiplying both sides of (1) by $e^{-\beta x}$ and setting $z^\#(x) = e^{-\beta x} z(x)$, $Z^\#(x) = e^{-\beta x} Z(x)$, the proper renewal equation $Z^\#(x) = z^\#(x) + \int_0^x z^\#(x-y) dF^\#(y)$ is obtained. The Key Renewal Theorem then yields

$$\lim_{x\to\infty} e^{-\beta x} Z(x) = \frac{1}{m^\#} \int_0^\infty z^\#(y) dy,$$

which establishes that, asymptotically, Z grows exponentially with rate β. We should point out that the defective case is not entirely similar. While formally one again tries to identify $\beta > 0$ so that $\int_0^\infty e^{\beta x} dF(x) = 1$, this may or may not be possible according to whether the distribution function $\frac{1}{\|F\|} F(x)$ is *light-tailed* or *heavy-tailed*. In the former case one proceeds just as in the excessive case. (For more details see Feller [1968, 1971, Vol. 2, Chap. 11]). This type of analysis is characteristic of the applications of renewal theory to areas such as population dynamics, the theory of collective insurance risk, and to the economic theory or replacement and depreciation (Jorgenson 1974; Feldstein and Rothchild 1974).

Alternating renewal processes arise in a natural way in many situations, like queueing systems and reliability of industrial equipment, where working(busy) periods (X) interchange with idle periods (T). Consider a sequence of random vectors with non-negative coordinates (T_i, X_i), $i = 1, 2, \ldots$. It defines an *alternating renewal sequence* (S_n, S'_{n+1}) as follows $S_0 = 0$, $S'_n = S_{n-1} + T_n$, $S_n = S'_n + X_n = S_{n-1} + (T_n + X_n)$, $n = 1, 2, \ldots$. An interpretation in terms of the reliability theory is the following. There are two types of renewal events: S_n is the moment when the installation of a new element begins (The installation takes time T_n); S'_{n+1} is the moment when the installation ends and the new element starts working. (The working period has length X_n). The renewal process $N(t) = \sup\{n : S_n \leq t\}$ counts the pairs of renewal events in the interval $[0, t]$. The processes $\sigma_t = \max\{0, t - S'_{N(t)+1}\}$ – *spent working time* and $\tau_t = \min\{S_{N(t)+1} - t, X_{N(t)+1}\}$ – *residual working time* generalize the lifetime processes A_t and B_t. Their properties are derived in Mitov and Yanev (2001) in the infinite mean cycle case.

The central place that renewal theory holds in the analysis of stochastic systems is due to the concept of *regeneration*. Let $\{X_t; t \in \mathbb{R}^+\}$ be a process with values in $\mathcal{S}$ (e.g. a Euclidean space $\mathbb{R}^d$) and sample paths that are càdlàg (right-continuous with left-hand limits) a.s.. Such a process is called *regenerative* with respect to a (possibly delayed) renewal process $\{S_n\}$, defined on the same probability space, if, for each $n \in \mathbb{N}$ the *post* S_n *process* $(\{X_{S_n+t}\}_{t\geq 0}, \{S_{n+k} - S_n\}_{k\in\mathbb{N}})$ is independent of $\{S_0, S_1, \ldots, S_n\}$ and its distribution does not depend on n, i.e. $(\{X_{S_n+t}\}_{t\geq 0}, \{S_{n+k} - S_n\}_{k\in\mathbb{N}}) \stackrel{d}{=} (\{X_{S_0+t}\}_{t\geq 0}, \{S_k - S_0\}_{k\in\mathbb{N}})$ for all n. The existence of an embedded, non-lattice renewal process with respect to which the process $\{X_t\}$ is regenerative, together with the finiteness of the mean $m := E[S_1 - S_0]$ is enough to guarantee the existence of a "stationary version," say $\{\tilde{X}_t\}$, to which $\{X_t\}$ converges as t goes to infinity. The statistics of $\{\tilde{X}_t\}$ can be determined by analyzing the behavior of $\{X_t\}$ over any *regenerative cycle*, i.e. a random time interval of the form $[S_n, S_{n+1})$. If $k \in \mathbb{N}$, $t_i \in \mathbb{R}^+$, $i = 1, 2, \ldots, k$, and $f : \mathcal{S}^k \to \mathbb{R}$ is any bounded, continuous function then

$$Ef(\tilde{X}_{t_1}, \ldots, \tilde{X}_{t_k}) = \frac{1}{m} E \int_{S_0}^{S_1} f(X_{t_1+t}, \ldots, X_{t_k+t}) dt.$$

Nowadays, our view of whole areas of probability, including parts of the theory of ▶Markov processes is influenced by renewal theoretic tools and related concepts of regeneration. The analysis of many stochastic models is greatly facilitated if one identifies certain embedded points in time that occur according to a renewal process and with respect to which the process is regenerative. The fact that these regeneration cycles are independent, identically distributed, also facilitates the statistical analysis of the simulation output of regenerative systems.

A detailed representation of the renewal theory and its applications could be found, for instance, in the following books Asmussen (2003), Bingham et al. (1987), Feller (1968, 1971), and Resnick (1992).

Cross References

▶Point Processes
▶Poisson Processes
▶Queueing Theory
▶Statistical Inference for Stochastic Processes
▶Stochastic Processes: Classification

References and Further Reading

Asmussen S (2003) Applied probability and queues, 2 edn. Springer, New York

Anderson KK, Athreya KB (1987) A renewal theorem in the infinite mean case. Ann Probab 15:388–393

Bingham NH, Goldie CM, Teugels JL (1987) Regular variation. Cambridge University Press, Cambridge

Dynkin EB (1955) Limit theorems for sums of independent random quantities. Izves Akad Nauk U.S.S.R 19:247–266

Erickson KB (1970) Strong renewal theorems with infinite mean. Trans Am Math Soc 151:263–291

Feldshtein M, Rotschild M (1974) Toward an economic theory of replacement investment, Econometrica 42(3):393–423
Feller W (1968, 1971) An introduction to probability theory and its applications, vol 1 and 2. Wiley, New York
Garsia A, Lamperti J (1962) A discrete renewal theorem with infinite mean. Comment Math Helv 37:221–234
Jorgenson DW (1974) Investment and production: a review. In: Intriligator M, Kendrick D (eds) Frontiers of quantitative economics, vol 2. Amsterdam, North-Holland, pp 341–366
Lamperti J (1962) An invariance principle in renewal theory. Ann Math Stat 33:685–696
Lindvall T (1992) Lectures on the coupling method. Wiley, New York
Lorden G (1970) On the excess over the boundary. Ann Math Stat 41:520–527
Mitov KV, Yanev NM (2001) Limit theorems for alternating renewal processes in the infinite mean case. Adv Appl Probab 33:896–911
Resnick S (1992) Adventures in stochastic processes. Birkhäuser, Boston

Repeated Measures

Geert Molenberghs
Professor
Universiteit Hasselt & Katholieke Universiteit Leuven, Leuven, Belgium

Repeated measures are obtained whenever a specific response is measured repeatedly in a set of units. Examples are hearing thresholds measured on both ears of a set of subjects, birth weights of all litter members in a toxicological animal experiment, or weekly blood pressure measurements in a group of treated patients. The last example is different from the first two examples in the sense that the time dimension puts a strict ordering on the obtained measurements within subjects. The resulting data are therefore often called longitudinal data. Obviously, a correct statistical analysis of repeated measures or longitudinal data can only be based on models which explicitly take into account the clustered nature of the data. More specifically, valid models should account for the fact that repeated measures within subjects are allowed to be correlated. For this reason, classical (generalized) linear regression models are not applicable in this context. An additional complication arises from the highly unbalanced structure of many data sets encountered in practice. Indeed, the number of available measurements per unit is often very different between units, and, in the case of longitudinal data, measurements may have been taken at arbitrary time points, or subjects may have left the study prematurely, for a number of reasons (sometimes known but mostly unknown). A large number of models have been proposed in the statistical literature, during the last few decades. Overviews are given in Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005).

About the Author

For biography *see* the entry ►Linear Mixed Models.

Cross References

►Nonlinear Mixed Effects Models
►Research Designs
►Sample Survey Methods
►Statistical Analysis of Longitudinal and Correlated Data

References and Further Reading

Brown H, Prescott R (1999) Applied mixed models in medicine. Wiley, New York
Crowder MJ, Hand DJ (1990) Analysis of repeated measures. Chapman & Hall, London
Davidian M, Giltinan DM (1995) Nonlinear models for repeated measurement data. Chapman & Hall, London
Davis CS (2002) Statistical methods for the analysis of repeated measurements. Springer, New York
Demidenko E (2004) Mixed models: theory and applications. Wiley, New York
Diggle PJ, Heagerty PJ, Liang KY, Zeger SL (2002) Analysis of longitudinal data, 2nd edn. Oxford University Press, Oxford
Fahrmeir L, Tutz G (2002) Multivariate statistical modelling based on generalized linear models, 2nd edn. Springer, New York
Fitzmaurice GM, Davidian M, Verbeke G, Molenberghs G (2009) Longitudinal data analysis. Handbook. Wiley, Hoboken
Goldstein H (1995) Multilevel statistical models. Edward Arnold, London
Hand DJ, Crowder MJ (1995) Practical longitudinal data analysis. Chapman & Hall, London
Hedeker D, Gibbons RD (2006) Longitudinal data analysis. Wiley, New York
Kshirsagar AM, Smith WB (1995) Growth curves. Marcel Dekker, New York
Leyland AH, Goldstein H (2001) Multilevel modelling of health statistics. Wiley, Chichester
Lindsey JK (1993) Models for repeated measurements. Oxford University Press, Oxford
Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2005) SAS for mixed models, 2nd edn. SAS Press, Cary
Longford NT (1993) Random coefficient models. Oxford University Press, Oxford
Molenberghs G, Verbeke G (2005) Models for discrete longitudinal data. Springer, New York
Pinheiro JC, Bates DM (2000) Mixed effects models in S and S-Plus. Springer, New York
Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York
Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer Series in Statistics. Springer, New York
Vonesh EF, Chinchilli VM (1997) Linear and non-linear models for the analysis of repeated measurements. Marcel Dekker, Basel

Weiss RE (2005) Modeling longitudinal data. Springer, New York
West BT, Welch KB, Gałecki AT (2007) Linear mixed models: a practical guide using statistical software. Chapman & Hall/CRC Press, Boca Raton
Wu H, Zhang J-T (2006) Nonparametric regression methods for longitudinal data analysis. Wiley, New York
Wu L (2010) Mixed effects models for complex data. Chapman & Hall/CRC Press, Boca Raton

Representative Samples

Ksenija Dumicic
Full Professor, Head of Department of Statistics, Faculty of Economics and Business
University of Zagreb, Zagreb, Croatia

According to Lavrakas (2008), a representative sample is one that ensures external validity in relationship to the population of interest the sample is meant to represent. In addition, it should be said that a representative sample is a probability sample, so, *sampling errors* for estimates can be calculated and the *estimates* from the sample survey can be generalized with confidence to the sampling population.

Random selection, i.e., being objective and unbiased, is an essential element of *survey sampling*. There are many factors that affect the representativeness of a sample, though traditionally attention has mostly been paid to sample design and coverage. More recently, the focus has extended to the nonresponse issues.

Zarkovic (1956), Kruskal and Mosteller (1980), Bellhouse (1988), Kish (2002), and Rao (2005) wrote histories of random sampling methods as representative methods. The statistical literature gives examples of all the meanings for "representative sampling," such as: general acceptance for data; absence of selective forces; miniature of the population; typical or ideal cases; and proper coverage of the population. Kruskal and Mosteller (1979) added the following new meanings: representative sampling as a specific sampling method; representative sampling as permitting unbiased estimation; and representative sampling as sufficient to serve a particular purpose. Occasionally, it is also determined as a vague term.

The development of modern *sampling theory* started in around 1895 when the Norwegian statistician A.N. Kiaer, the first director of Statistics Norway, published his *Representative Method* and was the first to promote "the representative method" over the ►*census* as a complete enumeration. Kiaer stated that if a sample was representative with respect to variables for which the population distribution was known, it would also be representative with respect to other survey variables. For Kiaer, a representative sample is a "miniature" of the actual population, though the selection of units is based on *purposive selection*, according to a rational scheme based on general results of previous investigations. He presented his thoughts at a meeting of the International Statistical Institute in Bern in 1895. Many famous statisticians did not agree on Kiaer's new approach, as no measure of the accuracy of the estimates could be obtained. A basic problem was that the representative method lacked a formal theory of inference.

It was Sir A.L. Bowley, an English statistician, who pioneered the use of *simple random sampling*, for which the accuracy measures of estimates could be computed. He introduced *stratified random sampling* with *proportional allocation*, leading to a representative sample with equal *inclusion probabilities*. By the 1920s, the representative method was widely used. In 1924, the International Statistical Institute played a prominent role with its formation of a committee to report on the representative method. In 1935, Polish scientist J. Neyman published his now famous paper (Neyman 1934). He developed a new theory and laid the theoretical foundations for design-based sampling or the probability sampling approach to inference from survey samples. He showed that stratified random sampling is preferable to ►balanced sampling and introduced the *optimal allocation* of units based on efficiency in his theory of stratified random sampling without replacement, by relaxing the condition of equal inclusion probabilities for sampling units. In 1943, M.H. Hansen and W.N. Hurwitz published their theory of *multistage cluster samples*. In 1944, W.G. Madow and L.H. Madow conducted the first theoretical study of the precision of *systematic sampling*. The classical theory of survey sampling was more or less completed in 1952. Horvitz and Thompson (1952) completed the classical theory, and the *random sampling* approach was almost unanimously accepted. Most of the classical books on *sampling* were also published by then: W.G. Cochran in 1953 (see the last edition: Cochran 1977), Deming (1950), Hansen et al. (1953a, b), and Yates (1949). Later, a great contribution to probability sampling was given by Kish (1965).

The representative method is applied in *survey research*, *both social and business*, in *official statistics*, for *public opinion polling*, in *market research*, etc. It is also applied for *audit sampling* and *statistical quality control*.

The sampling technique used by G. Gallup was *quota sampling* for opinion polling. Gallup's approach was in great contrast with that of *Literary Digest* magazine, the

leading polling organization at that time. This magazine conducted regular "America Speaks" polls with a *convenient sample* of the sample size near to two million people. The presidential election of 1936 turned out to be decisive for both approaches (Rao 2005). Gallup correctly predicted using a sample size of 3,000 that the candidate Alf Landon would beat Franklin Roosevelt. It seemed to be strange how could a prediction based on such a large sample be wrong? The explanation was the fatal flaw in the *Literary Digest*'s sampling mechanism. The automobile registration lists and telephone directories applied were not representative samples. In the 1930s, cars and telephones were typically owned by the middle and upper classes. More well-to-do Americans tended to vote Republican and the less well-to-do were inclined to vote Democrat. Therefore, Republicans were over represented in the *Literary Digest* sample. As a result of this famous historical error, opinion researchers learned that the *manner of selection* of a sample is more important than the *sample size*. Also, among nonprobability sample designs, a *quota sample*, if well designed, would be the most similar to a probability sample as a representative one.

In a sample survey, researchers must judge whether the sample is actually representative of the target population. The best way of ensuring a representative sample is to have a complete sampling frame (i.e., directory, list or map) covering all the elements in the population, and to know that each and every element (e.g., person, household, enterprise, etc.) on the list has a nonzero probability (equal or unequal) of being included in the sample. Furthermore, it is necessary to use random selection to draw elements from the sampling frame into the sample based either on a random number generator or on systematic selection procedure. Also, it is essential to collect complete data from every single sampled element.

Completely up-to-dated sampling frames of the populations of interest are very rare. If there are elements in the target population with a zero probability of selection, sample estimates cannot be generalized to these elements. For example, if unemployed persons belong to the population of interest, but were not registered as unemployed, then they would have a zero probability of inclusion in the sample. Further, very modern Internet surveys, Tele-Voting and Push Polling samples are not based on solid sampling frames and instead use non-representative sample designs. As such, the results in such surveys cannot be generalized to the overall population and users should be aware that these are nothing more than amusement techniques.

First, to judge the representativeness of a sample, the use of some prior knowledge about the population main variables structures is recommended, for comparison with the sample structures. Occasionally, an extra random sample is helpful. Further, to correct for biases, survey researchers apply post-stratification. Post-stratification is the process of weighting some of the respondents in the responding sample relative to others, so that the characteristics of the responding sample are essentially equal to those of the target population for those characteristics that can be controlled to complete coverage data (e.g., age, gender, educational level, geography, etc.). Applying post-stratification adjustments reduces the bias due to *noncoverage and nonresponse*. And finally, it is necessary to limit the conclusions to those elements in the sampling frame to only those with a nonzero probability of inclusion. In other words, to avoid biases, researchers need to estimate coverage, and both unit and item-nonresponse.

About the Author

Dr. Ksenija Dumicic is Professor, and, since 2006 Head of Department of Statistics, Faculty of Economics and Business, University of Zagreb, Croatia. She is founder of the postgraduate studies, Statistical Methods for Economic Analysis and Forecasting. She has been a member of Editorial Boards for several journals in Croatia, Serbia and Bosnia and Herzegovina. She has authored and coauthored more than 80 papers and two books in statistical research methodology. She specializes in statistical sample survey methods and has supervised over 10 postgraduate students.

Cross References

►Balanced Sampling
►Non-probability Sampling Survey Methods
►Nonresponse in Surveys
►Nonresponse in Web Surveys
►Sample Survey Methods
►Telephone Sampling: Frames and Selection Techniques

References and Further Reading

Bellhouse DR (1988) A brief history of random sampling methods. In: Krishnaiah PR, Rao CR (eds) Handbook of statistics: sampling. Elsevier, Amsterdam, pp 1–14
Cochran WG (1977) Sampling tehniques, 3rd edn. Wiley, New York
Deming WE (1950) Some theory of sampling. Wiley, New York
Hansen MH, Hurwitz WN, Madow WG (1953a) Sample survey methods and theory, volume I. Wiley, New York
Hansen MH, Hurwitz WN, Madow WG (1953b) Sample survey methods and theory, volume II. Wiley, New York
Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685
Kish L (1965) Survey sampling. Wiley, New York
Kish L (2002) New paradigms (models) for probability sampling. In: Survey methodology, vol 28(1), Statistics Canada, Catalogue No. 12001XIE, pp 31–34
Kruskal W, Mosteller F (1979) Representative sampling, III: the current statistical literature. Int Stat Rev 47:245–265

Kruskal WH, Mosteller F (1980) Representative sampling IV: the history of the concept in statistics, 1895–1939. Int Stat Rev 48:169–195
Lavrakas PJ (2008) Encyclopedia of survey research methods, volume 2. SAGE Publications, California
Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J Roy Stat Soc 97:558–606
Rao JNK (2005) Interplay between sample survey theory and practice: an appraisal. In: Survey methodology, vol 31(2), Statistics Canada, Catalogue No. 12001XPB, pp 117–138
Yates F (1949) Sampling methods for censuses and surveys. Griffin, London
Zarkovic SS (1956) Note on the history of sampling methods in Russia. J Roy Stat Soc A 119:336–338

Research Designs

Roxana Toma, G. David Garson
North Carolina State University, Raleigh, NC, USA

Introduction

Research design is a term broadly referring to any plan for gathering data systematically in such a way as to be able to arrive at conclusions. On the subject selection dimension, designs may be experimental, quasi-experimental, or non-experimental. On the measurement dimension, designs may be between-subjects or within-subjects.

Experimental studies are characterized by ►randomization of subjects into treatment and control groups. Control groups may receive no treatment or some standard treatment, where "treatment" is exposure to some stimulus. Randomization serves to control for variables which are not included explicitly in the study. In quasi-experimental designs, treatment and comparison groups are not composed of randomized subjects, even if data are gathered through random sampling. In the absence of randomization, control for confounding variables must be accomplished explicitly through statistical techniques. Finally, a design is non-experimental if there is systematic collection of data with respect to topics of interest but there is no randomization of subjects as in experimental studies nor statistical controls as in quasi-experimental designs. Most case study designs exemplify this category.

By type of measurement, between-subjects designs are the most common. The researcher is comparing between subjects who experience different treatments. There are different subjects for each level of the independent variable(s). Any given subject is exposed to only one level and comparisons are made between subjects' reactions or effects. In contrast, in within-subjects designs the same subjects are measured for each level of the independent variable, as in before-after studies or panel studies. Similar subjects, as in matched pair's designs, are of the same type. When subjects are measured more than once, within-subjects designs are also called repeated measures designs. Since the subjects are the same for all levels of the independent variable(s), they are their own controls (i.e., subject variables are controlled). However, there is greater danger to validity in the form of carryover effects due to exposure to earlier levels in the treatment sequence (e.g., practice, fatigue, attention) and there is danger of attrition in the sample. Counterbalancing is a common but not foolproof strategy to address carryover effects: e.g., half the subjects get treatment *A* first, then *B*, while the other half get B first, then *A*.

Factorial and Block Designs

Factorial designs use categorical independent variables to establish groups. For instance, in a two factor design, the independent variables might be information type (fiction, non-fiction) and media type (television, print, Internet), generating two times three = six categories. A factorial design is "fully crossed" if there is a group for every possible combination of factors (independent variables). An "incomplete" factorial design, leaving out some of the groups, may be preferred if some combinations of values of factors are nonsensical or of no theoretical interest. In experimental designs, an equal number of subjects are assigned randomly to each of the six possible groups (e.g., to the fiction-television group). The researcher might then measure subjects on information retention. A null outcome would be indicated by the average retention score being the same for all six groups of the factorial design. Unequal mean retention scores would indicate a main effect of information type or media type, and/or an interaction effect of both. Quasi-experimental designs may also be factorial, but groups are established by stratified random sampling, not randomization of subjects, entailing the need for more complex and explicit statistical controls and possibly less conclusive results.

Balanced designs are simply factorial designs where there are equal numbers of cases in each subgroup (cell) of the design, assuring that the factors are independent of one another (but not necessarily the covariates). Unbalanced designs have unequal n's in the cells formed by the intersection of the factors.

Randomized block designs stratify the subjects and for each strata, a factorial design is run. This is typically done when the researcher is aware of nuisance factors that

need to be controlled (e.g., there might be an air conditioned room stratum and a no air conditioning stratum) or if there were other mitigating structural factors known in advance (e.g., strata might be different cities). That is, the blocking variables which stratify the sample are factors which are considered to be control variables, not independent variables as they would be in a simple factorial design. Randomized block designs seek to control for the effects of main factors and their interactions, controlling for the blocking variable(s).

Nested designs have two or more factors, but the levels of one factor are never repeated as levels in the other factor(s). This happens in hierarchical designs, for instance, when a forester samples trees, then samples seedlings of each sampled tree for survival rates. The seedlings are unique to each tree and represent a random factor. Likewise, a researcher could sample drug companies, then could sample drug products for quality within each sampled company. This contrasts with crossed designs of ordinary two-way (or higher) analysis of variance, in which the levels of one factor appear as levels in another factor (e.g., tests may appear as levels across schools). We can get the mean of different tests by averaging across schools, but we cannot get the mean survival rate of different seedlings across trees because each tree has its own unique seedlings. Likewise, we cannot compute the mean quality rating for a drug product across companies because each company has its own unique set of products. Latin square and Graeco-Latin square designs discussed below are nested designs.

Random Versus Fixed Effects Designs

Most designs are fixed effects models, meaning that data are collected on all categories of the independent variables. In random effects models (also called random factors models), in contrast, data are collected only for a sample of categories. There is replaceability, meaning that the levels of the factor are randomly or arbitrarily selected and could be replaced by other, equally acceptable levels. The purpose of random effects modeling is generalizability – the researcher wishes to generalize findings beyond the particular, randomly or arbitrarily selected levels in the study. For instance, a researcher may study the effect of item order in a ►questionnaire. Six items could be ordered 720 ways. However, the researcher may limit him- or herself to the study of a sample of six of these 720 ways. The random effects model in this case would test the null hypothesis that the effects of ordering are zero. Note that "mixed factorial design" is also possible simply by having a random effects model with a fixed factor and a random factor.

Treatment by replication design is a common random effects model. The treatment is a fixed factor, such as exposure to different types of public advertising, while the replication factor is represented by the particular respondents who are treated. Sometimes it is possible and advisable to simplify analysis from a hierarchical design to a simple treatment by replication design by shifting the unit of analysis. An example would be to use class averages rather than student averages in a design in which students represent a random factor nested within teachers as another random factor (the shift drops the student random factor from analysis). Note also that the greater the variance of the random effect variable, the more levels are needed (e.g., more subjects in replication) to test the fixed (treatment) factor at a given alpha level of significance.

Common Experimental Designs

A very large number of research designs have been devised for experimental design. Though not exclusive to experimental design, the most prevalent examples are outlined below.

Completely randomized designs assign an equal number of subjects randomly to each of the cells formed by the factors. In the quasi-experimental mode, where there is no control by randomization, the researcher must measure and employ controls explicitly by using covariates.

Latin square designs extend block designs to control for two categorical variables. This design requires that the researcher assume all interaction effects are zero. Normally, if one had three variables, each of which could assume four values, then one would need $4^3 = 64$ observations just to have one observation for every possible combination. Under Latin square design, however, the number of necessary observations is reduced to $4^2 = 16$ because the third variable is nested. For instance, suppose there are 4 teachers, 4 classes, and 4 textbooks. The 16 groups in the design would be the 16 different class-textbook pairs. Each teacher would teach in each of the four classes, using a different text each time. Each class would be taught by each of the four different teachers, using a different text each time. However, only 16 of the 64 possible teacher-class-textbook combinations would be represented in the design because textbooks represent a nested factor, with each class and each teacher being exposed to a given textbook only once. Eliminating all but 16 cells from the complete (crossed) design requires the researcher to assume that there are no significant teacher-textbook or class-textbook interaction effects, only the main effects for teacher, class, and textbook.

Graeco-Latin square designs extend Latin square block designs to control for three categorical variables.

Split-plot designs. Like randomized complete block designs, in split plot designs there is still a blocking factor but each block is split into two segments and segments are assigned to the blocks in random order. Within any segment, treatments are assigned in random order. For instance, in a study of health improvement effects, the blocking factor might be in the form of three age groups, treatment in the form of three levels of dosage of medicine, with the segmentation variable being two brands of medicine. Splitting the three age blocks yields six segments, with each age group having a Brand A and Brand B segment. Each of the six segments is homogenous by brand. Split-plot designs are used when homogeneity rather than randomization within blocks is required (in agriculture, for instance, equipment considerations could dictate that any given plot segment only receive one brand of fertilizer).

Split-plot repeated measures designs can be used when the same subjects are measured more than once. In a typical split-plot repeated measures design, subjects are measured on some variable over a number of trials. Subjects are also split by some grouping variable. In this design, the between-subjects factor is the group (treatment or control) and the repeated measure is, for example, the test scores for two trials. The resulting statistical output will include a main treatment effect (reflecting being in the control or treatment group) and a group-by-trials interaction effect (reflecting treatment effect on posttest scores, taking pretest scores into account).

Common Quasi-Experimental Designs

As with experimental designs, numerous types of quasi-experimental designs exist, many enumerated in the classic work of Cook and Campbell (1979).

One-group posttest-only design. This design lacks a pretest baseline or a comparison group, making it impossible to come to reliable conclusions about a treatment effect.

Posttest-only design with nonequivalent comparison groups. In this common social science design, it is also impossible to come to reliable conclusions about treatment effect based solely on posttest information on two nonequivalent groups since effects may be due to treatment or to nonequivalencies between the groups.

Posttest-only design with predicted higher-order interactions. The presence of an interaction effect creates two or more expectations compared to the one-expectation one-group posttest-only design. Because there are more expectations, there is greater verification of the treatment effect but the explanation accounting for the interaction is more complex and therefore may be less reliable.

One-group pretest-posttest design. This is a common but flawed design subject to such threats to validity as history (events intervening between pretest and posttest), maturation (changes in the subjects that would have occurred anyway), regression toward the mean (the tendency of extremes to revert toward averages), testing (the learning effect on the posttest of having taken the pretest), and the like.

Two-group pretest-posttest design using an untreated control group. If a comparison group which does not receive treatment is added to what otherwise would be a one-group pretest-posttest design, threats to validity are greatly reduced. This is the classic experimental design but in quasi-experimental design, since the groups are not equivalent, there is still the possibility of selection bias.

Double pretest design. The researcher can strengthen pretest-posttest designs by having two (or more) pretest measures to establish if there is a trend in the data independent of the treatment effect measured by the posttest.

Regression-discontinuity design. If there is a treatment effect, then the slope of the regression line relating scores before and after treatment would be the same, but there would be a discontinuous jump in the intercept (and possibly also change in slope) following treatment. This design is extended in the simple interrupted time series design in which there are multiple pretests and posttests. The trend found in multiple pretests can be compared to the trend found in multiple posttests to assess whether apparent post-treatment improvement may simply be an extrapolation of a maturation effect which was leading toward improvement anyway.

Regression point displacement design. In this design there is a treatment group (e.g., a county) and a large number of comparison groups (e.g., other counties in the state). For instance, in a study of the effect of an after-school intervention on juvenile crime, the researcher might regress juvenile crime rates on median income in the pretest condition and note the position of the test county in the regression scattergram. In the posttest condition, the regression is re-run and the location of the test county is noted. If displaced on the scattergram, the researcher concludes that the intervention had an effect. This type of design assumes no misspecification of the model and assumes an invariant relation of independents to dependents between pretest and posttest.

Other designs. Cook and Campbell (1979) discussed other research designs for which space does not permit discussion here. These include nonequivalent dependent variables pretest-posttest designs, removed-treatment pretest-posttest designs, repeated-treatment designs, switching replications designs, reversed-treatment pretest-posttest nonequivalent comparison groups designs,

cohort designs with cyclical turnover, four-group designs with pretest-posttest and posttest-only groups, interrupted time series designs with a nonequivalent no-treatment comparison group, interrupted time series designs with nonequivalent dependent variables, interrupted time series designs with removed treatment, interrupted time series designs with multiple replications, and interrupted time series designs with switching replications.

About the Author

G. David Garson is a Professor of public administration at North Carolina State University, where he teaches graduate research methodology. He is Editor of the *Social Science Computer Review* and is author of *Statnotes: Topics in Multivariate Analysis* (http://faculty.chass.ncsu.edu/garson/PA765/statnote.htm), an online text utilized by over a million scholars and researchers each year. His recently authored or edited books include *Handbook of Research on Public Information Technology* (2008; 3rd ed. 2010), *Patriotic Information Systems: Privacy, Access, and Security Issues of Bush Information Policy* (2008), *Modern Public Information Technology Systems* (2007), and *Public Information Technology and E-Governance: Managing the Virtual State* (2006; 2nd ed. 2010). He may be contacted at david_garson@ncsu.edu.

Cross References

►Chi-Square Test: Analysis of Contingency Tables
►Clinical Trials: An Overview
►Design of Experiments: A Pattern of Progress
►Designs for Generalized Linear Models
►Experimental Design: An Introduction
►Factorial Experiments
►Farmer Participatory Research Designs
►Graphical Analysis of Variance
►Incomplete Block Designs
►Interaction
►Medical Research, Statistics in
►Multilevel Analysis
►Optimum Experimental Design
►Randomization
►Repeated Measures
►Selection of Appropriate Statistical Methods in Developing Countries
►Statistical Design of Experiments (DOE)
►Student's t-Tests
►Uniform Experimental Design

References and Further Reading

Bordens K, Abbott BB (2008) Research design and methods: a process approach, 7th edn. McGraw-Hill, New York
Cook TD, Campbell DT (1979) Quasi-experimentation: design and analysis issues for field settings. Houghton-Mifflin, Boston
Creswell JW (2008) Research design: qualitative, quantitative, and mixed methods approaches, 3rd edn. Sage Publications, Thousand Oaks, CA
Leedy P, Ormrod JE (2009) Practical research: planning and design, 9th edn. Prentice-Hall, Upper Saddle River, NJ
Levin IP (1999) Relating statistics and experimental design. Sage Publications, Thousand Oaks, CA. Quantitative Applications in the Social Sciences series #125
Pedhazur EJ, Schmelkin LP (1991) Measurement, design, and analysis: an integrated approach. Lawrence Erlbaum Assoc, Mahwah, NJ
Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton-Mifflin, Boston

Residuals

Samprit Chatterjee
Professor Emeritus of Statistics
Graduate School of Business Administration
Professor
New York University, New York, NY, USA

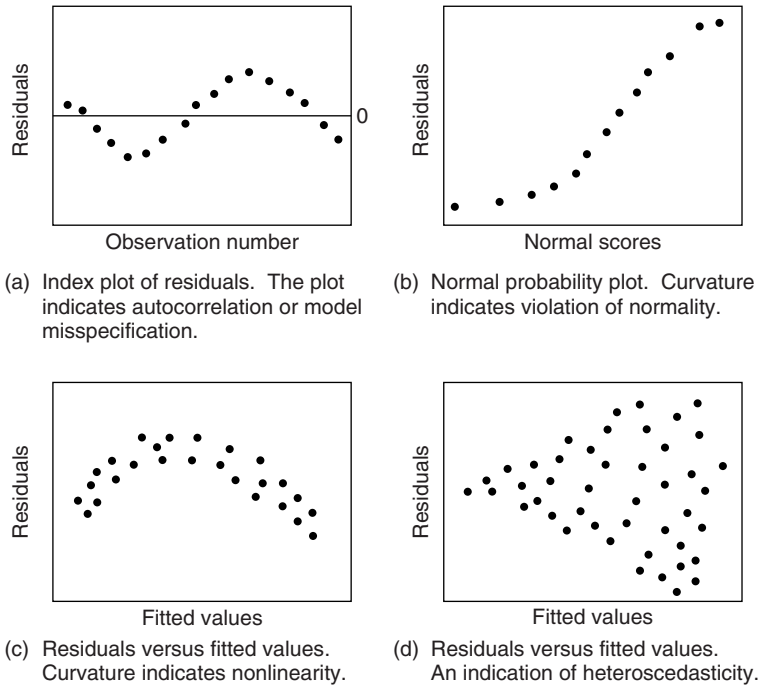
Residual is an important concept in statistical model building. Residual is defined as the difference between an observed value (Y) and the value fitted by a statistical model $\hat{Y}$

$$\text{Residual}_i = Y_i - \hat{Y}_i$$

A large value of the residual (positive or negative) shows the model does not fit the particular data point. The pattern of the residuals will often reveal the inadequacy of the fitted model. A plot of the residuals, often called the residual plot, is an important tool in regression model building. In this article we will concentrate on the role of residuals in regression analysis.

With n observations in a regression data set there will be n residuals. The sum of the n residuals from a least squares fit will be zero. Instead of working with the residuals as defined above we usually work with the standardized residuals. The residuals are scaled so they have unit standard deviation. We will call the standardized residuals for brevity residual. We will be talking about residuals obtained from least squares fit in our discussion.

The magnitude and the pattern of the distribution of residuals will reveal a great deal about the adequacy of the model describing the data. For moderate sized data the residuals can be thought of as normal deviates (mean 0, and Standard deviation 1). Large residuals, with values greater than 2 or 2.5, are called ►outliers. The data points



(a) Index plot of residuals. The plot indicates autocorrelation or model misspecification.

(b) Normal probability plot. Curvature indicates violation of normality.

(c) Residuals versus fitted values. Curvature indicates nonlinearity.

(d) Residuals versus fitted values. An indication of heteroscedasticity.

Residuals. Fig. 1 Several configurations illustrating possible violations of model assumptions

corresponding to the outliers are not well fitted by the model. These points should be examined carefully, as they often represent transcription errors or contamination of the data. By contamination of data we mean an observation which comes from another population. As an example consider a data set of weekly production of a factory. Most weeks have 5 workdays, but there may be few weeks with only 4 workdays. The weeks with 4 workdays will not be well fitted by the model, and those weeks will be outliers.

The residual plot should show no structure. The distribution should appear random. Instead of trying to describe random structure (an impossible task!) we provide examples of some commonly observed structures of residual plots, and indicate the model deficiencies they indicate.

The four graphs depicted in Fig. 1 give some of the commonly observed pattern of residual plots which indicate model deficiencies. Plot (c) indicates the data has a nonlinear component which is not included in the specified mode. Inclusion of a squared term in the model will remove the structure from the residuals.

The graph (a) indicates that the successive values are correlated, a common feature of time series data. This pattern may also arise if the model is misspecified. Methods for removing auto regression have to be adopted. Working with successive differences is a good first step.

The graph (d) indicates that the error variance is not constant and increases with size. The data is often classified as heteroscedastic. This is often referred to as the size effect. To account for the size effect, sometimes we work with logarithms of the data, use weighted least squares, or introduce a variable which reflects size.

The graph (b) is the normal plot of the residuals and is used to assess the normality of the residuals. This plot is not very effective for small sample sizes.

The residual plots are one of the most effective diagnostic plots for model fitting. No regression analysis is complete without a residual plot analysis.

About the Author

Samprit Chatterjee received his B.A. in 1958 from Calcutta University, and in 1967 he received a PhD from Harvard University. He was Research Assistant to Professor W.G. Cochran of Harvard University. He has been a visiting professor at Stanford, Sloan School of Management, Harvard School of Public Health, ETH (Zurich), University of Tampere, and University of Auckland (New Zealand). Since 1973, he has been Professor of Statistics, Graduate School of Business Administration, New York University. Professor Chatterjee has (co-)authored about 60 publications, including successful textbook *Regression*

Analysis by Example (with Ali Hadi, 4th edition, John Wiley & Sons, 2006).

Cross References

►Absolute Penalty Estimation
►Bootstrap Methods
►Gauss-Markov Theorem
►Generalized Linear Models
►Graphical Analysis of Variance
►Heteroscedasticity
►Influential Observations
►Jarque-Bera Test
►Least Absolute Residuals Procedure
►Linear Regression Models
►Outliers
►Regression Diagnostics
►Simple Linear Regression
►Structural Time Series Models
►Time Series Regression
►Vector Autoregressive Models

References and Further Reading

Chatterjee S, Hadi A (1988) Sensitivity analysis in linear regression. Wiley

Chatterjee S, Hadi A (2006) Regression analysis by example, 4th edn. Wiley

Response Surface Methodology

André I. Khuri
Professor Emeritus
University of Florida, Gainesville, FL, USA

Response surface methodology (RSM) is an area of statistics that incorporates the use of design and analysis of experiments along with model fitting of a response of interest denoted by y. One of the main objectives of RSM is the determination of operating conditions on a group of control (or input) variables that yield optimal response values over a certain region of interest denoted by $\mathcal{R}$.

In a typical response surface (RS) investigation, several factors are first identified by the experimenter as having possible effects on the response y. In some experiments, the number of such factors may be large. In this case, factor screening is carried out in order to eliminate factors deemed to be unimportant. This represents the first stage in the RS investigation. The execution of this stage requires the use of an initial design which consists of a number of specified settings of the control variables. Each set of such settings is used to produce a value on the response y. A low-degree polynomial model, usually chosen to be of the first degree, is then fitted to the resulting data set. Following factor screening, additional experiments are carried out which lead to a new region of experimentation where the actual exploration of the response will take place. By this we mean fitting a suitable polynomial model of degree higher than the one used in the initial screening stage. Such a model can be expressed as

$$y = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta} + \epsilon, \tag{1}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)'$ is a vector of k of control variables representing the levels of the factors that were retained after the initial screening, $\boldsymbol{f}(\boldsymbol{x})$ is a vector function of $\boldsymbol{x}$ whose elements consist of powers and cross products of powers of $x_1, x_2, \ldots, x_k$ up to a certain degree denoted by d. Typically, $d = 2$ or higher depending on the adequacy of model (1). Furthermore, $\boldsymbol{\beta}$ is a vector of p unknown parameters and ϵ is a random experimental error term assumed to have a zero mean. A commonly used form of model (1) is the second-degree model,

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \epsilon. \tag{2}$$

In this case, the elements of $\boldsymbol{\beta}$ consist of β_0, the β_i's, β_{ij}'s, and β_{ii}'s $(i, j = 1, 2, \ldots, k,\ i < j)$. The quantity $\boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta}$ in model (1) is called the mean response at $\boldsymbol{x}$ and is denoted by $\eta(\boldsymbol{x})$. Thus,

$$\eta(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\boldsymbol{\beta}. \tag{3}$$

In order to estimate the parameter vector $\boldsymbol{\beta}$, a series of n experiments is carried out in each of which the response y is measured at specified settings of $x_1, x_2, \ldots, x_k$. Let $\boldsymbol{x}_u = (x_{u1}, x_{u2}, \ldots, x_{uk})'$, where x_{ui} is the setting of x_i at the uth experimental run, and let y_u denote the corresponding response value $(i = 1, 2, \ldots, k,\ u = 1, 2, \ldots, n)$. From model (1) we then have

$$y_u = \boldsymbol{f}'(\boldsymbol{x}_u)\boldsymbol{\beta} + \epsilon_u,\ \ u = 1, 2, \ldots, n. \tag{4}$$

Model (4) can be expressed in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, $\boldsymbol{X}$ is a matrix of order $n \times p$ whose uth row is $\boldsymbol{f}'(\boldsymbol{x}_u)$, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$. The first

column of $\boldsymbol{X}$ is $\mathbf{1}_n$, the column of n ones. The $n \times k$ matrix,

$$\boldsymbol{D} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \tag{6}$$

whose rows consist of the settings of $x_1, x_2, \dots, x_k$ used at the n experimental runs is called the *design matrix*. If $\boldsymbol{\epsilon}$ is assumed to have a zero mean and a variance–covariance matrix $\sigma^2 \boldsymbol{I}_n$, where σ^2 is an unknown variance component and $\boldsymbol{I}_n$ is the matrix of ones of order $n \times n$, then $\boldsymbol{\beta}$ is estimated by the ordinary least-squares estimator,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{7}$$

The variance–covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2. \tag{8}$$

Using formula (3), an estimate of the mean response, $\eta(\boldsymbol{x})$, at a point $\boldsymbol{x}$ in the region of interest, $\mathcal{R}$, is given by

$$\hat{\eta}(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\beta}},$$

which is also known as the *predicted response* at $\boldsymbol{x}$ and is denoted by $\hat{y}(\boldsymbol{x})$. Thus,

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{f}'(\boldsymbol{x})\hat{\boldsymbol{\beta}}. \tag{9}$$

The variance of $\hat{y}(\boldsymbol{x})$ is then of the form

$$\mathrm{Var}[\,\hat{y}(\boldsymbol{x})] = \sigma^2 \boldsymbol{f}'(\boldsymbol{x})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{f}(\boldsymbol{x}). \tag{10}$$

This is called the *prediction variance*. The process of selecting the design matrix $\boldsymbol{D}$ and the subsequent fitting of model (1) represents the second stage of the RS investigation.

The third stage involves the determination of optimum operating conditions on the control variables, $x_1, x_2, \dots, x_k$, that yield either maximum or minimum values of $\hat{y}(\boldsymbol{x})$ over the region $\mathcal{R}$. This is a very important stage since it amounts to determining the settings of the control variables that should be used in order to obtain "best" values for the response. For example, if y represents the yield of some chemical product, and if the corresponding control variables consist of x_1 = reaction temperature and x_2 = length of time of the reaction, then it would be of interest to determine the settings of x_1 and x_2 that result in a maximum yield.

The proper choice of the design matrix $\boldsymbol{D}$ given in formula (6) is very important. This is true because $\boldsymbol{D}$ is used to predict the response and determine its prediction variance (see formula (10)). The size of the latter quantity has to be small in order to get good quality predictions. This is particularly true since the optimization of $\hat{y}(\boldsymbol{x})$ in formula (9) leads to the determination of optimum operating conditions on $x_1, x_2, \dots, x_k$ in the third stage of a RS investigation.

Several criteria are available for the choice of the design $\boldsymbol{D}$. Some of these criteria pertain to the prediction variance, such as *D-optimality* and *G-optimality*. A review of such criteria can be found in several textbooks such as Khuri and Cornell (1996, Chap. 12), Atkinson and Donev (1992, Chap. 10) and Myers and Montgomery (1995, Sect. 8.2.1), among others. Other design criteria deal with the minimization of the bias caused by fitting the wrong model as explained in Box and Draper (1959, 1963).

If model (1) is of the first degree, that is,

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \epsilon, \tag{11}$$

then common designs for fitting the model are 2^k factorial, Plackett-Burman, and simplex designs. These are referred to as *first-order designs*. A coverage of such designs can be found in, for example, Khuri and Cornell (1996, Chap. 3) and Myers and Montgomery (1995, Chaps. 3, 4, and 7). On the other hand, if model (1) is of the second degree, as shown in formula (2), then common second-order designs include 3^k factorial, the central composite design, and the Box-Behnken design. A coverage of these designs can be found in, for example, Khuri and Cornell (1996, Chap. 4), which also includes reference to other lesser-known second-order designs (see also Myers and Montgomery 1995, Sect. 7.4).

Methods for the determination of optimum conditions on the control variables depend on the nature of the fitted model in (1). If it is of the second degree, as in model (2), then the method of ridge analysis can be used to optimize $\hat{y}(\boldsymbol{x})$ in (9). This method was introduced by Hoerl (1959) and later formalized by Draper (1963) (see also Khuri and Cornell 1996, Chap. 5). A more recent modification of this method that takes into account the size of the prediction variance was given by Khuri and Myers (1979).

Historically, the development of RSM was initiated by the work of Box and Wilson (1951) which introduced the sequential approach in a RS investigation. The article by Box and Hunter (1957) is considered to be a key paper, which along with the one by Box and Wilson (1951), provided an outline of the basic principles of RSM. Several review articles were also written about RSM. These include those by Mead and Pike (1975), Myers

et al. (1989, 2004), and Myers (1999). In addition, a comprehensive coverage of RSM can be found in the books by Khuri and Cornell (1996), Myers and Montgomery (1995), and Box and Draper (2007).

New developments and modeling trends were introduced into the RSM literature in the late 1970s. They provided further extensions of the classical techniques used in RSM. Some of these developments include the *analysis of multiresponse experiments*, which deals with several response variables that are measured for each setting of a group of control variables (see Khuri 1996a), the *response surface approach to robust parameter design* (see, for example, Myers et al. 1992), *response surface models with random effects* (see Khuri 1996b, 2006). Furthermore, in the design area, several graphical techniques were introduced for comparing response surface designs. These include the use of *variance dispersion graphs*, as in Giovannitti-Jensen and Myers (1989), the *quantile plots* of the prediction variance, as in Khuri et al. (1996), and the *fraction of design space plots* by Zahran et al. (2003). The main advantage of the graphical approach is its ability to explore the prediction capability of a response surface design throughout the region of interest, $\mathcal{R}$. By contrast, standard design optimality criteria, such as *D*- or *G-optimality*, use single-valued criteria functions to evaluate a given design. This does not give adequate information about the design's performance at various locations inside the region $\mathcal{R}$.

About the Author

Dr. André I. Khuri is Professor Emeritus, Department of Statistics, University of Florida, Gainesville, Florida, USA. He is a Fellow of the American Statistical Association (since 1992) and an Elected Member of the International Statistical Institute (since 1989). He holds two PhDs, one in mathematics (University of Florida, 1969) and one in statistics (Virginia Tech, 1976). He has published more than 100 papers in statistics journals, in addition to 5 books, including *Response Surfaces* (1987, Dekker; 2nd edition, 1996) with John Cornell, *Statistical Tests for Mixed Linear Models* (1998, Wiley) with Thomas Mathew and Bimal Sinha, *Advanced Calculus with Applications in Statistics* (1993, Wiley; 2nd edition, 2003), *Response Surface Methodology and Related Topics* (2006, an edited book, World Scientific), and *Linear Model Methodology* (2009, Chapman & Hall/CRC). Professor Khuri was Associate Editor of *Technometrics* (1983–1992) and *Journal of Statistical Planning and Inference* (1995–2003). Currently, he is Editorial Advisor for *Journal of Probability and Statistical Science* (since 2003).

Cross References

►Optimal Designs for Estimating Slopes
►Optimum Experimental Design
►Statistical Design of Experiments (DOE)

References and Further Reading

Atkinson AC, Donev AN (1992) Optimun experimental designs. Oxford University Press, New York

Box GEP, Draper NR (1959) A basis for the selection of a response surface design. J Am Stat Assoc 54:622–654

Box GEP, Draper NR (1963) The choice of a second order rotatable design. Biometrika 50:335–352

Box GEP, Draper NR (2007) Response surfaces, mixtures, and ridge analyses, 2nd edn. Wiley, Hoboken

Box GEP, Hunter JS (1957) Multifactor experimental designs for exploring response surfaces. Ann Math Stat 28: 195–241

Box GEP, Wilson KB (1951) On the experimental attainment of optimum conditions (with discussion). J R Stat Soc B13: 1–45

Draper NR (1963) Ridge analysis of response surfaces. Technometrics 5:469–479

Giovannitti-Jensen A, Myers RH (1989) Graphical assessment of the prediction capability of response surface designs. Technometrics 31:159–171

Hoerl AE (1959) Optimum solution of many variables equations. Chem Eng Prog 55:69–78

Khuri AI (1996a) Multiresponse surface methodology. In: Ghosh S, Rao CR (eds) Handbook of statistics, vol 13. Elsevier Science B. V., Amsterdam, pp 377–406

Khuri AI (1996b) Response surface models with mixed effects. J Qual Technol 28:177–186

Khuri AI (2006) Mixed response surface models with heterogeneous within-block error variances. Technometrics 48: 206–218

Khuri AI, Cornell JA (1996) Response surfaces, 2nd edn. Dekker, New York

Khuri AI, Myers RH (1979) Modified ridge analysis. Technometrics 21:467–473

Khuri AI, Kim HJ, Um Y (1996) Quantile plots of the prediction variance for response surface designs. Comput Stat Data Anal 22:395–407

Mead R, Pike DJ (1975) A review of response surface methodology from a biometric viewpoint. Biometrics 31:803–851

Myers RH (1999) Response surface methodology – current status and future directions. J Qual Technol 31:30–44

Myers RH, Montgomery DC (1995) Response surface methodology. Wiley, New York

Myers RH, Khuri AI, Carter WH (1989) Response surface methodology: 1966–1988. Technometrics 31:137–157

Myers RH, Khuri AI, Vining GG (1992) Response surface alternatives to the Taguchi robust parameter design approach. Am Stat 46:131–139

Myers RH, Montgomery DC, Vining GG, Borror CM, Kowalski SM (2004) Response surface methodology: a retrospective and literature survey. J Qual Technol 36:53–77

Zahran A, Anderson-Cook CM, Myers RH (2003) Fraction of the design space to assess prediction capability of response surface designs. J Qual Technol 35:377–386

Ridge and Surrogate Ridge Regressions

Ali S. Hadi
Professor and Vice Provost
The American University in Cairo, Cairo, Egypt
Emeritus Professor
Cornell University, Ithaca, NY, USA

Ridge regression is a method for the estimation of the parameters of a linear regression model (see ▶Linear Regression Models) which is useful when the predictor variables are highly *collinear*, that is, when there is a strong linear relationship among the predictor variables. Hoerl (1959) named the method *ridge regression* because of its similarity to ridge analysis used in his earlier work to study second-order response surfaces in many variables. Some standard references for ridge regression are Hoerl and Kennard (1970, 1976), Belsley et al. (1980), and Chatterjee and Hadi (2006).

The standard linear regression model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations on the response variable, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is an $n \times p$ matrix of n observations on p predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors. It is usual to assume that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2\mathbf{I}_n$, where σ^2 is unknown constant and $\mathbf{I}_n$ is the identity matrix of order n.

Without loss of generality, we also assume that $\mathbf{Y}$ and the columns of $\mathbf{X}$ are centered and scaled to have unit length so that $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{Y}$ are matrices of correlation coefficients. If a variable $\mathbf{V}$ is not centered or scaled, it's i-th element, v_i, can be replaced by $(v_i - \bar{v})/\sqrt{\sum_{i=1}^n (v_i - \bar{v})^2}$.

An estimate for $\boldsymbol{\beta}$ is obtained by minimizing

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{2}$$

where $\|\cdot\|$ denotes the Euclidean (or L_2) norm. The minimization of this ordinary least squares (OLS) problem leads to the so-called system of normal equations,

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}. \tag{3}$$

Provided that $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ exists, the solution of this system of linear equations is given by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{4}$$

with $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

If collinearity is present, the linear system in (3) is said to be *ill-conditioned* and $\hat{\boldsymbol{\beta}}$ in (4) can be unstable, that is, a slight change in the data can result in a substantial change in the values of the estimated regression coefficients. Furthermore, collinearity usually inflates the variance of $\hat{\boldsymbol{\beta}}$ and this in turn deflates the t-statistics for testing the significance of the regression coefficients, which can lead to the wrong conclusion that the coefficients of some important predictors are statistically insignificant.

A measure for assessing the condition of the linear system in (3) is the condition number of $\mathbf{X}$, which is defined as $\kappa = \sqrt{\lambda_1/\lambda_p}$, where λ_1 and λ_p are the largest and smallest eigenvalues of $\mathbf{X}^T\mathbf{X}$, respectively. Large values of the condition number indicate ill-conditioned system. A measure for assessing the effect of collinearity on variance inflation is the *variance inflation factor* (VIF). For the j-th predictor variable X_j, the VIF is the j-th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. It can be shown that $\mathrm{VIF}_j = 1/\left(1 - R_j^2\right)$, where R_j^2 is the multiple correlation coefficient when X_j is regressed on all other predictor variables. When X_j has a strong linear relationship with all other predictors, R_j^2 would be very large (close to 1), causing VIF_j to be very large. As a rule of thumb, values of variance inflation factors greater than 10 are indicative of the presence of collinearity.

To obtain a stable (*regularized*) solution, we replace the problem in (2) by minimizing

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + k\|\boldsymbol{\beta}\|^2, \tag{5}$$

for some value of $k > 0$, suitably chosen by the user. The explicit solution of the problem in (5) is

$$\hat{\boldsymbol{\beta}}(k) = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}, \tag{6}$$

where $\mathbf{I}_p$ is the identity matrix of order p. The expected value and variance of $\hat{\boldsymbol{\beta}}(k)$ are

$$E(\hat{\boldsymbol{\beta}}(k)) = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \tag{7}$$

and

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + k\,\mathbf{I}_p)^{-1}. \tag{8}$$

In statistics, the solution $\hat{\boldsymbol{\beta}}(k)$ is known as *ridge regression* (see Hoerl 1962) and the ridge parameter k is a penalizing factor. But more generally, it is known as the Tikhonov regularization (TR) method (Tikhonov 1943) and the factor k is known as the Tikhonov factor.

By comparing (4) and (6), one can see that the ridge estimator is obtained by adding a small positive quantity k to each of the diagonal elements of the matrix $\mathbf{X}^T\mathbf{X}$. Clearly, when $k = 0$, the ridge estimator in (6) becomes the OLS estimator in (4). It is clear from (7) that for $k > 0$, ridge estimators are biased for $\boldsymbol{\beta}$.

The variance inflation factors as a function of k are the diagonal elements of the matrix $(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}$. Hoerl and Kennard (1970) show that there exists a value of $k > 0$ such that

$$E[(\hat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta})] < E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})], \quad (9)$$

which means that the total mean square error of the ridge estimators are less than that of the OLS.

The choice of the ridge parameter k is therefore important. The optimal value of k is difficult to find, but there exists several alternative methods for estimating k. First, an appropriate value of k can be found graphically by examining the *ridge trace*, which is a simultaneous plot of the elements of $\hat{\boldsymbol{\beta}}(k)$ versus k (usually between 0 and 1). The smallest value of k, for which (a) the estimated vector of regression coefficients, $\hat{\boldsymbol{\beta}}(k)$, is stable, (b) the variance inflation factors are less than 10 (close to 1), and (c) the residual sum of squares is close to its minimum value, is chosen and used in (6) to obtain the ridge estimators.

Second, numerical methods for estimating k have been proposed. For example, Hoerl et al. (1975) suggest estimating k by $\hat{k} = p\hat{\sigma}^2/(\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}})$, where $\hat{\sigma}^2 = SSE/(n-p)$ and SSE is the OLS residual sum of squares. Other numerical methods have also been suggested; see, for example, Lawless and Wang (1976), Wahba et al. (1979), Hoerl and Kennard (1981), Masuo (1988), Khalaf and Shukur (2005), and Dorugade and Kashid (2010).

Forms of ridge regression other than (6) are possible. For example the ridge parameter k (which is a scalar) can be replaced by a diagonal matrix with possibly different diagonal elements, or even with a full $p \times p$ matrix, but these alternatives are less common in practice.

More recently, Jensen and Ramirez (2008) cast some doubt about the ability of ridge estimators to actually improve the condition of an ill-conditioned linear system and provide stable estimated regression coefficients and smaller variance inflation factors. Note that the ridge estimator in (6) is the solution of the linear system

$$(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}, \quad (10)$$

which replaces the system of normal equations in (3). The condition number of the matrix $(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)$ on the left-hand side of (10) is $\sqrt{(\lambda_1 + k)/(\lambda_p + k)}$, which is smaller than the condition number of $(\mathbf{X}^T\mathbf{X})$, which is $\sqrt{\lambda_1/\lambda_p}$. Thus adding k to each of the diagonal elements of $\mathbf{X}^T\mathbf{X}$ improves its condition. But the matrix $\mathbf{X}$ on the right-hand side of (10) remains ill-conditioned. To also improve the condition of the right-hand side of (10), Jensen and Ramirez (2008) propose replacing the ill-conditioned regression model in (1) by the *surrogate* but less ill-conditioned model

$$\mathbf{Y} = \mathbf{X}_k\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (11)$$

where $\mathbf{X}_k = \mathbf{U}(\boldsymbol{\Lambda} + k\mathbf{I}_p)^{1/2}\mathbf{V}^T$, the matrices $\mathbf{U}$ and $\mathbf{V}$ are obtained from the *singular-value decomposition* of $\mathbf{X} = \mathbf{UDV}^T$ (see, e.g., Golub and van Loan 1989) with $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$, and $\mathbf{D}$ is a diagonal matrix containing the corresponding ordered singular values of $\mathbf{X}$. Note that the square of the singular values of $\mathbf{X}$ are the eigenvalues of $\mathbf{X}^T\mathbf{X}$, that is, $\mathbf{D}^2 = \boldsymbol{\Lambda}$. Because $\mathbf{X}_k^T\mathbf{X}_k = \mathbf{X}^T\mathbf{X} + k\mathbf{I}_p$, the least squares estimator of the regression coefficients in (11) is the solution of the linear system

$$(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)\boldsymbol{\beta} = \mathbf{X}_k^T\mathbf{Y}, \quad (12)$$

which is given by

$$\hat{\boldsymbol{\beta}}_s(k) = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}_k^T\mathbf{Y}. \quad (13)$$

Jensen and Ramirez (2008) study the properties of the surrogate ridge regression estimator, $\hat{\boldsymbol{\beta}}_s(k)$, in (13), and using a case study they demonstrate that the surrogate estimator is more conditioned than the classical ridge estimator, $\hat{\boldsymbol{\beta}}(k)$, in (6). For example, they observe that (a) the condition of the variance of $||\hat{\boldsymbol{\beta}}_s(k)||$ is monotonically increasing in k and (b) the maximum variance inflation factor is monotonically decreasing in k. These properties do not hold for the classical ridge estimator, $\hat{\boldsymbol{\beta}}(k)$.

About the Author

Professor Hadi is the Vice Provost and the Director of Graduate Studies and Research, the American University in Cairo (AUC). He is also a Stephen H. Weiss Presidential Fellow and Professor Emeritus, Cornell University, USA. He is the Founding Director of the Actuarial Science Program at AUC (2004–present). Dr. Hadi is the Editor-in-Chief, *International Statistical Review* (2009–present) and Co-founding Editor, *Journal of Economic and Social Research* (1998–present). He is an Elected Fellow of the American Statistical Association (1997) and Elected Member of the International Statistical Institute (1998). He has authored/coauthored nearly 100 articles in international refereed journals, and has published 5 books including, *Regression Analysis by Example* (with Samprit Chatterjee, Wiley, 4th edition, 2006).

Cross References

►Absolute Penalty Estimation
►Linear Regression Models
►Multicollinearity
►Multivariate Statistical Analysis
►Partial Least Squares Regression Versus Other Methods
►Properties of Estimators

References and Further Reading

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York

Chatterjee S, Hadi AS (2006) Regression analysis by example, 4th edn. Wiley, New York

Dorugade AV, Kashid DN (2010) Alternative method for choosing ridge parameter for regression. Appl Math Sci 4:447–456

Golub GH, van Loan C (1989) Matrix computations. John Hopkins, Baltimore

Hoerl AE (1959) Optimum solution of many variables. Chem Eng Prog 55:69–78

Hoerl AE (1962) Application of ridge analysis to regression problems. Chem Eng Prog 58:54–59

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–68

Hoerl AE, Kennard RW (1976) Ridge regression: iterative estimation of the biasing parameter. Commun Stat, Theory Methods A5:77–88

Hoerl AE, Kennard RW (1981) Ridge regression – 1980: advances, algorithms, and applications. Am J Math Manag Sci 1:5–83

Hoerl AE, Kennard RW, Baldwin KF (1975) Ridge regression: some simulations. Commun Stat, Theory Methods 4:105–123

Jensen DR, Ramirez DE (2008) Anomalies in the foundations of ridge regression. Int Stat Rev 76:89–105

Khalaf G, Shukur G (2005) Choosing ridge parameter for regression problem. Commun Stat, Theory Methods 34:1177–1182

Lawless JF, Wang P (1976) A simulation study of ridge and other regression estimators. Commun Stat, Theory Methods 14:1589–1604

Masuo N (1988) On the almost unbiased ridge regression estimation. Commun Stat, Simul 17:729–743

Tikhonov AN (1943) On the stability problems. Dokl Akad Nauk SSSR 39:195–198

Wahba G, Golub GH, Health CG (1979) Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21:215–223

Rise of Statistics in the Twenty First Century

Jon R. Kettenring
Drew University, Madison, NJ, USA

Introduction

As the end of the first decade of the twenty first century approaches, it is fair to say that statistics as a profession is on the rebound in ways that matter. In recent years there have been recurring laments about missed opportunities and lack of respect for statistics and statisticians. Yet, statisticians are on the move, identifying and embracing new opportunities, and being increasingly expansive in defining the scope of their field and its relationship to the world at large.

A sign of progress is the growing recognition by outsiders of the importance of the statistics discipline. For example, a recent editorial in *Science* magazine (Long and Alpern 2009), based on a new report, "Scientific Foundations for Future Physicians," argues that "students should arrive at medical school prepared in the sciences, including some areas not currently required, such as statistics and biochemistry."

Even more recently, this dramatic headline appeared on the front page of *The New York Times*: "For today's graduate, just one word: statistics." The accompanying article (Lohr 2009) talked about "the rising stature of statisticians," "the new breed of statisticians" who analyze "vast troves of data," and how the "data explosion" is "open[ing] up new frontiers" for statistics.

The goal in this essay is to discuss a few disparate factors that are relevant to my thesis about the current rise of and strong future for statistics and statisticians.

Renewal

One of the great qualities of the statistics profession, and an important reason for optimism about the future, is its tradition of introspection, self-assessment, and adjustment, leading to renewal of how we view our field, how we teach our subject, and how we interact with others. In part, this reflects the healthy questioning that statisticians engage in whenever confronted with a new problem. We want to know the whole story. We are politely but usefully skeptical of everything we are told. We examine all assumptions. We strive for quality data and supportable conclusions drawn from sound analyses. We know how to unlock underlying truths from complex and noisy circumstances. We recognize our limitations but also are able to develop new methods as needed.

Education

►Statistics education has received considerable attention and improvement during the last 50 years. It is even an accepted area for research – and we've learned a lot about how to teach students more effectively. Statistics courses are now frequently found in high schools in the U.S. Over 100,000 high school students take the Advanced Placement Examination in Statistics annually. Introductory courses for college students are often taught by specialists who have devoted careers to perfecting ways of breaking novices gently and carefully into the wonders of statistical thinking. The wide availability of statistical software has allowed students to experience first hand what it is like to analyze real data.

Consulting courses have been a popular way to expose graduate students to current data analysis and modeling

problems. Going a step further, there are now success stories that involve the orchestrated merging of statistics education, consulting, and research. A well-developed example of this synergistic approach, from the University of California at Riverside, is described by Jeske et al. (2007). As a result of such efforts, students are entering the workforce with wider experiences and broader skill sets.

Another plus has been the evolution of postdoctoral programs in statistics. Once rarities, they are increasingly common in academia, government, industry and research institutes such as the National Institute of Statistical Sciences (NISS) in the U.S.A. In the past 18 years, NISS has engaged more than 60 postdoctoral students in cross-disciplinary research projects.

Still, the potential for *substantial* improvements remains, and a few specific ones are spelled out in Lindsay et al. (2004). The recent eye-catching proposal by Brown and Kass (2009), based on their experience working together in neuroscience, calls for strong reforms based on "deepening cross-disciplinary involvement" and "a broad vision of the discipline of statistics." The aggressive changes that they propose amount to a culture change (comment by Gibbs and Reid in the discussion) and won't come easily (comment by Johnstone) but are necessary to keep up with "big science" (comment by Nolan and Temple Lang).

Cross-Disciplinary Research

Statistics has always been driven in part by applications. It is only in the last few decades that full-blown cross-disciplinary research involving statistics has become widely accepted within the profession. Now it is not only part of the culture but it is also spurring growth as statisticians respond to new data problems such as those posed by neuroscience. Other fruitful fields are easy to tick off as well, e.g., ▶bioinformatics, healthcare, life sciences, climate change, the environment, manufacturing, business strategy, privacy and confidentiality, bioterrorism, and national defense. Increasingly, the associated problems involve the analysis of massive datasets, i.e., ones of extraordinary size and complexity. When confronted with such challenges, teamwork is perhaps the most effective strategy and is very likely to trump purely statistical ones (Kettenring 2009).

A variety of other examples of cross-disciplinary work are listed in Lindsay et al. (2004), under the heading of statistics in science and industry, to illustrate the now common "interplay between statistics and other scientific disciplines." It is also worth noting that funding opportunities for such crossover activities have been on the rise.

Credentials

Within the American Statistical Association (ASA) there have been strong debates for at least 15 years about the need for credentialing (as in accreditation based on experience or certification based on testing) of professional statisticians. In a perfect world there would be no need for more than a suitable academic degree, but ours is not so neat and tidy. Practitioners with no degrees in statistics but excellent records of accomplishment often have difficulty achieving the stature that they deserve or require for success in their careers. Similar problems are encountered by those who work in fringe areas or for small employers where the professionalism of statisticians is misunderstood or underappreciated. We also face the unfortunate companion situation of practitioners who claim competency in statistics but lack it. This can result in damaging malpractice.

Similar concerns were no doubt behind earlier movements by the Royal Statistical Society, the Statistical Society of Canada, and the Statistical Society of Australia Inc. to provide accreditation programs that help to differentiate practitioners who are good at statistics from those who only claim to be. Motivated in part by the apparent success of these ventures, the ASA has taken several steps along the same path, culminating in approval in August 2009 of a plan (Bock et al. 2009) to launch its own *voluntary* accreditation program for professional statisticians.

In a randomized survey of ASA members, 41% reported that they would apply to such a program were it offered because it would provide evidence of competency and a credential useful for employment, among other factors. This change in thinking from "why do we need such a thing" to "maybe the time is right" illustrates that the field is evolving not only in theoretical and philosophical ways but also in very pragmatic ones aimed at meeting the needs of practitioners operating in very competitive environments.

Time will tell, but if these credentialing programs fulfill their potential, they will have served a very useful purpose by helping to legitimize and support a broad class of professionals who are highly qualified to practice statistics.

Journals

Journals are an essential component of the statistics infrastructure. They serve as a collective record of historical and current developments in the field. JSTOR, e.g., provides a very important centralized archive and point of access for more than 40 of the leading ones in statistics and probability. It includes journals such as *Biometrika*, the *Journal of the American Statistical Association (JASA)*, the *Journals of the Royal Statistical Society (JRSS)*, and *Technometrics*. Yet there are many more. The Current Index

to Statistics lists over 160 "core" journals on its website, www.statindex.org/CIS/news/CIS_core_journals.pdf. As of April 2009 the website, http://w4.stern.nyu.edu/ioms/research.cfm?doc_id=3532, hosted by the Stern School at New York University, included over 200 names under the heading of statistics and probability. Its breadth reflects the soft boundary view of statistics. Examples include *Analytical Chemistry, Econometrica*, the *Journal of Machine Learning Research*, and *Water Resources Research*.

Comparing journals across fields is dicey business, but it is tempting to see where one stands. In Lindsay et al. (2004) it was observed that *JASA* was "far and away the most cited mathematical science journal" for the period 1991–2001. I've also taken note of more recent data on the website www.eigenfactor.org, where 68 journals in probability and statistics are ranked based on cross-journal citation patterns, along with nearly 8,000 others. Citations of a journal to itself are not counted. The journals are quantified by their "Eigenfactor Score," which measures "the journal's total importance to the scientific community", and the "Article influence Score," which measures "the average influence, per article, of the papers in a journal" (Bergstrom et al. 2008). For the most recent year available, 2007, based on citations to the previous five years, the top four in the probability and statistics category are *JASA*, *Statistics in Medicine*, the *Annals of Statistics*, and *JRSS Series B* by the first measure, and *JRSS Series B*, *JASA*, the *Annals of Statistics*, and *Biostatistics* by the second. (The last title is a convenient reminder of the enormous success story of ►biostatistics as a subfield that has led the way in growth and career opportunities.)

The median Article Influence Score is 0.45 across all journals and 0.70 for those in the probability and statistics category. In comparison, the medians are 0.72 for neuroscience, 0.62 for mathematics, 0.60 for psychology, 0.58 for economics, and 0.59 for physics. The point is that statistics journals are publishing articles of relatively broad interest and high influence, at least by this measure. The vitality of our better journals provides a strong backbone for the future of statistics research, practice, and education.

Holistic Statistics

In Kettenring (1997), under the heading of "holistic statistics," I asked whether statistics in the twenty first century should be equated so strongly to the more traditional core topics of statistics as it had been in the past and followed with these points:

- There is a natural tension between narrowly focused pursuits of science vs. broader ones that favor synthesis and interdisciplinarity.
- The core of statistics should be nourished by surrounding itself with vigorous areas of application.
- Broad-minded statisticians are needed to work across boundaries and operate in fast-paced environments.
- A more inclusive definition of statistics would better reflect its strong interdisciplinary character.
- Such an inclusive interpretation of statistics is where the future lies.

In similar spirit, Hand (2009) talks about the importance of "greater statistics" (Chambers 1993) as an overarching discipline that deals with (quoting Chambers) "everything related to learning from data." It is this expansive view of statistics that I intended in the title of this essay and what Hand has in mind when he talks about the "magic" of modern statistics.

Wrap Up

Taking a bit of license with a popular adage, we can safely say that the future of statistics isn't what it used to be. These are meant to be encouraging words for students looking for a field of study that is full of life and much needed in a modern information age that is swamped with data and in need of help on what to do about it. Or, as the distinguished economist Hal Varian put it in Lohr (2009), "I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding."

Acknowledgments

The author was President of the American Statistical Association in 1997. He thanks Daniel Jeske, Alan Karr, and David Morganstein for their suggestions.

About the Author

Dr. Jon R. Kettenring is Past President of the American Statistical Association (1997). He has been Director of RISE since 2008. Previously, he was Executive Director of the Mathematical Sciences Research Center at Telcordia Technologies. Throughout his career, he has maintained a strong interest in statistics research and its application to solve real problems in the telecommunications industry. He is a Fellow of ASA and AAAS and an Elected Member of the International Statistical Institute. The National Institute of Statistical Sciences sponsored The Future of Data Analysis Conference in his honor at Avaya Labs Research, Basking Ridge, NJ (September 30, 2005).

Cross References

►Careers in Statistics
►Online Statistics Education
►Role of Statistics

►Role of Statistics in Advancing Quantitative Education
►Role of Statistics: Developing Country Perspective
►Statistics Education
►Statistics: An Overview

References and Further Reading

Bergstrom CT, West JD, Wiseman MA (2008) The eigenfactor metrics. J Neurosci 28:11433–11434
Bock ME, Hoerl R, Kettenring J, Kirkendall N, Mason R, Morganstein D, Nair V, O'Neill R, Oppenheimer L, Wasserstein R (2009) Report to the ASA board of directors by the individual accreditation proposal review group. www.amstat.org/news/pdfs/Kettenring_AccreditationReport.pdf
Brown EM, Kass RE (2009) What is statistics? Am Stat 63: 105–123
Chambers JM (1993) Greater or lesser statistics: a choice for future research. Stat Comput 3:182–184
Hand DJ (2009) Modern statistics: the myth and the magic (with discussion). J R Stat Soc A 172:287–306
Jeske DR, Lesch SM, Deng H (2007) The merging of statistics education, consulting and research: a case study. J Statist Educ 15. www.amstat.org/publications/jse/v16n3/jeske.html
Kettenring JR (1997) Shaping statistics for success in the 21st century. J Am Stat Assoc 92:1229–1234
Kettenring JR (2009) Massive datasets. WIRES Comput Stat 1:25–32
Lindsay BG, Kettenring JR, Siegmund DO (2004) A report on the future of statistics (with discussion). Stat Sci 19:387–412
Lohr S (2009) For today's graduate, just one word: statistics. The New York Times A1 and A3
Long S, Alpern R (2009) Science for future physicians. Science 324:1241

Risk Analysis

Michael R. Greenberg
Professor
Rutgers University, New Brunswick, NJ, USA

Introduction

Risk analysis originated in safety and systems engineering and following the events at the Three Mile Island nuclear facilities in the United States developed into an interdisciplinary approach to better understand and manage hazards. It has been applied to many human and ecological health issues, such as air borne spread of biological agents, destruction of the United States chemical weapons stockpile, cyber attacks, facility safety, food contamination, hazardous waste management, medical decision-making, nuclear power and waste management, and natural hazards such as earthquakes, floods, and tornadoes. Several of the ideas and models can be extended to economic, social, and even political risk.

Risk analysis is divided into risk assessment and risk management, although feedback loops exist among the stages. To provide continuity to this entry, the author uses the example of a terrorist planning to kill bus riders.

Risk Assessment

Risk assessors try to answer three questions.

1. What can go wrong?
2. What are the chances that something with serious consequences will go wrong?
3. What are the consequences if something does go wrong?

This so-called "triplet" of questions (Kaplan and Garrick 1981; Garrick 1984) can be written as follows:

$$R = (S, P, C)$$

where R is the risk; S is a hypothesized risk scenario event of what can go wrong; P is the probability of that scenario occurring; and C are the consequences.

Scenarios

Analysts create risk scenarios. For example, the terrorist could board a bus and detonate a bomb or leave a bomb near a stop and detonate it remotely. When there are thousands of potential triggering scenarios, analysts identify the worst consequences and then they work backward to scenarios that could produce them.

Analysts use fault trees or event trees to build out risk assessment scenarios. Event trees start with an event and follow it through branches. Some of the branches lead to insignificant consequences, while others end in serious outcomes. Fault tree analyses begin with the end state and work backwards to identify the event or sequence of events that will trigger it. They also are used together.

Likelihood

Quantification of the ►likelihood of events was the major improvement introduced by risk assessment to safety analysis. Analysts develop a probability distribution of the likelihood of events, and then apply them to the trees. This step, the author believes, is the most difficult challenge faced by risk assessors (Hora 2007; Aven and Renn 2009; Committee on Methodological Improvements to the Department of Homeland Securities Biological Agent Risk Analysis, National Research Council 2007; Cox 2009; Dillon et al. 2009).

When there is no deliberate intent driving an event, for instance, a bus has a flat tire and crashes, then estimating probabilities from historical records and experiences make sense. The problem is how to estimate the likelihood

of a deliberate attack. Do we assume that the terrorists are intent as well as capable and will adjust to countermeasures? If a terrorist is assumed to achieve optimal or near optimal success, then we need to game theoretic or more generally agent-based modeling to support these estimates.

Eliciting likelihood estimates from experts is part science and part art. Expert opinion can be obtained through surveys. The results may not translate directly into an absolute measurement of likelihood, but perhaps an ordinal scale that can be used to set priorities. The key is training the experts to participate in the survey (Hora 2007; Aven and Renn 2009; Committee on Methodological Improvements to the Department of Homeland Securities Biological Agent Risk Analysis, National Research Council 2007).

Consequences

Estimating consequences is part three of the risk assessment process. In the case of the bus passengers, the worst outcomes would be death of all the riders and for a transit system it would also be public fear of using the bus. When all the branches are followed, the results include estimates of the number of deaths, injuries, physical damage to assets, environmental effects, and local/regional economic impacts. These consequences are ranked with regard to severity.

Summarizing, risk assessments produce a list of risks. What to do about them is the responsibility of risk managers.

Risk Management

The author offers the following three questions for risk managers:

1. How can consequences be prevented or reduced?
2. How can recovery be enhanced, if the scenario occurs?
3. How can key local officials, expert staff, and the public be informed to reduce concern and increase trust and confidence?

Prevention

Risk managers try to contain risk within an acceptable (typically regulatory-based) level by implementing mitigation measures. These options are engineered and behavioral, and collectively they can be seen as the essence of a multi-criteria decision-making model (MCDM) (Chankong and Haimes 2008), where prevention options $O_i(i = 1, 2, \ldots, m)$ are defined and the decision-maker(s) evaluate them based on selected criteria $C_j(j = 1, 2, \ldots, n)$. In the case of our at risk bus passengers, the company can monitor the bus stops, and train the drivers can look for suspicious behavior (they are already trained to deal with rowdy and intoxicated riders).

Prevention is partly based on engineering options, life cycle cost, union contracts, ethical and other considerations, and requires blending of quantitative and qualitative data into a multi-criteria decision making framework. The audience for this volume recognizes the problems of scarcity of data, lack of knowledge, and subjectivity. Various mathematical and statistical techniques are available in the literature to deal with decision-making with uncertainties (Chankong and Haimes 2008; Heinz Center 2000; Skidmore and Toya 2002; Greenberg et al. 2007; Bedford and Cooke 2001; Edwards et al. 2007).

Two difficult challenges for risk managers are the time and space dimensions. In fact, we do not know a great deal about the geographical and temporal impacts of risk-related events (Zinn 2009). The direct consequences of a bus explosion event include impacts on the passengers and the driver. But indirect effects could include fear of using the bus, leading to a loss of revenue for the system and for businesses that depend on it, and to induced income effects caused by job losses. That is, when people lose their jobs, they begin to reduce their purchases.

The local impact is the area directly impacted by the event. Regional impacts occur in surrounding areas that are affected by direct losses. State, national and international impacts are felt as economic consequence ripples across the landscape. Some of these impacts are felt immediately or within a month or two of the event. Others are intermediate in length and measured in months and even a year or two out from the events. If the event is large enough there will also be long-term impacts that can be measured for many years. For example, the author has studied the impact of large scale loss of energy supply, leading to loss of confidence in the region and relocation out of the area. Yet we also have learned that a devastated region will receive funds from insurance companies, not-for-profits, and government agencies (Singpurwalla 2006), so negative consequences may be less than had been anticipated.

Economic impact tools allow us to estimate some of the consequences of such risk management decisions, albeit each of these has important data requirements, limitations, and capabilities (Modarres et al. 2010). The key to successfully using the economic models is the willingness of analysts to probe deeply into the events and through the stages that follow. Using sophisticated models without first understanding the event is a waste of time and money.

Recovery

Even the best efforts to prevent risk events sometimes fail. Every risk manager needs a plan to respond to events. Like

the risk mitigation response, these options are both human and engineered. The response options $R_i(i = 1, 2, \ldots, m)$ are defined and the decision-maker(s) evaluate them based on selected criteria $C_j(j = 1, 2, \ldots, n)$. In the case of our bus passengers, for example, police would cordon off the area, ambulances would arrive, and the injured would be moved to a nearest health care facility that is able to treat those that are alive.

Communications

At some major risk-related events, there are misunderstandings about who is in charge, who should perform what function, and sometimes the results are tragic. Firemen and police should be fully equipped and trained to deal with hazards. Transit workers and upper-level managers should know how to cope with passengers who show signs of suspicious behavior, and how to prevent panic rather than contribute to it. A good deal of research is focusing on crisis communications specifically and risk communications more generally, and principles have been articulated about how to manage risk events. However, it will take a systematic and ongoing effort to diffuse these suggestions to managers and to front-line employees.

Summary

Risk analysis is a multidisciplinary field that includes researchers trained in physics, chemistry, biology, engineering, mathematics, economics, psychology, geography, sociology, communications, political science and others. This essay touches on the six key questions that risk analysts try to answer. Some good books are available (Bedford and Cooke 2001; Edwards et al. 2007; Zinn 2009; Singpurwalla 2006; Modarres et al. 2010). However, in a rapidly moving field like this, most books are out of date quickly. The author recommends consulting two journals: *Risk Analysis: an International Journal* and the *Journal of Risk Research*.

Acknowledgments

I would like to thank colleagues of many years for helping me learn about risk analysis, especially Vicky Bier, Tony Cox, John Garrick, Bernard Goldstein, Ortwin Renn, Paul Slovic, and Yacov Haimes.

About the Author

Dr. Michael R. Greenberg is Professor and Associate Dean of the Faculty of the Edward J. Bloustein School of Planning and Public Policy of Rutgers University. He is Director, National Center for Neighborhood and Brownfields, and Director, Rutgers National Transportation Security Center of Excellence. He has been a member of National Research Council Committees that focus on waste management, such as the destruction of the U.S. chemical weapons stockpile and nuclear weapons. Professor Greenberg has contributed more than 500 publications to scientific journals like *Cancer Research*, *American Journal of Epidemiology*, *Risk Analysis*, *American Journal of Public Health*, and public interest ones like *Urban Affairs Review*, *Housing Policy Debate*, *Society*, *the Sciences*, and *Public Interest*. He has received awards for research from the United States Environmental Protection Agency, the Society for Professional Journalists, the Public Health Association, the Association of American Geographers, and Society for Risk Analysis. In 2003, he received the Distinguished Career Achievement Award, International Society for Risk Analysis. He has supervised about 70 PhD students. Currently, Professor Greenberg is the Editor-in-Chief, *Risk Analysis: An International Journal*.

Cross References

►Actuarial Methods
►Banking, Statistics in
►Bias Analysis
►Insurance, Statistics in
►Likelihood
►Quantitative Risk Management
►Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts

References and Further Reading

Aven T, Renn O (2009) The role of quantitative risk assessment for characterizing risk and uncertainty and delineating appropriate risk management options, special emphasis on terrorism risk. Risk Anal 29(4):587–599

Bedford T, Cooke R (2001) Probabilistic risk analysis: foundations and methods. Cambridge University Press, Cambridge, UK

Chankong V, Haimes Y (2008) Multiobjective decision making: theory and methodology. Dover, New York

Committee on Methodological Improvements to the Department of Homeland Securities Biological Agent Risk Analysis, National Research Council (2007) Interim report on methodological improvements to the Department of Homeland Security's biological agent risk analysis. National Academy, Washington, DC

Cox LA Jr (2009) Improving risk-based decision-making for terrorism applications. Risk Anal 29(3):336–341

Dillon R, Liebe R, Bestafka T (2009) Risk-based decision-making for terrorism applications. Risk Anal 29(3):321–335

Edwards W, Miles R Jr, von Winterfeldt D (2007) Advances in Decision Analysis. Cambridge University Press, Cambridge, UK

Garrick BJ (1984) Recent case studies and advances in probabilistic risk assessments. Risk Anal 4:262–279

Greenberg M, Lahr M, Mantell N, Felder N (2007) Understanding the economic costs and benefits of catastrophes and their aftermath: a review and suggestions for the as-federal government. Risk Anal 27(1):83–96

Heinz Center for Science, Economics and the Environment (2000) The hidden costs of coastal hazards: implications for risk assessment and mitigation. Island Press, Washington, DC
Hora S (2007) Eliciting probabilities from experts. In: Edwards W, Miles R, von Winterfeldt D (eds) Advances in decision analysis. Cambridge University Press, Cambridge, UK, pp 129–153
Kaplan S, Garrick BJ (1981) On the quantitative definition of risk. Risk Anal 1(1):11–27
Modarres M, Kaminskiy M, Krivtsov V (2010) Reliability Engineering and Risk Analysis. Taylor & Francis Group, Boca Raton, Florida, FL
Singpurwalla N (2006) Reliability and risk. Wiley, New Jersey
Skidmore M, Toya H (2002) Do natural disasters promote long-run growth? Economic Inquiry 40:664–687
Zinn J (2009) Social Theories of Risk and Uncertainty: An Introduction. Blackwell, Oxford, UK

Robust Inference

Elvezio Ronchetti
Professor
University of Geneva, Geneva, Switzerland

►Robust statistics deals with deviations from ideal parametric models and their dangers for the statistical procedures derived under the assumed model. Its primary goal is the development of procedures which are still reliable and reasonably efficient under small deviations from the model, i.e., when the underlying distribution lies in a neighborhood of the assumed model. Robust statistics is then an extension of parametric statistics, taking into account that parametric models are at best only approximations to reality. The field is now some 50 years old. Indeed one can consider Tukey (1960), Huber (1964), and Hampel (1968) the fundamental papers which laid the foundations of modern robust statistics. Book-length expositions can be found in Huber (1981, 2nd edition by Huber and Ronchetti 2009), Hampel et al. (1986), Maronna et al. (2006).

More specifically, in robust testing one would like the level of a test to be stable under small, arbitrary departures from the distribution at the null hypothesis (*robustness of validity*). Moreover, the test should still have good power under small arbitrary departures from specified alternatives (*robustness of efficiency*). For confidence intervals, these criteria correspond to stable coverage probability and length of the confidence interval.

Many classical tests do not satisfy these criteria. An extreme case of nonrobustness is the F-test for comparing two variances. Box (1953) showed that the level of this test becomes large in the presence of tiny deviations from the normality assumption (see Hampel et al. 1986; 188–189). Well known classical tests exhibit robustness problems too. The classical t-test and F-test for linear models are relatively robust with respect to the level, but they lack robustness of efficiency with respect to small departures from the normality assumption on the errors (cf. Hampel 1973; Schrader and Hettmansperger 1980; Ronchetti 1982; Heritier et al. 2009: 35). Nonparametric tests are attractive since they have an exact level under symmetric distributions and good robustness of efficiency. However, the distribution free property of their level is affected by asymmetric contamination (cf. Hampel et al. 1986: 201). Even ►randomization tests which keep an exact level, are not robust with respect to the power if they are based on a non-robust test statistic like the mean.

The first approach to formalize the robustness problem was Huber's (1964, 1981) minimax theory, where the statistical problem is viewed as a game between the Nature (which chooses a distribution in the neighborhood of the model) and the statistician (who chooses a statistical procedure in a given class). The statistician achieves robustness by constructing a minimax procedure which minimizes a loss criterion at the worst possible distribution in the neighborhood. More specifically, in the problem of testing a simple hypothesis against a simple alternative, Huber (1965, 1981) found the test which maximizes the minimum power over a neighborhood of the alternative, under the side condition that the maximum level over a neighborhood of the hypothesis is bounded. The solution to this problem which is an extension of ►Neyman-Pearson Lemma, is the censored likelihood ratio test. It can be interpreted in the framework of capacities (Huber and Strassen 1973) and it leads to exact finite sample minimax confidence intervals for a location parameter (Huber 1968). While Huber's minimax theory is one of the key ideas in robust statistics and leads to elegant and exact finite sample results, it seems difficult to extend it to general parametric models, when no invariance structure is available.

The infinitesimal approach introduced in Hampel (1968) in the framework of estimation, offers an alternative for more complex models. The idea is to view the quantities of interest (for instance the bias or the variance of an estimator) as functionals of the underlying distribution and to use their linear approximations to study their behavior in a neighborhood of the ideal model. A key tool is a derivative of such a functional, the influence function (Hampel 1974) which describes the local stability of the functional.

To illustrate the idea in the framework of testing, consider a parametric model $\{F_\theta\}$, where θ is a real parameter and a test statistic T_n which can be written (at least asymptotically) as a functional $T(F_n)$ of the empirical distribution function F_n. Let $H_0 : \theta = \theta_0$ be the null hypothesis and $\theta_n = \theta_0 + \Delta/\sqrt{n}$ a sequence of alternatives. We consider a neighborhood of distributions $F_{\epsilon,\theta,n} = (1-\epsilon/\sqrt{n})F_\theta + (\epsilon/\sqrt{n})G$, where G is an arbitrary distribution and we can view the asymptotic level α of the test as a functional of a distribution in the neighborhood. Then by a von Mises expansion of α around F_{θ_0}, where $\alpha(F_{\theta_0}) = \alpha_0$, the nominal level of the test, the asymptotic level and (similarly) the asymptotic power under contamination can be expressed as

$$\lim_{n\to\infty} \alpha(F_{\epsilon,\theta_0,n}) = \alpha_0 + \epsilon \int IF(x;\alpha,F_{\theta_0})dG(x) + o(\epsilon), \tag{1}$$

$$\lim_{n\to\infty} \beta(F_{\epsilon,\theta_n,n}) = \beta_0 + \epsilon \int IF(x;\beta,F_{\theta_0})dG(x) + o(\epsilon), \tag{2}$$

where

$$IF(x;\alpha,F_{\theta_0}) = \phi(\Phi^{-1}(1-\alpha_0))IF(x;T,F_{\theta_0}) \Big/ [V(F_{\theta_0},T)]^{1/2},$$

$$IF(x;\beta,F_{\theta_0}) = \phi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E})IF(x;T,F_{\theta_0}) \Big/ [V(F_{\theta_0},T)]^{1/2},$$

$\alpha_0 = \alpha(F_{\theta_0})$ is the nominal asymptotic level, $\beta_0 = 1 - \Phi(\Phi^{-1}(1-\alpha_0) - \Delta\sqrt{E})$ is the nominal asymptotic power, $E = [\xi'(\theta_0)]^2 \Big/ V(F_{\theta_0},T)$ is Pitman's efficacy of the test, $\xi(\theta) = T(F_\theta)$, $V(F_{\theta_0},T) = \int IF(x;T,F_{\theta_0})^2 dF_{\theta_0}(x)$ is the asymptotic variance of T, and $\Phi^{-1}(1-\alpha_0)$ is the $1-\alpha_0$ quantile of the standard normal distribution Φ and ϕ is its density (see Ronchetti 1979; Rousseeuw and Ronchetti 1979). More details can be found in Markatou and Ronchetti (1997) and Huber and Ronchetti (2009, Chap. 13).

Therefore, bounding the influence function of the the test statistic T from *above* will ensure *robustness of validity* and bounding it from *below* will ensure *robustness of efficiency*. This is in agreement with the exact finite sample result about the structure of the censored likelihood ratio test obtained using the minimax approach.

In the multivariate case and for general parametric models, the classical theory provides three asymptotically equivalent tests, Wald, score, and likelihood ratio test, which are asymptotically uniformly most powerful with respect to a sequence of contiguous alternatives. If the parameter of the model is estimated by a robust estimator such as an M-estimator T_n defined by the estimating equation $\sum_{i=1}^n \psi(x_i;T_n) = 0$, natural extensions of the three classical tests can be constructed by replacing the score function of the model by the function ψ. This leads to formulas similar to (1) and (2) and to optimal bounded influence tests (see Heritier and Ronchetti 1994).

About the Author

Professor Elvezio Ronchetti is Past Vice-President of the Swiss Statistical Association (1988–1991). He was Chair, Department of Econometrics, University of Geneva (2001–2007). He is an Elected Fellow of the American Statistical Association (2001) and of the International Statistical Institute (2008). Currently he is an Associate Editor, *Journal of the American Statistical Association* (2005–present) and Director of the Master of Science and PhD Program in Statistics, University of Geneva (2009–present). He is the co-author (with F.R. Hampel, P.J. Rousseeuw, and W.A. Stahel) of the well known text *Robust Statistics : The Approach Based on Influence Functions* (Wiley, New York, 1986, translated also into Russian), and of the 2nd edition of Huber's classic *Robust Statistics* (with P. J. Huber, Wiley 2009.)

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Confidence Interval
►Multivariate Technique: Robustness
►Neyman-Pearson Lemma
►Nonparametric Statistical Inference
►Power Analysis
►Randomization Tests
►Robust Regression Estimation in Generalized Linear Models
►Robust Statistical Methods
►Robust Statistics
►Statistical Inference
►Statistical Inference: An Overview
►Student's t-Tests
►Tests for Homogeneity of Variance

References and Further Reading

Box GEP (1953) Non-normality and tests on variances. Biometrika 40:318–335
Hampel FR (1968) Contribution to the theory of robust estimation. PhD Thesis, University of California, Berkeley
Hampel FR (1973) Robust estimation: a condensed partial survey. Z Wahrsch Verwandte Geb 27:87–104
Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69:383–393
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York

Heritier S, Ronchetti E (1994) Robust bounded-influence tests in general parametric models. J Am Stat Assoc 89:897–904
Heritier S, Cantoni E, Copt S, Victoria-Feser M-P (2009) Robust methods in biostatistics. Wiley, Chichester
Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101
Huber PJ (1965) A robust version of the probability ratio test. Ann Math Stat 36:1753–1758
Huber PJ (1968) Robust confidence limits. Z Wahrsch Verwandte Geb 10:269–278
Huber PJ (1981) Robust statistics. Wiley, New York
Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, New York
Huber PJ, Strassen V (1973) Minimax tests and the Neyman-Pearson lemma for capacities. Ann Stat 1:251–263, 2:223–224
Markatou M, Ronchetti E (1997) Robust inference: the approach based on influence functions. In: Maddala GS, Rao CR (eds) Handbook of Statistics, vol 15. North Holland, Amsterdam, pp 49–75
Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, New York
Ronchetti E (1979) Robustheitseigenschaften von Tests. Diploma Thesis, ETH Zürich, Switzerland
Ronchetti E (1982) Robust testing in linear models: The infinitesimal approach. PhD Thesis, ETH Zürich, Switzerland
Rousseeuw PJ, Ronchetti E (1979) The influence curve for tests. Research report 21. Fachgruppe für Statistik, ETH Zürich, Switzerland
Schrader RM, Hettmansperger TP (1980) Robust analysis of variance based upon a likelihood ratio criterion. Biometrika 67:93–101
Tukey JW (1960) A survey of sampling from contaminated distributions. In: Olkin I (ed) Contributions to probability and statistics. Stanford University Press, Stanford, pp 448–485

Robust Regression Estimation in Generalized Linear Models

Nor Aishah Hamzah[1], Mohammed Nasser[2]
[1]Professor
University of Malaya, Kuala Lumpur, Malaysia
[2]Professor
University of Rajshahi, Rajshahi, Bangladesh

The idea of ▶generalized linear models (GLM) generated by Nelder and Wedderburn (1972) seeks to extend the domain of applicability of the linear model by relaxing the normality assumption. In particular, GLM can be used to model the relationship between the explanatory variable, X, and a function of the mean, μ_i, of a continuous or discrete responses. More precisely, GLM assumes that $g(\mu_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$, where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is the p-vector of unknown parameters and $g(\cdot)$ is the link function that determines the scale on which linearity is assumed. Models of this type include logistic and probit regression, Poisson regression, linear regression with known variance, and certain models for lifetime data.

Specifically, let $Y_1, Y_2, \ldots, Y_n$, be n independent random variables drawn from the exponential family with density (or probability function)

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \qquad (1)$$

for some specific functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$. Here, $E(Y_i) = \mu_i = b'(\theta_i)$ and $\mathrm{var}(Y_i) = b''(\theta_i)a(\phi)$ with usual notation of derivative.

The most common method of estimating the unknown parameter, β, is that of maximum likelihood estimation (MLE) or quasi-likelihood methods (QMLE), which are equivalent if $g(\cdot)$ is the canonical link such as the logit function for the ▶logistic regression, the log function for ▶Poisson regression, or the identity function for the Normal regression. That is, when $g(\mu_i) = \theta_i$, the MLE and QMLE estimator of β are the solutions of the p-system of equations:

$$\sum_{i=1}^{n}(y_i - \mu_i)x_{ij} = 0, \quad j = 1, \ldots, p. \qquad (2)$$

The estimator defined by (2) can be viewed as an M-estimator with score function

$$\psi(y_i; \beta) = (y_i - \mu_i)x_i \qquad (3)$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Since the score function defined by (3) is proportional to x and y, the maximum possible influence in both the x and y spaces are unbounded. When y is categorical, the problem of unbounded influence in x remains and in addition, the breakdown possibility by inliers arises (Albert and Anderson 1984). As such, the corresponding estimator of β based on (2) is therefore non-robust. Any attempt to improve the estimation of such β should limit such influences. Two basic approaches are usually employed in order to address the problems stated above, that is: (a) diagnostics and (b) robust estimation.

Diagnostic Measures

In most diagnostics approaches, the MLE is first employed and subsequently diagnostics tools are used to identify potential influential observations. For details on diagnostic measures, readers are referred to the published works of Pregibon (1981, 1982), McCullagh and Nelder (1989), Johnson (1985), Williams (1987), Pierce and Schafer (1986), Thomas and Cook (1990), and Adimari and Ventura (2001).

While these techniques have been quite successful in identifying individual influential points, its generalization to jointly influential points cannot guarantee success. The development of a robust method in the early 1980s provides an option that offers automatic protection against anomalous data. A recent trend in diagnostic research is (a) to detect wild observations by using the classical diagnostic method after initially deploying the robust method (Imon and Hadi 2008) or (b) to use robust method in any case (Cantoni and Ronchetti 2001; Serigne and Ronchetti 2009).

Robust Estimation

Since the score function in (3) is subject to influence of outlying observation, both in the X and y, appropriate robust estimations are those of the GM-estimates. These include the Mallows-type (Pregibon 1979) and Schweppe-type (Stefanski et al. 1986; Künsch et al. 1989). The proposed methods are discussed here. Let

$$\ell(\theta_i, y_i) = \log f(y_i; \theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \quad (4)$$

and define the i-th deviance as $d_i = d_i(\theta_i) = 2\{\ell(\tilde{\theta}_i, y_i) - \ell(\theta_i, y_i)$, where $\tilde{\theta}_i$ is the MLE based on observation y_i alone, that is, $\tilde{\theta}_i = (b')^{-1}(y_i)$. The deviance d_i can be interpreted as a measure of disagreement of the i-th observation and the fitted model. Thus, MLE that aims at maximizing the likelihood function also aims at minimizing the deviances, specifically minimizing $M(\beta) = \sum_{i=1}^{n} d_i(\theta)$.

In an attempt to robustify the MLE, the first modification of the MLE introduced by Pregibon (1979) is to replace the minimization criterion with $M(\beta) = \sum_{i=1}^{n} \rho(d_i)$.

The function $\rho(\cdot)$ acts as a filter that limits the contribution of extreme observations in determining the fits to the data. Minimizing the criterion above can be obtained by finding the root solutions to the following score function

$$\sum_{i=1}^{n} \psi(d_i) = \sum_{i=1}^{n} w_i s_i x_{ij} = 0, \quad j = 1, \ldots, p, \quad (5)$$

with $s_i = \partial\ell(\theta_i, y_i)/\partial\eta_i$, and $w_i (0 \le w_i \le 1)$ given by $w(d_i) = \partial\rho(d_i)/\partial d_i$. Note that this is simply the weighted version of the maximum likelihood score equations with data-dependent weights.

Mallows-Type GM Estimate

Based on Huber's loss function, the corresponding weight function $w_i = \min\{1, (H/d_i)^{1/2}\}$ with adjustable tuning constant H, which aims at achieving some specified efficiency, can be used (Pregibon 1982). By solving (5), one can obtain a class of Mallows M-estimates. This type of estimation is resistant to poorly fitted data, but not to extreme observations in the covariate space that may exert undue influence on the fit.

Schweppe-Type GM Estimate

Extending the results obtained by Krasker and Welsch (1982) and Stefanski et al. (1986), Künsch et al. (1989) proposed bounded influence estimators that are also conditionally Fisher-consistent. Subject to a bound b on the measure of sensitivity $\gamma_\psi (\gamma_\psi \le b < \infty)$, the following modification to the score function was proposed:

$$\psi_{BI} = \left\{ y - \mu - c\left(x^T\beta, \frac{b}{(x^T B^{-1} x)^{1/2}} \right) \right\} w_b(|r(y, x, \beta, B)|(x^T B^{-1} x)^{1/2}) x^T$$

where $c(\cdot, \cdot)$ and B are the respective bias-correction term and dispersion matrix chosen so that the estimates are conditionally Fisher-consistent with bounded influence, with weight function of the form $w_b(a) = \min\{1, b/a\}$ based on Huber's loss function. As in Schweppe-type GM estimates, $w_b(\cdot)$ downweight observations with a high product of corrected residuals and leverage. Details on the terms used here can be found elsewhere (see, e.g., Huber (1981) on infinitesimal sensitivity).

Besides the general approach in robust estimation in GLM several researchers put forward various other estimators for specific case of GLM. For example, when y follows a Gamma distribution with log link function, Bianco et al. (2005) considered redescending M-estimators and showed that the estimators are Fisher-consistent without any correction term. In the logistic model, Carrol and Pederson (1993) proposed weighted MLE to robustify estimators, Bianco and Yohai (1996) extended the work of Morgenthaler (1992) and Pregibon (1982) on M-estimators while Croux and Haesbroeck (2003) developed a fast algorithm to execute Bianco–Yohai estimators. Gervini (2005) presented robust adaptive estimators and recently Hobza et al. (2008) opened a new line proposing robust median estimators in ▶logistic regression (see also Hamzah 1995). The robust Poisson regression model (RPR) (see ▶Poisson Regression) was proposed by Tsou (2006) for the inference about regression parameters for more general count data; here one need not worry about the correctness of the Poisson assumption.

Cross References

▶Generalized Linear Models
▶Influential Observations
▶Outliers

►Regression Diagnostics
►Robust Statistics

References and Further Reading

Adimari G, Ventura L (2001) Robust inference for generalized linear models with application to logistic regression. Stat Probab Lett 55:413–419
Albert A, Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1):1–10
Bianco A, Yohai V (1996) Robust estimation in the logistic regression model. In: Rieder H (ed) Robust statistics, data analysis, and computer intensive methods. Lecture notes in statistics, vol 109, Springer, New York, pp 17–34
Bianco AM, Garcia Ben M, Yohai VJ (2005) Robust estimation for linear regression with asymmetric error. Can J Stat 33: 511–528
Carrol RJ, Pederson S (1993) On robustness in logistic regression model. J Roy Stat Soc B 55:693–706
Cantoni E, Ronchetti E (2001) Robust inference for generalized linear models. J Am Stat Assoc 96:1022–1030
Croux C, Haesbroeck G (2003) Implementing the Bianco and Yohai estimator for logistic regression. Comput Stat Data Anal 44:273–295
Gervini D (2005) Robust adaptive estimators for binary regression models. J Stat Plan Infer 131:297–311
Hobza T, Pardo L, Vajda I (2008) Robust median estimator in logistic regression. J Stat Plan Infer 138:3822–3840
Hamzah NA (1995) Robust regression estimation in generalized linear models, University of Bristol, Ph.D. thesis
Huber PJ (1981) Robust Statistics. Wiley, New York
Imon AHMR, Hadi AS (2008) Identification of multiple outliers in logistic regression. Commun Stat Theory Meth 37(11): 1697–1709
Johnson W (1985) Influence measures for logistic regression: Another point of view. Biometrika 72:59–65
Krasker WS, Welsch RE (1982) Efficient bounded-influence regression estimation. J Am Stat Assoc 77:595–604
Künsch H, Stefanski L, Carroll RJ (1989) Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. J Am Stat Assoc 84:460–466
McCullagh P, Nelder JA (1989) Generalized Linear Models. Chapman and Hall, London
Morgenthaler S (1992) Least-absolute-deviations fits for generalized linear models. Biometrika 79:747–754
Nelder JA, Wedderburn RWM (1972) Generalized Linear Models. J Roy Stat Soc A 135:370–384
Pierce DA, Schafer DW (1986) Residual in generalized linear model. J Am Stat Assoc 81:977–990
Pregibon D (1979) Data analytic methods for generalized linear models. University of Toronto, Ph. D. thesis
Pregibon D (1981) Logistic regression diagnostics. Ann Stat 9:705–724
Pregibon D (1982) Resistant fits for some commonly used logistic models with medical applications. Biometrics 38: 485–498
Serigne NL, Ronchetti E (2009) Robust and accurate inference for generalized linear models. J Multivariate Anal 100:2126–2136
Stefanski L, Carroll RJ, Ruppert D (1986) Optimally bounded score functions for generalized linear models, with applications to logistic regression. Biometrika 73:413–425
Thomas W, Cook RD (1990) Assessing influence on predictions from generalized linear models. Technometrics 32:59–65
Tsou T-S, Poisson R (2006) regression. Journal of Statistical Planning and Inference 136:3173–3186
Williams DA (1987) Generalized linear model diagnostics using the deviance and single case deletions. Appl Stat 36:181–191

Robust Statistical Methods

Ricardo Maronna
Professor
University of La Plata and C.I.C.P.B.A., La Plata, Buenos Aires, Argentina

Outliers

The following Table (Hand et al. 1994: 278) contains 20 measurements of the speed of light in suitable units (km/s minus 299000) from the classical experiments performed by Michelson and Morley in 1887.

880	880	880	860	720
720	620	860	970	950
880	910	850	870	840
840	850	840	840	840

We may represent our data as

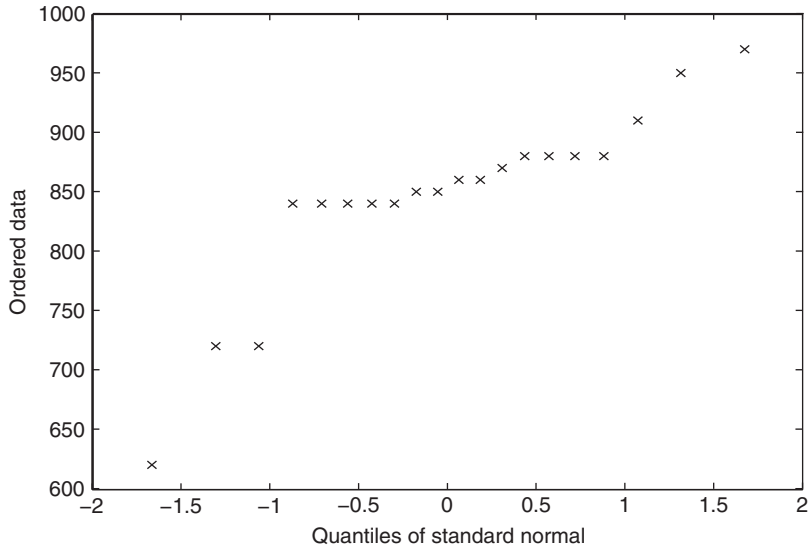
$$x_i = \mu + u_i,\ i = 1, \ldots, n \qquad (1)$$

where $n = 20$, μ is the true (unknown) speed value and u_i are random observation errors. We want a point estimate $\widehat{\mu}$ and a ►confidence interval for μ.

Figure 1 is the normal QQ-plot of the data. The three smallest observations clearly stand out from the rest. The central part of the plot is approximately linear, and therefore we may say that the data are "approximately normal."

The left-hand half of the following Table shows the sample mean and standard deviation (SD) of the complete data and also of the data without the three smallest observations (the right-hand half will be described below).

	Mean	SD	Median	MADN
Complete data	845.0	79.1	855.0	29.6
3 obs. omitted	872.9	38.5	860.0	29.6



Robust Statistical Methods. Fig. 1 Speed of light: normal QQ-plot of data

We see that these three observations inflate the SD and diminish the mean.

The confidence intervals with level 0.95 for the mean with the complete data and with the three outliers removed are respectively [807.94, 882.06] and [853.14 892.74].

We see that even data from a carefully controlled experiment may contain atypical observations ("outliers") which may overly influence the conclusions from the experiment. Although the proportion of outliers is low (3/20=15%) they have a serious influence.

The oldest approach to deal with this problem is to employ some diagnostic tool to detect ▶outliers, delete them, and then recompute the statistics of interest. Barnett and Lewis (1998) is a useful source of methods for outlier detection.

Using a good outlier diagnostic is clearly better than doing nothing, but has its drawbacks:

- Deletion requires a subjective decision. When is an observation "outlying enough" to be deleted?
- The user or the author of the data may feel uneasy about deleting observations
- There is a risk of deleting "good" observations, which results in underestimating data variability
- Since the results depend on the user's subjective decisions, it is difficult to determine the statistical behavior of the complete procedure.

Robust statistical methods are procedures that require no subjective decisions from the user, and that

- give approximately the same results as classical methods when there are no atypical observations, and
- are only slightly affected by a small or moderate proportion of atypical observations.

The sample median $\mathrm{Med}(\mathbf{x})$ is a robust alternative to the median. The median absolute deviation from the median $\mathrm{MAD}(\mathbf{x}) = \mathrm{Med}(|\mathbf{x}\text{-}\mathrm{Med}(\mathbf{x})|)$ is a robust dispersion estimate. The normalized MAD: $\mathrm{MADN}(\mathbf{x}) = \mathrm{MAD}(\mathbf{x})/0.675$ is a robust alternative to the SD; for large normal samples MADN and SD are approximately equal. The right-hand half of the Table above shows the sample median and MADN for the complete data and the data with the three smallest observations omitted. We see that the median has only a small change, and that MADN remains the same.

Then, why not always use the median instead of the mean? To answer this question we have to analyze the behavior of the estimates at a given model. Assume that u_i are normal: $\mathrm{N}(0, \sigma^2)$. Then the sample mean has variance $\mathrm{Var}(\bar{x}) = \sigma^2/n$, while for large n the sample median has $\mathrm{Var}(\mathrm{Med}(\mathbf{x})) \approx 1.571\sigma^2/n$ (proofs for all results can be found in Maronna et al. 2006). We say that the sample median has asymptotic efficiency 1/1.571 = 0.636 at the normal. This means that we have to pay a high price for the median's robustness. We may make requirement (1) above more precise by stating that we want an estimate with a high efficiency at the normal, while keeping condition (2). We now consider two approaches to attain this goal.

M Estimates

Let u_i have a positive density function f, so that x_i in (1) has density $f(x-\mu)$. Then the maximum likelihood estimate (MLE) of μ is the solution of

$$\prod_{i=1}^{n} f(x_i-\mu) = \max.$$

Taking logs we get

$$\sum_{i=1}^{n} \rho(x_i-\mu) = \min \tag{2}$$

where $\rho = -\log(x)$. If f N$(0,1)$ we have $\rho(x) = \left(x^2+\log(2\pi)\right)/2$. Note that using this ρ is equivalent to using $\rho(x) = x^2$, which yields $\widehat{\mu} = \bar{x}$. If f is the double exponential density $f(x) = 0.5\exp(-|x|)$ we get likewise $\rho(x) = |x|$, which yields $\widehat{\mu} = \text{Med}(\mathbf{x})$.

An M estimate is defined through (2) where $\rho(x)$ is a given function (which does not necessarily correspond to a MLE). To fulfill (1) it has to be approximately quadratic for small x; to fulfill (2) it must increase more slowly than x^2 for large x. An important case is the Huber ρ-function

$$\rho(x) = \begin{cases} x^2 & \text{for } |x| \le k \\ 2k|x| - k^2 & \text{for } |x1 > k. \end{cases}$$

Figure 2 plots ρ for $k = 2$.

The limit cases $k \to \infty$ and $k \to 0$ correspond respectively to x^2 and $|x|$, and therefore the estimate is an intermediate between the mean and the median. Differentiating (2) we get that $\widehat{\mu}$ is a solution to the *estimating equation*

$$\sum_{i=1}^{n} \psi(x_i-\widehat{\mu}) = 0 \tag{3}$$

where $\psi = \rho'$. For the Huber function we have that (up to a constant)

$$\psi(x) = \begin{cases} -k & \text{for } x < -k \\ x & \text{for } |x1 \le k. \\ k & \text{for } x > k \end{cases}$$

Figure 3 displays ψ for $k = 2$.

The boundedness of ψ makes the estimate robust.

It can be shown that for any symmetric distribution of the u_i, for large n the distribution of $\widehat{\mu}$ is approximately N$(\mu, v/n)$ where the asymptotic variance v is given by

$$v = \frac{\text{E}\psi(x-\mu)^2}{[\text{E}\psi'(x-\mu)]^2}.$$

The following Table gives the normal efficiencies of the Huber estimate for different values of k.

k	Efficiency
0	0.64
1.0	0.90
1.4	0.95
∞	1.00

It is seen that $k = 1.4$ yields a high efficiency.

Define now the *weight function* W as $W(x) = \psi(x)/x$. For the Huber function we have

$$W(x) = \begin{cases} 1 & \text{for } |x| \le k \\ k/|x| & \text{for } |x| > k. \end{cases}$$

Figure 4 plots Huber's W :

We may rewrite (3) as

$$\sum_{i=1}^{n} W(x_i-\widehat{\mu})(x_i-\widehat{\mu}) = 0$$

and therefore

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{4}$$

where $w_i = W(x_i-\widehat{\mu})$. This shows that a location M estimate can be thought of as a weighted mean with weights w_i, where observations distant from the "bulk" of the data receive smaller weights.

Note however that (4) is not an explicit formula for $\widehat{\mu}$, since w_i depends on both x_i and $\widehat{\mu}$. It can however be used as a basis for the iterative numerical computing of $\widehat{\mu}$.

L Estimates

A different approach to robust location estimates is based on the ordered observations $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ ("order statistics"). The simplest is the *α-trimmed mean*. For $\alpha \in [0, 0.5]$ let $m = [\alpha(n-1)]$ where $[.]$ stands for the integer part. Then the α-trimmed mean is defined as

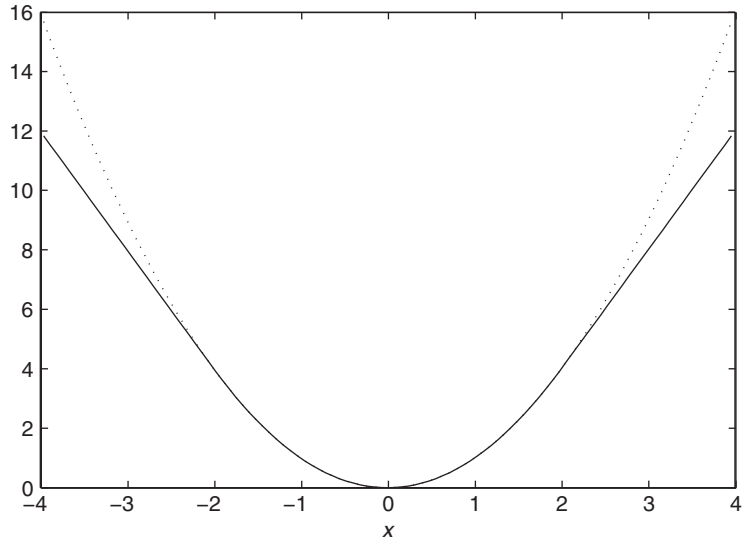
$$\bar{x}_\alpha = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} x_{(i)}. \tag{5}$$

That is, a proportion α of the largest and smallest observations are deleted. It can be shown that for $\alpha = 0.25$ the efficiency of $\bar{x}_\alpha$ is 0.83, although it seems that we are "deleting" half of the sample!. The reason is that $\bar{x}_\alpha$ is actually a function of *all* observations, even of those that do not appear in (5).

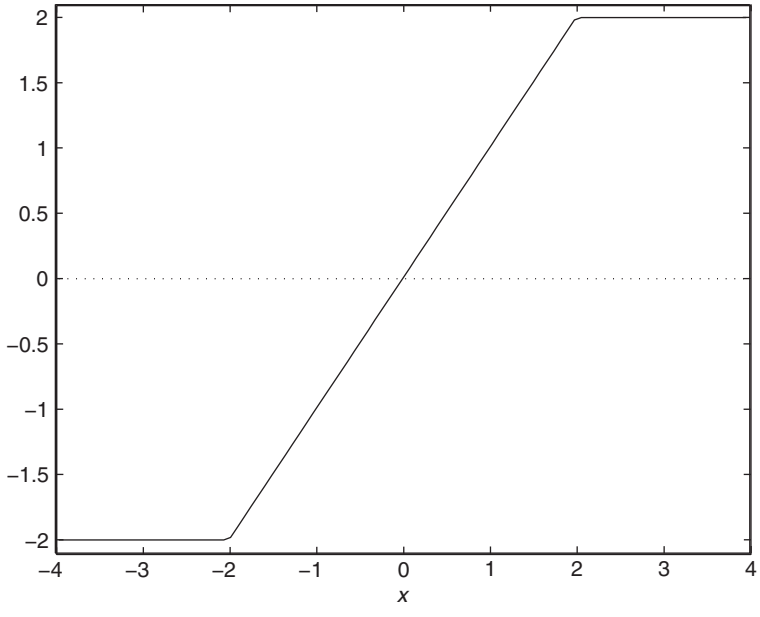
In general, L estimates are linear combinations of ▶order statistics:

$$\widehat{\mu} = \sum_{i=1}^{n} a_i x_{(i)}$$

where the a_i are constants such that $a_i = a_{n-1+i}$ and $\sum_{i=1}^{n} a_i = 1$.



Robust Statistical Methods. Fig. 2 Huber $\rho(x)$ with $K = 2$ (full line) and x^2 (dotted line)



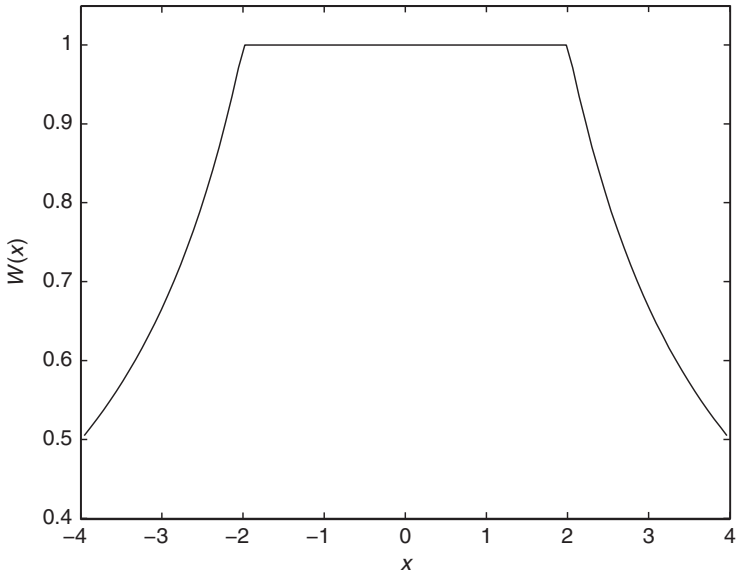
Robust Statistical Methods. Fig. 3 Huber's ψ for $k = 2$

Although L estimates seem simpler than M estimates, they are difficult to generalize to regression and multivariate analysis. On the other hand, M estimates can be generalized to more complex situations.

General Considerations

The present exposition attempts to give the reader a flavor of what robust methods are, through the incomplete treatment of a very simple situation. It is based on the author's experience and personal preferences.

The book by Maronna et al. (2006) contains a general and up to date account of robust methods. The classic book by Huber (1981) and the recent one by Jurecková and Picek (2006) contain more theoretical material. Hampel et al. (1986) gives a particular approach to robustness. Rousseeuw and Leroy (1987) deal (although not exclusively) with an important approach to robust regression.



Robust Statistical Methods. Fig. 4 Huber's weight function for $k = 2$

About the Author

Dr. Ricardo Maronna is Consulting Professor, University of La Plata, and Researcher at C.I.C.P.B.A., both in La Plata, Argentina. He has been three times Head of the Mathematics Department of the Faculty of Exact Sciences of the University of La Plata. He is Past President of the Statistical Society of Argentina (2003–2004). He is the author or co-author of 40 papers on statistical methods and their applications, and of the highly praised book *Robust Statistics: Theory and Methods* (with R.D. Martin and V.J. Yohai, John Wiley and Sons, 2006). "This book belongs on the desk of every statistician working in robust statistics, and the authors are to be congratulated for providing the profession with a much-needed and valuable resource for teaching and research." (*Tyler, David E. (2008), Journal of the American Statistical Association*, **103:** June 2008, p. 889.)

Cross References

►Adaptive Linear Regression
►Adaptive Methods
►Mean Median and Mode
►Multivariate Outliers
►Multivariate Technique: Robustness
►Order Statistics
►Outliers
►Robust Inference
►Robust Statistics

References and Further Reading

Barnett V, Lewis T (1998) Outliers in statistical data, 3rd edn. Wiley, New York

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York

Hand DJ, Daly F, Dunn AD, McConway KJ, Ostrowski E (1994) A handbook of small data sets. Chapman & Hall, London

Huber PJ (1981) Robust statistics. Wiley, New York

Jurecková J, Picek J (2006) Robust statistical methods with R. Chapman & Hall, London

Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics: theory and methods. Wiley, New York

Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York

Robust Statistics

Peter J. Huber
Klosters, Switzerland

Introduction

The term "robust" was introduced into the statistical literature by Box (1953). By then, robust methods such as trimmed means, had been in sporadic use for well over a century, see for example Anonymous (1821). However, Tukey (1960) was the first person to recognize the

extreme sensitivity of some conventional statistical procedures to seemingly minor deviations from the assumptions, and to give an eye-opening example. His example, and his realization that statistical methods optimized for the conventional Gaussian model are unstable under small perturbations were crucial for the subsequent theoretical developments initiated by Huber (1964) and Hampel (1968).

In the 1960s robust methods still were considered "dirty" by most. Therefore, to promote their reception in the statistical community it was crucial to mathematize the approach: one had to prove optimality properties, as was done by Huber's minimax results (1964, 1965, 1968), and to give a formal definition of qualitative robustness in topological terms, as was done by Hampel (1968, 1971). The first book-length treatment of theoretical robustness was that by Huber (1981, 2nd edition by Huber and Ronchetti 2009).

M-Estimates and Influence Functions

With Huber (1964) we may formalize a robust estimation problem as a game between the Statistician and Nature. Nature can choose any distribution within some uncertainty region, say an ε-contamination neighborhood of the Gaussian distribution (i.e., a fraction ε of the observations comes from an arbitrary distribution). The Statistician can choose any M-estimate, that is, an estimate defined as the solution $\hat{\theta}$ of an equation of the form

$$\sum \psi(x_i, \theta) = 0, \tag{1}$$

where ψ is an arbitrary function. If $\psi(x, \theta) = (\partial/\partial\theta) \log f(x, \theta)$ is the logarithmic derivative of a probability density, then $\hat{\theta}$ is the maximum likelihood estimate. The Statistician aims to minimize the worst-case asymptotic variance of the estimate.

It can be seen from (1) that in large samples the influence of the ith observation toward the value of $\hat{\theta}$ is proportional to $\psi(x_i, \theta)$. Hampel (1968; 1974, see also Hampel et al. 1986) generalized this notion through his *influence curve* (or *influence function*) to more general types of estimators. In the case of M-estimates the influence function is proportional to $\psi(x, \theta)$. Arguably, the influence function is the most useful heuristic tool of robustness. To limit the influence of gross errors, the influence function should be bounded, and a simple method for constructing a robust M-estimate is to choose for ψ a truncated version of the logarithmic derivative of the idealized model density.

In simple cases, in particular the estimation of a one-dimensional location parameter, the game between the Statistician and Nature has an explicit asymptotic minimax solution: find the *least favorable* distribution (i.e., minimizing Fisher information) within the uncertainty region. This is the minimax strategy for Nature. The asymptotic minimax strategy for the Statistician then is the maximum likelihood estimate for the least favorable distribution. In fact, error distributions occurring in practice are well modeled by least favorable distributions corresponding to contamination rates between 1% and 10%, better than by the Gaussian model itself.

Note that bounding the influence provides safety not only against ▶outliers ("gross errors"), but also against all other types of contamination. All three approaches: the simple-minded truncation of the logarithmic derivative, the asymptotic minimax solution, and the finite sample minimax solution (see below) lead to qualitatively identical ψ-functions.

Robustness, Large Deviations and Diagnostics

By 1970 John Tukey's interests had changed their focus, he scorned models, and for him, robust methods now were supposed to have a good performance for the widest possible variety of (mostly longtailed) distributions. His shift of the meaning of the word "robust" inevitably created some confusion. I still hold (with Tukey 1960, Huber 1964 and Hampel 1968) that robust statistics should be classified with parametric statistics, and that robustness primarily should be concerned with safeguarding against ill effects caused by finite but small deviations from an idealized model, with emphasis on the words *small* and *model*. Interpretation of the results in terms of a model becomes difficult if one leaves the neighborhood of that model. Good properties far away from the model should be regarded as a bonus rather than as a must.

The concern with large deviations (see ▶Large Deviations and Applications) has caused a concomitant confusion between the complementary roles of diagnostics and robustness. The purpose of robustness is to *safeguard* against deviations from the assumptions, while the purpose of diagnostics is to *identify* and *interpret* such deviations. Robustness is concerned in particular with deviations that are near or below the limits of detectability. Safeguards against those can be achieved in a mechanical, almost blind fashion, even if the sparsity of data may prevent you from going beyond. Diagnostics on the other hand comes into play with larger deviations; it is an art, requiring insight into the processes generating the data.

The Breakdown Point

The standard interpretation of contamination models is that a dominant fraction $1 - \varepsilon$ of the data consists of "good"

observations that follow the idealized model, while a small fraction ε of "bad" observations does not.

The breakdown point ε^* is the smallest fraction ε of bad observations that may cause an estimator to take on arbitrarily large aberrant values. This concept is a very simple, but extremely useful global characteristic of a robust procedure. Hampel (1968) had given it an asymptotic definition, but actually, it is most useful in small sample situations (Donoho and Huber 1983).

Robust statistical procedures should have a reasonably high breakdown point (i.e., at least in the range of 10% to 25%). A higher values is desirable – if it comes for free and does not unduly impair performance at the model. Indeed, robust M-estimates of one-dimensional parameters in large samples typically approach the maximum possible breakdown point of 50%. This is not so in higher dimensions: M-estimates of d-dimensional location parameters and covariance matrices have a disappointingly low breakdown point $\varepsilon^* \leq 1/(d+1)$, see Maronna (1976). For a while this limit was thought to hold generally for all affine equivariant estimators, but then it was found that with the help of projection pursuit methods it is possible to construct estimators approaching an asymptotic breakdown point of 50%, see Donoho and Huber (1983). However, these estimators are overly pessimistic by having a low efficiency at the model, and they are very computer intensive.

Over the years it has become fashionable to strive for the highest possible breakdown point, particularly in regression situations, where observations that are influential through their position in factor space (the so-called "leverage points") present peculiar problems. While a proof that the theoretical maximum of 50% can be attained is interesting and theoretically important, the corresponding procedures in my opinion suffer from what I have called the Souped-up Car Syndrome (Huber 2009): they optimize one aspect to the detriment of others. For example, the high breakdown point S-estimators of regression even lack the crucial stability attribute of robust procedures (Davies 1993, Section 1.6).

With high values of ε, alternative interpretations of contamination models become important, transcending the ubiquitous presence of a small fraction of gross errors. The data may be a mixture of two or more sets with different (e.g., ethnic) origins, and the task no longer is to ignore a small discordant minority of gross errors (a robustness problem), but to disentangle larger mixture components (a diagnostic problem). High breakdown point procedures can be used for diagnostic purposes, namely to identify a dominant mixture component, but they need not provide the best possible approach.

Bayesian Robustness

The term "robust" had been coined by a Bayesian (Box 1953). Ironically, while there is a fairly large literature in the form of journal articles – see, for example, Berger's (1994) overview – Bayesianism never quite assimilated the concept. The reason seems to be that for an orthodox Bayesian statistician probabilities exist only in his mind, and that he therefore cannot separate the model (i.e., his belief) from the true underlying situation. For a pragmatic Bayesian like Box, robustness was a property of the model (which he was willing to adjust in order to achieve robustness), while for a pragmatic frequentist like Tukey, it was a property of the procedure (and he would tamper with the data by trimming or weighting them to achieve robustness). To me as a decision theorist, the dispute between Box and Tukey about the proper robustness concept was a question of the chicken and the egg: which comes first, the least favorable model of Nature, or the robust minimax procedure of the Statistician? See Huber and Ronchetti (2009), Chapter 15, in particular p. 325.

Finite Sample Results and Robust Tests

In his decision theoretic formalization, Huber (1964) had imposed an unpleasant restriction on Nature by allowing only symmetric contaminations. It seems to be little known that this restriction is irrelevant; it can be removed by an approach through finite sample robust tests, Huber (1965, 1968). The extension of robust tests beyond the single-parameter case, however, is difficult; see Huber and Ronchetti (2009), Chapter 13.

Heuristic Aspects of Robustness

There are no rigorous optimality results available once one leaves the single-parameter case. Admittedly, the perceived need for mathematical rigor and proven optimality properties has faded away after the 1960s. But at least, one should subject one's procedures to a worst case analysis in some neighborhood of the model. Even this is difficult and rigorously feasible only in few cases. A good heuristic alternative is a combination of infinitesimal approaches (influence function or shrinking neighborhoods) with breakdown point aspects. Shrinking neighborhoods were first dealt with by Huber-Carol (1970) in her thesis, and a comprehensive treatment was given by Rieder (1994).

In my opinion the crucial attribute of robust methods is stability under small perturbations of the model. I am tempted to claim that robustness is not a collection of procedures, but rather a state of mind: a statistician should keep in mind that *all* aspects of a data analytic setup (experimental design, data collection, models, procedures)

should be such that minor deviations from the assumptions cannot have large effects on the results (a robustness problem), and that major deviations can be discovered (a diagnostic problem). Compromises are unavoidable. For example, the so-called "optimal" linear regression designs, which evenly distribute the observations on the d corners of a $(d-1)$-dimensional simplex, on one hand lack redundancy to spot deviations of the response surface from linearity, and on the other hand, already subliminal deviations from linearity may impair optimality to such an extent that the "naive" design (which distributes the observations evenly over the entire design space) is superior. Moreover, if there is a problem at a single corner of the simplex, affecting half of the observations there, then this can cause breakdown, leading to a breakdown point no better than $\varepsilon^* \cong 1/(2d)$. See Huber and Ronchetti (2009), Chapter 9, and Chapter 11, p. 285.

About the Author

Professor Huber is a Fellow of the American Academy of Arts and Sciences. He received a Humboldt Award in 1988. He was a Professor of statistics at ETH Zurich (Switzerland), Harvard University, Massachusetts Institute of Technology, and the University of Bayreuth (Germany). Peter Huber has published four books and over 70 papers on statistics and data analysis, including the fundamental paper on robust statistics "Robust Estimation of a Location Parameter" (Annals of Mathematical Statistics, (1964) Volume 35, Number 1, 73–101), and the text *Robust Statistics* (Wiley, 1981; republished in paperback 2004). In addition to his fundamental results in robust statistics, Peter Huber made important contributions to computational statistics, strategies in data analysis, and applications of statistics in fields such as crystallography, EEGs, and human growth curves.

Cross References

►Bayesian Statistics
►Functional Derivatives in Statistics: Asymptotics and Robustness
►Imprecise Probability
►Large Deviations and Applications
►Misuse of Statistics
►Multivariate Technique: Robustness
►Optimality and Robustness in Statistical Forecasting
►Robust Inference
►Robust Statistical Methods
►Statistical Fallacies: Misconceptions, and Myths

References and Further Reading

Anonymous (1821) Dissertation sur la recherche du milieu le plus probable. Ann Math Pures et Appl 12:181–204
Berger JO (1994) An overview of robust Bayesian analysis. Test 3: 5–124
Box GEP (1953) Non-normality and tests on variances. Biometrika 40:318–335
Davies PL (1993) Aspects of robust linear regression. Ann Stat 21:1843–1899
Donoho DL, Huber PJ (1983) The notion of breakdown point. In: Bickel PJ, Doksum KA, Hodges JL (eds) A festschrift for Erich L. Lehmann. Wadsworth, Belmont
Hampel FR (1968) Contributions to the theory of robust estimation, Ph.D. Thesis. University of California, Berkeley
Hampel FR (1971) A general qualitative definition of robustness. Ann Math Stat 42:1887–1896
Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 62:1179–1186
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics. The approach based on influence. Wiley, New York
Huber PJ (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101
Huber PJ (1965) A robust version of the probability ratio test. Ann Math Stat 36:1753–1758
Huber PJ (1968) Robust confidence limits. Z Wahrscheinlichkeitstheorie Verw Gebiete 10:269–278
Huber PJ (1981) Robust statistics. Wiley, New York
Huber PJ (2009) On the non-optimality of optimal procedures. In: Rojo J (ed) Optimality. The third E. L. Lehmann symposium. Institute of Mathematical Statistics, Lecture Notes Vol. 57. Beachwood, Ohio, USA, pp 31–46
Huber PJ, Ronchetti EM (2009) Robust statistics, 2nd edn. Wiley, New York
Huber-Carol C (1970) Etude asymptotique de tests robustes, Ph.D. Thesis, Eidgen. Technische Hochschule, Zürich
Maronna RA (1976) Robust M-estimators of multivariate location and scatter. Ann Stat 4:51–67
Rieder H (1994) Robust asymptotic statistics. Springer, Berlin
Tukey JW (1960) A survey of sampling from contaminated distributions. In: Olkin I (ed) Contributions to probability and statistics, Stanford University Press, Stanford

ROC Curves

Lino Sant
Professor, Head of Department of Statistics & Operations Research
University of Malta, Msida, Malta

Classification problems, arising in different forms within various contexts, have stimulated a lot of statistical research with a thread of development stretching back to Fisher's discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues

and Problems) and leading right to the core of statistical learning theory. Along this line ROC (Receiver Operating Characteristic) has come to occupy a privileged position. Weaving within its theory a central role for two classification errors types, it manages to give a statistically sound way of evaluating the diagnostic accuracy of classifier variables.

ROC saw its birth within signal detection theory (Green and Swets 1966). It was cultivated for a time by researchers in psychophysics and later on much promoted within the biomedical sciences (Pepe 2003; Zhou et al. 2002). Interest in the technique and the theoretical tools it offers has extended to many areas these days. The problem it addresses is fairly simple:

A population Π of entities, be they individuals, signal emitters, images, ecosystems, whatever, is made up of two disjoint subpopulations: $\Pi = M \cup N$. An attribute of relevance, say a biomarker like BMI, intensity of an electrical signal, or environmental variable like Air Quality Index, is being measured across both subpopulations. This attribute will be modeled by random variable X with values on a continuous or ordinal scale. F is the probability distribution of X restricted to M with probability density function f, G that restricted to N with density g. The classification problem is that of determining appurtenance to one subpopulation of an object whose X-reading was x. Assuming the mean corresponding to F is smaller than that of G, it is natural to set up some number c and declare the object to belong to M if $x < c$, or to N if $x \geq c$. With reference to the figure below, the location of c determines a number of classification probabilities.

Corresponding to this rule we have two consequences for each decisions: assigning object with value $x < c$ to M incurs the risk of committing an error whose probability is denoted by false negative fraction (*FNF*) $P[X < c|N]$ or else being correct with probability called true negative fraction (*TNF*) $P[X < c|M]$. The "negative" epithet comes from the medical context where F would correspond to a healthy group who are not afflicted by some disease under study. Conversely, assigning object with value $x \geq c$ to N incurs the risk of committing an error, whose probability is called the false positive fraction (*FPF*) $P[X > c|M]$, or else being correct with probability called true positive fraction (*TPF*) $P[X > c|N]$.

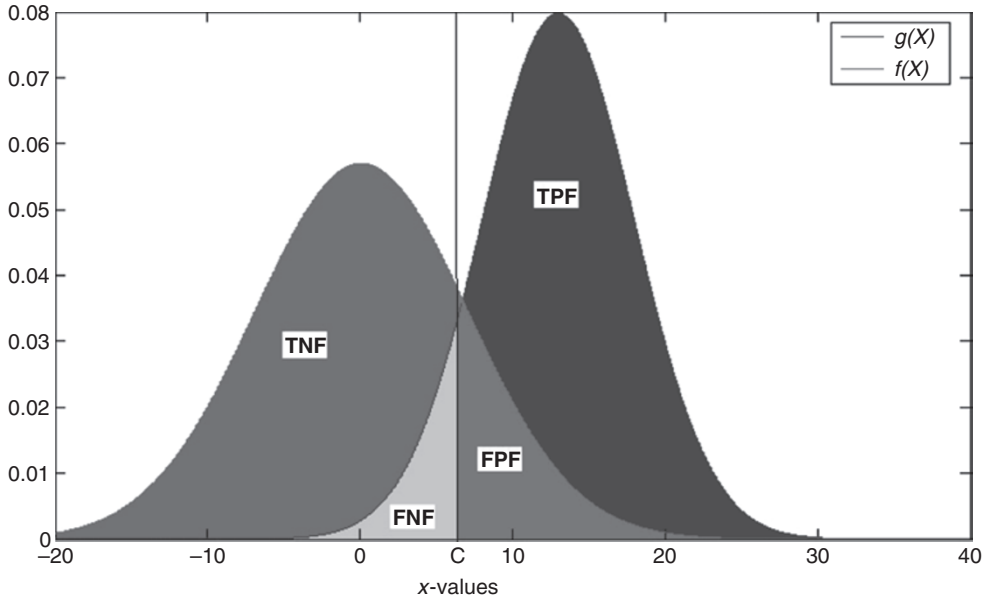
The performance of a classifier variable, in particular its diagnostic accuracy, can be studied in depth by looking at the graph of the *sensitivity* (another name for *TPF*) against the 1-*specificity* (another name for *TNF*) of the classifier for each possible value of c. This graph is called the receiver operating characteristic curve, *ROC*. Using distribution functions and hiding c implicitly we have: $ROC(t) = 1 - G(F^{-1}(1-t))$ for $0 \leq t \leq 1$ and c is given by: $c = G^{-1}(1 - ROC(t))$. A typical *ROC* curve is shown in the figure below.

The higher the graph reaches toward the top left corner the better the classifier behaves. One way of gauging this property is through the area under the curve, denoted *AUC*, and defined as: $AUC = \int_0^1 ROC(t)dt$. This quantity corresponds to the probability that a randomly selected pair of objects, one from each subpopulation, is correctly classified by a test using the classifier. This statistic allows comparisons to be made between classifiers. Classifiers with large AUC are to be preferred. The above analysis can be suitably adapted to random variables with discrete distributions (Figs. 1 and 2).

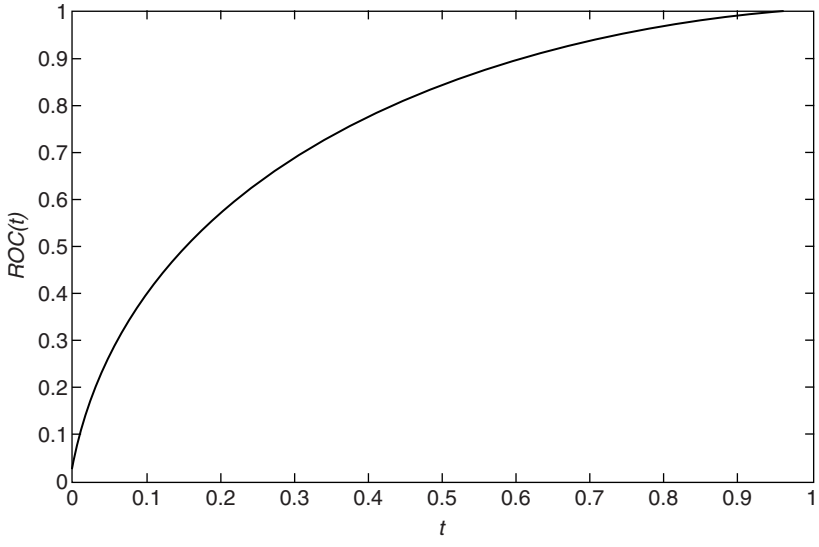
In practice all the population quantities above are not known explicitly. They have to be estimated from actual data. The estimation procedure starts with the procurement of samples of size m (resp. n) selected from subpopulation M (resp. N). The values obtained are used to obtain optimal values of c as well as to compare different classifiers. *ROC* curves can be estimated from such data using a number of techniques which vary across the whole spectrum of estimation techniques. There are parametric methods using known underlying distribution types. The sample from M (resp. N) gives estimates for the corresponding parameters and an *ROC* curve can be derived from the definition above using distribution functions explicitly.

Nonparametric models using empirical distributions are popular in areas where identification of the underlying distributions has not been definitively established. Using results from ▶empirical processes and asymptotic theory a number of very useful statistical results have been obtained for nonparametric models. The most popular method, called the binormal model, is in fact semiparametric. It derives a lot of its sampling distributional results from the Komlós–Major–Tusnády Brownian bridge construction (Hsieh and Turnbull 1996). Though it assumes underlying normal distributions, it can be shown to be valid in cases where the distributions can be transformed to normal distributions. Furthermore the method has shown itself to be robust to departures from normality.

A large number of other estimation techniques have been proposed in the literature like minimum distance and Bayesian estimators. The former are defined relative to some specific metric, or penalty function if you will, on some suitable space of probability distributions. This idea ties up nicely with the hypothesis testing aspect of *ROC* theory. In practice good values of cut-off point c



ROC Curves. Fig. 1 Superimposed graphs for pdf's *f* and *g*



ROC Curves. Fig. 2 A typical ROC graph for continuous distributions

obtained from reliable *ROC* curve estimators would be needed. "Good" varies from one application to the other, but in general it means values which minimize costs related to consequences following from taking the wrong decision, which are tied up to probabilities *FNF* and *TPF*. So in general we need to take care of some penalty function, say linear function: $\alpha_0 + \alpha_{TP}\boldsymbol{P}[TP] + \alpha_{TN}\boldsymbol{P}[TN] + \alpha_{FP}\boldsymbol{P}[FP] + \alpha_{FN}\boldsymbol{P}[FN]$ where the coefficients α_{AB} give the costs corresponding to eventuality *AB*.

ROC was, and still is, extensively used and developed within the biomedical sciences. One important current line of research tries to locate canonical theory within a GLM context. Nevertheless these last thirty years have seen an enormous amount of interest in the technique

amongst computer science researchers interested in disciplines related to statistical classification and machine learning (Krzanowski and Hand 2009).

Cross References

►Discriminant Analysis: An Overview
►Nonparametric Predictive Inference
►Pattern Recognition, Aspects of
►Statistical Signal Processing

References and Further Reading

Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York
Hsieh F, Turnbull BW (1996) Nonparametric and semi-parametric estimation of the receiver operating characteristic curve. Ann Stat 24(1):25–40
Krzanowski WJ, Hand J (2009) ROC curves for continuous data. CRC/Chapman and Hall, Boca Raton
Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. University Press, Oxford
Zhou KH, Obuchowski NA, McClish DK (2002) Statistical methods in diagnostic medicine. Wiley, New York

Role of Statistics

Ashok Sahai[1] , Miodrag Lovric[2]
[1]Professor
St. Augustine Campus of the University of the West Indies at Trinidad, St. Augustine, Trinidad and Tobago
[2]Professor
University of Kragujevac, Kragujevac, Serbia

► *"Modern statistics, like telescopes, microscopes, X-rays, radar, and medical scans, enables us to see things invisible to the naked eye. Modern statistics enables us to see through the mists and confusion of the world about us, to grasp the underlying reality."*

David Hand

Introduction

Despite some recent vigorously promulgated criticisms of statistical methods (particularly significance tests), methodological limitations, and misuses of statistics (see Ziliak and McCloskey 2008; Hurlbert and Lombardi 2009; Marcus 2009; especially Siegfried 2010, among others), we are the ones still "living in the golden age of statistics" (Efron 1998).

Statistics play a vital role in collecting, summarizing, analyzing, and interpreting data in almost all branches of science such as agriculture, astronomy, biology, business, chemistry, economics, education, engineering, genetics, government, medicine, pharmacy, physics, psychology, sociology, etc. Statistical concepts and principles are ubiquitous in science: "as researchers, we use them to design experiments, analyze data, report results, and interpret the published findings of others" (Curran-Everett et al. 1998). Statistical analysis has become an indispensable and fundamental component and vehicle of modern research.

Why is there such a dependence on statistical methods? One of the possible reasons is that statistical thinking has a universal value, as a process that recognizes that variation is present in all phenomena and that the study of variation leads to new knowledge and better decisions. According to Suppes (2007), "the new work, the new concepts, the new efforts, always lead initially, and, often for a long time, to uncertain results. It is…only by an understanding of probability and statistics that a philosopher of science can come to appreciate, in any sort of sophisticated way, the nature of uncertainty that is at the heart of contemporary science…Without statistical methods, it is often impossible to convert the natural, seemingly confused, uncertainty of many results in science into highly probable ones." Straf, in his presidential ASA address (2003), points out that statistics is special "not only because it advances discoveries across the breadth of scientific disciplines and advances the development of technologies, but also because it has an important connection to the human side of scientific and technological development." According to him, the role of statistics is "to increase our understanding, to promote human welfare, and to improve our quality of life and well-being by advancing the discovery and effective use of knowledge from data."

The Importance of Statistics

Since this Encyclopedia contains many entries on the specific role of statistics in different sciences, we will list here only several selected sources underlying the importance of statistics and its versatile usefulness. For obtaining a further appreciation of the role of statistics, the interested reader is referred to those entries, list of references and is urged to "virtually attend" a lecture given by Sir David Cox, by downloading the video file "The Role of Statistics in Science and Technology." Additionally, readers (including all researchers and writers of introductory textbooks on statistics) are advised to read carefully the elucidating paper written by James Brewer (1985) on myths and misconceptions in statistics textbooks.

(a) **Climate research**. Zwiers and Storch (2004) emphasize the importance of statistical methods "for a whole gamut of activities that contribute to the ultimate synthesis of climate knowledge, ranging from the collection of primary data, to the interpretation and analysis of the resulting high-level data sets" (see also ▶Statistics and Climate Change).

(b) **Economics and social studies**. Statistical analysis has proved useful in the solution of a variety of economic problems such as production, consumption, distribution of income and wealth, wages, prices, profits, savings, expenditure, investment, unemployment, poverty, etc. "Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences. This particular dependence of social studies upon statistical methods has led to the unfortunate misapprehension that statistics is to be regarded as a branch of economics, whereas in truth, methods adequate to the treatment of economic data, in so far as these exist, have only been developed in the study of biology and the other sciences" (Fisher 1925).

(c) **Engineering**. According to Johnson et al. (2004, p. 7) "there are few areas where the impact of the recent growth of statistics has been felt more strongly than in engineering and industrial management. Indeed, it would be difficult to overestimate the contributions statistics has made to solving production problems, to the effective use of materials and labor, to basic research, and to the development of new products." Statistics in engineering can be effectively used to solve, for example, the following diversified tasks: "calculating the average length of the downtimes of a computer, collecting and presenting data on the numbers of persons attending seminars on solar energy, evaluating the effectiveness of commercial products, predicting the reliability of a rocket, or studying the vibrations of airplane wings" (op. cit., p. 5).

(d) **Genomics**. Ben-Hui Liu (1998) emphasizes that statistics is a tool to solve problems that cannot be solved only through biological observation or qualitative analysis and that this is especially true for the statistics used in genomic mapping.

(e) **Information systems**. Dudewicz and Karian (1999) indicate that the role of statistics in information systems and information technology in general "can be substantial, yielding more nearly optimal performance of problems at the emerging frontiers in all their aspects."

(f) **Kinetic theory of gases**. Von Mises (1930, p. 207) believes that "not even the tiniest little theorem in the kinetic theory of gases follows from classical physics alone, without additional assumptions of a statistical kind."

(g) **Medical research**. Statisticians are at the "forefront of medical research, helping to produce the evidence for new drugs or discovering links between health and disease and the way we lead our lives" (Oxford Brookes University web site (http://tech.brookes.ac.uk/teaching/pg/msc-in-medical-statistics). According to Sprent (2003) the role of statistics in medical research "starts at the planning stage of a clinical trial or laboratory experiment to establish the design and size of an experiment that will ensure a good prospect of detecting effects of clinical or scientific interest. Statistics is again used during the analysis of data (sample data) to make inferences valid in a wider population." Feinstein (2001) points out that the statistical citation of results has become one of the most common, striking phenomena of modern medical literature (see also ▶Medical Statistics and ▶Medical Research, Statistics in).

(h) **Ophthalmology**. Coleman (2009) believes that statistics play a vital role in "helping us to make decisions about new diagnostic tools and treatments and the care of our patients in the face of uncertainty because, when dealing with patients, we are never 100% certain about an outcome."

(i) **Pharmacogenomics**. Kirkwood (2003) argues that statistical theory and probability will play an expanded role in understanding genetic information through the development of new analytical methodology and the novel application of traditional statistical theory. He points out that "the combination of statistical applications and genomic technologies is a key to understanding the genetic differences that identify patients susceptible to disease, stratify patients by clinical outcome, indicate treatment response, or predict adverse event occurrences."

(j) **Policy and world development**. For example, Moore (1998), in his presidential address to the American Statistical Association (ASA), claimed that it is difficult to think of policy questions that have no statistical component, and argued that statistics is a general and fundamental method because data, variation, and chance are omnipresent in modern life. High-quality statistics also "improve the transparency and accountability of policy making, both of which are essential for good governance, by enabling electorates to judge the success of government policies and

to hold their government to account for those policies… Statistics play a vital role in poverty reduction and world development" (Paris21). However, many developing countries still lack the capacity to produce and analyze good-quality data and use the range of appropriate statistical techniques required to support effective development progress (see also the entries ►Promoting, Fostering and Development of Statistics in Developing Countries, The Role of Statistics – Developing Country Perspective and ►Selection of Appropriate Statistical Methods in Developing Countries).

(k) **Psychiatry**. Hand (1985) believes that statistics has a major role in modern psychiatry, and that "awareness and understanding of statistical concepts is of increasing importance to all psychiatrists, but especially those who wish to advance the field by undertaking research themselves" (see also ►Psychiatry, Statistics in).

(l) **Quality management**. The role of statistics includes control and improvement of the quality of industrial products, during and after the production process through statistical quality control (Srivastava and Rego 2008).

(m) **Quantum theory**. Karl Popper (2002, p. 217) argues that the concept that "quantum theory should be interpreted statistically was suggested by various aspects of the problem situation. Its most important task—the deduction of the atomic spectra—had to be regarded as a statistical task ever since Einstein's hypothesis of photons (or light-quanta)" (see also ►Statistical Inference for Quantum Systems).

(n) **Science**. Magder (2007) points out that the role of statistics in science should be to quantify the strength of evidence in a study so other scientists can integrate the new results with other information to make scientific judgments.

(o) **Seismology**. According to Vere-Jones, one of the pioneers of statistical seismology, "the last decade has seen an influx of new concepts, new data, and new procedures, which combine to make the present time as exciting as any for statistical seismology. New concepts include new mathematical structures, such as self-similarity, fractal growth and dimension and self-organizing criticality, for which existing statistical techniques, based as most of them are on assumptions of stationarity and ergodicity, are inappropriate. In this area at least, seismology is once more challenging the statisticians to enlarge and update their tool box" (Vere-Jones 2006).

(p) **Sociology**. Statistical methods have had a successful half-century in sociology, contributing to a greatly improved standard of scientific rigor in the discipline (Raftery 2001). The overall trend has been toward using more complex statistical methods to match the data, starting from cross-tabulation, measures of association, and log-linear models in the late 1940s; LISREL-type causal models and event-history analysis in the 1960s; and social networks, simulation models, etc. in the late 1980s (see also ►Sociology, Statistics in).

Statistics and Uncertainty

Human life always confronts challenging situations calling for decision-making under uncertainties. While the role of statistics is to minimize the uncertainty associated with the impugned phenomena under investigation, the uncertainty could be measured by the concept of probability. Probability is sine qua non for statistics and statistical modeling, the most important covariate that is omnipresent in all realistic situations challenging the scientists. In fact, the role of statistics encompasses the two fundamentally relevant areas of approximation (any model is an approximation of the real-life phenomenon) and that of optimization (to achieve the minimization of the "gap" between the model and reality).

The role of statistics could, very comprehensibly, be summarized as the "statistical game" being played by the statistician/scientist(s) using statistics against nature as the second player. And this statistical game is quite different from the well-known "zero-sum two person game" in the mathematical setup, inasmuch as the second player is not a conscious player trying to be strategic with the choice of the playing strategies of the first player (statistician/scientist(s) using statistics), and in that the loss incurred by the first player is not a gain for the second player, namely, nature (so that this game is not zero-sum). For example, nature will not cause rains if the statistician/person, guided by weather scientists predicting empirically (statistically), is not carrying an umbrella/raincoat. And vice versa if the person is not carrying the protection, that nature will cause the rain. Nature will cause/not cause the rain if it had to do so for whatever reasons not fully known to us/scientists.

The previous discussion is related to the quantum physics phenomenon. If we go at the microphysics level, as we would attempt to do with the help of a powerful microscope, any matter is not deterministic. Actually, the most decisive conceptual event of twentieth-century physics has been the discovery that the world is not deterministic (Hacking 1990, p. 1). In fact, as we know, at the microscopic level, as to whether or not there will be occupancy/a particle or the absence of it at a particular point in the space

occupied by the relevant matter at any specific point in time, the best physicist in the world, as of today, cannot tell. The best that one could do will be the statement of probability of occupancy reckoned empirically (i.e., based on the experimental data, and that too, only statistically and probabilistically) subject to approximation error.

Conclusion and Recommendation

We agree with Provost and Norman (1990, p. 43) that the 21st century will place even greater demands on society for statistical thinking throughout industry, government, and education.

However, if statistics aspire to be an essential element in the description and understanding of the actual phenomena in the world around us, it is an imperative that we, statisticians, conduct a critical evaluation of statistics in the first place. To achieve that, we need to begin with building a bridge between Bayesians and frequentists, may be with the help of a new Ronald Fisher. Equally importantly, we need to find a way to explain more clearly the usage of statistical methods, along with their advantages and disadvantages, to overcome the generalized confusion in the public and among many researchers over many statistical issues and also to educate statistical practitioners at all levels.

> *"A chisel in a skillful artist's hand can produce a beautiful sculpture and a scalpel in an experienced surgeon's hand can save a person's life. Similarly, statistical techniques used properly by an honest and knowledgeable scientist can be equally impressive at illuminating complex phenomena, thus promoting scientific understanding, and shortening the time between scientific discovery and its impact on societal problems. If misused, they can produce the counterproductive results… Such erroneous results, however, should not be viewed as a failing of Statistics"*
>
> (ASA unedited letter to the editor of the *Science News* in response to the "Odds Are, It's Wrong" paper.)

Acknowledgment

Professor Sahai dedicates this write-up to the fond memory of his most beloved and inspiring Professor, Late Dr. A. R. Roy (Stanford University).

Cross References

►Careers in Statistics
►Environmental Monitoring, Statistics Role in
►Frequentist Hypothesis Testing: A Defense
►Medical Research, Statistics in
►Misuse of Statistics
►Null-Hypothesis Significance Testing: Misconceptions
►Rise of Statistics in the Twenty First Century
►Role of Statistics in Advancing Quantitative Education
►Role of Statistics: Developing Country Perspective
►Statistics: An Overview

References and Further Reading

Brewer JK (1985) Behavioral statistics textbooks: source of myths and misconceptions? J Edu Stat, 10(3) Available at: http://www.jstor.org/stable/1164796 (Special issue: Teaching statistics)

Coleman AL (2009) The role of statistics in ophthalmology. Am J Ophthalmol 147(3):387–388

Cox D (2008) The role of statistics in science and technology. Video file, available at: http://video.google.com/videoplay?docid=1739298413105326425#

Curran-Everett D, Taylor S, Kafadar K (1998) Fundamental concepts in statistics: elucidation and illustration. J Appl Physiol 85: 775–786

Dudewicz EJ, Karian ZA (1999) The role of statistics in IS/IT: practical gains from mined data. Inform Syst Frontiers 1(3):259–266

Efron B (1998) R. A. Fisher in the 21st century. Stat Sci 13(2):95–122

Feinstein AR (2001) Principles of medical statistics. Chapman and Hall/CRC, London

Fisher R (1925) Statistical methods for research workers, Oliver and Boyd, Edinburgh

Hacking I (1990) The taming of chance. Cambridge University Press, Cambridge

Hand D (1985) The role of statistics in psychiatry. Psychol Med 15:471–476

Hand D (2008) Statistics: a very short introduction. Oxford University Press, Oxford

Hurlbert SH, Lombardi CM (2009) Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Ann Zool Fenn 46:311–349

Johnson R, Miller I, Freund J (2004) Miller & Freund's probability and statistics for engineers. 7th edn. Prentice Hall, Englewood Cliffs, NJ

Kirkwood SC (2003) The role of statistics in pharmacogenomics. J Japan Soc Comp Stat 15(2):3–13

Liu BH (1998) Statistical genomics: linkage, mapping, and QTL analysis. CRC Press, Boca Raton, p x

Magder L (2007) Against statistical inference: a commentary on the role of statistics in public health research, The 135th APHA annual meeting & exposition of APHA, Washington DC

Marcus A (2009) Fraud case rocks anesthesiology community: Mass. researcher implicated in falsification of data, other misdeeds. Anesthesiology News, 35, 3

Moore DS (1998) Statistics among the liberal arts. J Am Stat Assoc 93(444):1253–1259

Morrison D, Henkel R (eds) (2006) The significance test controversy: a reader. Aldine transaction, Piscataway, USA (reprint)

Paris21 (The partnership in statistics for development in the 21st Century) Counting down poverty: the role of statistics in world development. Available at http://www.paris21.org/documents/2532.pdf

Popper K (2002) The logic of scientific discovery. (trans: Logik der Forschung, Vienna, 1934). Routledge, London

Provost LP, Norman CL (1990) Variation through the ages. Quality Progress Special Issue on Variation, ASQC
Raftery AE (2001) Statistics in sociology, 1950–2000: a selective review. Sociol Methodol 31(1):1–45
Siegfried T (2010) Odds are, it's wrong: science fails to face the shortcomings of statistics, Science News, 177, 26
Sprent P (2003) Statistics in medical research. Swiss Med Wkly. 133(39–40), 522–529
Srivastava TN, Rego S (2008) Statistics for management, Tata McGraw Hill, New Delhi
Straf ML (2003) Statistics: the next generation. J Am Stat Assoc 98:461 (Presidential address)
Suppes P (2007) Statistical concepts in philosophy of science. Synthese 154:485–496
v Mises R (1930) Über kausale und statistische Gesetzmäßigkeit in der Physik. Die Naturwissenschaften 18(7):145–153
Vere-Jones D (2006) The development of statistical seismology: a personal experience. Tectonophysics 413:5–12
Ziliak ST, McCloskey DN (2008) The cult of statistical significance: how the standard error costs us jobs, justice, and lives. University of Michigan Press
Zwiers FW, Storch HV (2004) On the role of statistics in climate research. Int J Climatol 24:665–680

Role of Statistics in Advancing Quantitative Education

James J. Cochran
College of Business, Louisiana Tech University, Ruston, LA, USA

Introduction

Education policy makers and educators from all disciplines generally agree that as data have become more plentiful and more readily available, the importance of statistical literacy has grown (Utts 2003; Garfield and Ben-Zvi 2008). Since trends in availability of data show no signs of abating (indicators actually point to accelerating increases in data availability), the importance of teaching statistical thinking to students at all levels is difficult to overstate or overestimate. It is not an exaggeration to state that statistical/quantitative literacy is almost as critical to the future success of students as is reading literacy (Steen 2002). Thus, it is vitally important that both society and the statistics community understand the vital role of statistics in quantitative education.

Statistics' Role in Quantitative Education

Several phrases are used somewhat interchangeably in reference to an individual's ability to work with numbers and relationships and understand the implications of her or his results. While the generally accepted definitions of these phrases overlap, there are subtle but important differences. These phrases (and generic versions of their generally acceptable definitions) include:

- Numeracy – this phrase, first used in the 1959 Crowther Report on education in the United Kingdom (Jarman 1960), comprises the aptitude to use reason to solve sophisticated quantitative problems; a fundamental understanding of the scientific method; and the ability to communicate with others about everyday quantitative issues, questions, and concerns. In explaining this phrase, Steen (1990) wrote:
  - Numeracy is to mathematics as literacy is to language. Each represents a distinctive means of communication that is indispensable to civilized life.
- Quantitative Literacy – this phrase refers to minimal levels of comfort with, competency in, and disposition toward working with numerical data and concepts necessary to function at a reasonable level in society.
- Quantitative Reasoning – this phrase represents the manifestation of basic logic applied by an individual to the construction of rigorous and valid arguments as well as the evaluation of the rigor and validity of arguments made by others. Thus, quantitative reasoning emphasizes the higher-order analytic and critical thinking skills needed to understand, create, and cope with sophisticated arguments (which are frequently supported by quantitative data).

See Madison and Steen (2008) for a brief history of the evolution of these terms.

If one considers these three concepts to be the cornerstones of quantitative education, then their meanings must provide the basis of a broad definition of quantitative education. The definition of quantitative education that results from this perspective is *the effort to imbue students with numeracy, quantitative literacy, and quantitative reasoning.*

While quantitative education certainly subsumes statistics education, statistics education is without question a vital and critical component of quantitative education. The strict association most individuals place on quantitative literacy and mathematics, in combination with the manner in which mathematics is generally taught at lower levels of education, generates an overwhelming and ill-advised emphasis on thinking of quantitative issues deterministically. This emphasis on a deterministic treatment of quantitative problems reinforces the notion that a problem has a single correct answer and cultivates the more damaging conclusion that there is a single correct way to solve

a problem, which ultimately robs the student of the opportunity to fully comprehend and appreciate the nature of quantitative problems and problem solving. This issue can be confronted directly through the integration of statistics (and probability) into quantitative education at all levels. Integration of statistics and probability into quantitative education can also be used to address the common misconceptions that (1) quantitative concepts can only be memorized and cannot be understood by average students; (2) quantitative concepts have little relevance to everyday life; (3) quantitative approaches to problem solving lead to conclusive and consistent conclusions; and (4) quantitative analysis is a solitary activity to be pursued by individuals in isolation.

Florence Nightingale (Cook and Nash 1936) expressed her explicit recognition and appreciation of statistics' role in quantitative education when she stated:

> Statistics is the most important science in the whole world, for upon it depends the practical application of every other science and of every art; the one science essential to all political and social administration, all education, all organization based on experience, for it only gives results of our experience.

and

> To understand God's thoughts we must study statistics, for these are the measure of his purpose.

As a critically important component of quantitative education, statistics educators should strive to encourage students to develop numeracy, quantitative literacy, and quantitative reasoning skills. Thus, it is vitally important that statistical education focus on the components of these three objectives.

Important Objectives of Statistics Education and their Links to Quantitative Education

Statistics education contributes to quantitative education through its strong emphasis on the development of numeracy, quantitative literacy, and quantitative reasoning skills. Statistics education can address the various components of numeracy, quantitative literacy, and quantitative reasoning skills in a myriad of ways. In the following sections the author discusses several ways in which statistics education naturally and organically encourages students to develop each component of numeracy, quantitative literacy, and quantitative reasoning skills (identified in the preceding definitions).

Aptitude to use Reason to Solve Sophisticated Quantitative Problems

Statistics instructors have abundant opportunities to help students develop their aptitude to use reason to solve sophisticated quantitative problems. To solve problems in statistics, students must appreciate the fundamental difference between certainty and uncertainty as well as the ramifications of uncertainty; development of this appreciation certainly fosters the aptitude to use reason in resolving sophisticated quantitative problems. The student must use sophisticated logic and reason to understand the p-value as a conditional probability (the probability of taking a sample in a prescribed manner and collecting results at least as counter to the null hypothesis *given that the condition that the null hypothesis is true*). Similarly, comprehension of the distinctions between conditional and joint probability, precision and accuracy, and experimentation and observation also require extensive use of sophisticated reasoning and logic.

Fundamental Understanding of the Scientific Method

The *scientific method* is the logical and rational order of steps by which a scientist analytically and rigorously tests a conjecture and reaches a conclusion while minimizing the biases s/he introduces into a scientific inquiry. *Statistical inference* is unquestionably an overt embodiment of the scientific method. Indeed, the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE 2005) report of the American Statistical Association maintains that statistics is a problem-solving process that comprises four steps

- Formulation of questions – developing hypotheses and selecting appropriate analytic methods and decision making criteria;
- Data collection – making decisions on what data to collect and how to collect the data, as well as executing the actual process of collecting data;
- Data analysis – using descriptive and inferential approaches that lead to an understanding of what the sample data that have been collected can reveal about their population; and
- Interpretation of results – disseminating and explaining the results to the appropriate audience, considering implementation issues, and making suggestions to make similar future efforts more scientifically sound.

These four steps and the steps of the scientific method have a bijective relationship (this is why many consider statistics to be the purest science). By stressing this relationship in statistics courses, statistics educators can help students

understand the relevance of the scientific method to their lives. Furthermore, statistics educators have opportunities to reinforce the appreciation of the scientific method through coverage of the justifications for sampling, as it is the use of sample data in lieu of census data that necessitates the use of the scientific method.

Ability to Communicate about Everyday Quantitative Issues, Questions, and Concerns

Because statistics students are generally learning about concepts and ideas that are sophisticated and unfamiliar to them, statistics courses provide natural environments for the nurturing of communications skills. As a statistics instructor works with students to enable them to connect with and understand statistical concepts, s/he can also stress the importance of students' attempts to emulate the instructor's efforts to communicate; through these efforts students can develop the ability to explain statistical concepts, methodologies, and results with individuals who are unfamiliar with and intimidated by statistics. Students will naturally embrace this skill once they understand its desirability and marketability. A student who can correctly explain the underlying principle of statistical inference in an understandable manner or clarify the distinction between the concepts of association and causality, replication and repeated measurement, or parametric and nonparametric approaches will have great advantages in her/his academic and career pursuits.

Minimal Levels of Comfort with, Competency in, and Disposition toward Working with Numerical Data and Concepts Necessary to Function in Society

A sound background in basic statistics provides an individual with an important level of self-sufficiency with respect to numerical data and concepts. For example, statistics provides its users with systematic methods for dealing with variation; the quantitatively literate individual not only appreciates the ubiquity of variation but also understands the need to consider variation when interpreting observed phenomena and making decisions. Such an individual is capable of quantifying and explaining variability; she or he also recognizes that variability can be the result of patterns in the values of the variable of interest, relationships between the variable of interest and other variables, and/or randomness. Understanding of these notions leads directly to an understanding of randomness (and its importance) as well as the distinction between experimentation and observation (and the associated ramifications).

While every concept covered in an introductory statistics course provides statistics students with an occasion to become more comfortable with numerical concepts, perhaps none presents a greater opportunity (and challenge) than the notion of a central limit theorem (see ▶Central Limit Theorems). An instructor will surely fail to communicate with all but the most highly motivated students by explaining that:

- A central limit theorem is any weak-convergence theorem that expresses the tendency for a sum of several independent identically distributed random variables with a positive variance to converge in distribution to a member of a known and predictable family of distributions.

On the other hand, a statistics instructor can open her/his students' eyes to the elegance and beauty of this concept if s/he explains instead that:

- One version of the central limit theorem states that given a sample is taken from a population whose distribution has mean μ and variance σ^2, the distribution of the potential values of the resulting sample mean $\bar{X}$ approaches a normal distribution with a mean $\mu_{\bar{X}} = \mu$ and a variance $\sigma_{\bar{X}}{}^2 = {}^{\sigma^2}/_{n}$ as the sample size n increases.

The second explanation allows the statistics instructor to further elaborate on how the probability of collecting a sample with an extreme mean decreases rapidly as the sample size increases because an extreme sample mean can only result from a sample that consists primarily of extreme observations, and the probability of collecting a sample that consists primarily of extreme observations decreases rapidly as the number of observations in the sample increases. This explanation both appeals to and nurtures the student's levels of comfort with, competency in, and disposition toward working with numerical data and concepts.

Development/Enhancement of the Ability to Use the Higher-Order Analytic and Critical Thinking Skills Needed to Understand and Create Sophisticated Arguments

In properly designing and executing a statistical investigation, one must use higher-order analytic and critical thinking skills to create quantitatively and logically sophisticated arguments. Thus, by teaching the process of statistical investigation, a statistics instructor is implicitly assisting the student in the development and enhancement of these higher-order analytic and critical thinking skills. For example, because students generally think of mathematics deterministically, they tend to adhere to the rhetorical tactic of using examples to support an argument (Lawton 2009; Sotos et al. 2009). Because of the uncertainty

embedded in sample data (and so is embedded in any statistical investigation), one cannot use examples to support a null hypothesis; sample data is evaluated in terms of the strength of the evidence it provides *against the null hypothesis*. This distinction, between logical/rhetorical and empirical arguments, is initially difficult for students to comprehend. However, with clear and cogent explanations the statistics instructor can help the student understand the sophisticated argument behind this distinction; this certainly constitutes the development and enhancement of these higher-order analytic and critical thinking skills.

In another example, consider the introductory statistics student's strong tendency to fall victim to the *cum hoc ergo propter hoc* fallacy. When these students find a strong correlation between two random variables, they often immediately infer that a causal relationship exists between these two variables. Enhancement of their higher-order analytic and critical thinking skills is necessary to facilitate their understanding that correlation is a necessary but not sufficient condition for causality. The students must further refine these skills in order to achieve an understanding of the concepts of spurious correlation, reverse causation, two way causation, and common causal variables (examples of which can be found throughout the popular media). Thus, through the development of the ability to properly interpret the results of a statistical analysis (in this case, a simple correlation), a student is enhancing her/his ability to use the higher-order analytic and critical thinking skills needed to understand and create sophisticated arguments.

Conclusions

Statistics education has a critical role in each of the primary components of quantitative education. Through statistics education students can become more numerate and strengthen their analytic and critical thinking skills. Statistics instructors can and should increase their students' and the public's appreciation for statistics by closely aligning their course objectives with the broad definition of quantitative education.

About the Author

Dr. James J. Cochran is Associate Professor (the Bank of Ruston Endowed Research Professor) of Quantitative Analysis and Computational Modeling; Senior Scientist, Center For Information Assurance; and Senior Scientist/Analytic Group Director, Center For Secure Cyberspace at Louisiana Tech University. He has previously been on the faculty at Wright State University, Drexel University, Miami University, and the University of Cincinnati. He has also held the position of Visiting Scholar with Stanford University, the University of South Africa, and the Universidad de Talca. Professor Cochran has published over thirty articles in statistics and operations research journals. He is a Coeditor of the *Anthology of Statistics in Sports*, and is the founding Editor-in-Chief of the *Wiley Encyclopedia of Operations Research and Management Science*. Professor Cochran is also the Editor-in-Chief of *INFORMS Transactions on Education*. Dr Cochran was General Chair of the 2005 INFORMS Conference, and President of INFORM-ED (the INFORMS Education Forum) from 2002–2005 and the founding Chair of the INFORMS Section on OR in SpORts from 2004–2008. He has received the INFORMS Prize for the Teaching of OR/MS Practice and the American Statistical Association's Significant Contributor to Statistics in Sports Award (both in 2008). He established and has organized INFORMS' Teaching Effectiveness Colloquium series and annual Case Competition as well as the annual INFORMS/IFORS International Education Workshop series. He is also a founding co-chair of *Statistics Without Borders*, and was elected to the International Statistics Institute in 2005.

Cross References

►Careers in Statistics
►Online Statistics Education
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Statistical Literacy, Reasoning, and Thinking
►Statistics Education

References and Further Reading

Cook ET, Nash RN (1936) A Short Life of Florence Nightingale. Macmillan, New York

GAISE (2005) Guidelines for assessment and instruction in statistics education. Retrieved November 2, 2009 from www.amstat.org/Education/gaise/GAISECollege.htm

Garfield JB, Ben-Zvi D (2008) Developing students' statistical reasoning: connecting research and teaching practice. Springer, New York

Jarman TL (1960) Developments in English Education in 1959: The year of the Crowther Report. Internationale Zeitschrift für Erziehungswissenschaft/Revue Internationale l' Éducation 6(1):231–234

Lawton L (2009) An exercise for illustrating the logic of hypothesis testing. J Stat Educ 17(2). Available at: www.amstat.org/publications/jse/v17n2/lawton.html

Madison BL, Steen LA (2008) Evolution of numeracy and the national numeracy network. Numeracy 1(1). Available at: http://services.bepress.com/numeracy/vol1/iss1/art2

Sotos AEC, Vanhoof S, Van den Noortgate W, Onghena P (2009) How confident are students in their misconceptions about hypothesis tests? J Stat Educ 17(2). Available at: www.amstat.org/publications/jse/v17n2/castrosotos.html

Steen LA (1990) Numeracy. Daedalus 119(2):211–231
Steen LA (2002) Quantitative literacy: why numeracy matters for schools and colleges. Focus 22(2):8–9
Utts J (2003) What educated citizens should know about statistics and probability. Am Stat 57(2):74–79

Role of Statistics: Developing Country Perspective

Dimitri Sanga
Acting Director
African Centre for Statistics, Addis Ababa, Ethiopia

Statistics: A Part of the Enabling Environment for Development

The main role of statistical development is to help National Statistical Systems (NSS) efficiently produce good statistics. Good statistics are characterized, *inter alia*, by the quality (reliability, accuracy, accessibility, timeliness, etc.) with which they are produced. They are said to be good only to the extent that they meet users' needs. They need to be available to a broad range of public and private users and be trusted to be objective and reliable. In addition, they must meet all policy needs and inform the public so that the latter can evaluate the effectiveness of government's actions.

Good statistics are needed to assess, identify issues, support the choice of interventions, forecast the future, monitor progress and evaluate the results and impacts of policies and programmes. They provide a basis for good decision-making, support governments in identifying the best courses of action in addressing problems, are essential to manage the effective delivery of basic services, and are indispensable for accountability and transparency. They are also essential for providing a sound basis for the design, management, monitoring, and evaluation of national policy frameworks such as the Poverty Reduction Strategies (PRSs) and for monitoring progress towards national, sub regional, regional, and international development goals including the Millennium Development Goals initiatives (MDGs). Accordingly, good statistics are considered to be part of the enabling environment for development.

Initiatives Aimed at Supporting Developing Countries in Statistics

In recognition of the importance of statistics in their development process, developing countries have been struggling to provide their users with quality statistical information. However, the last decade of the twentieth century has witnessed an unprecedented increase in the demand for quality and timely statistics following an emergence of initiatives aimed at tackling development issues including those enshrined in the Millennium Declaration. In fact, there is increasing recognition that the successful implementation, evaluation, and monitoring of national, sub regional, regional, and international development agendas rely on the production and use of quality statistics. This has challenged already weak and vulnerable NSSs in developing countries.

In response to this challenge, several initiatives have been launched at the international level to support developing countries to meet their respective users' needs. Among these is the Marrakech Action Plan for Statistics (MAPS) adopted in 2004 during the Second Round Table on Managing for Development Results. It consists of a global agenda aimed at improving the availability and use of quality statistics in support of PRSs according to a well-defined budget covering a specific period of time. The MAPS recommends, *inter alia*, that every low-income country designs and implements a National Strategy for the Development of Statistics (NSDS) aimed at providing the country with strategic orientations and appropriate mechanisms to guide and accelerate the development of its statistical capacity in a sustainable manner.

Some Issues and Challenges Facing Developing Countries

Key issues and problems confronting statistical development in developing countries include: inadequate political commitment to statistical development especially at the national level; limited coordination, collaboration, networking and information sharing among stakeholders; inadequate resources (financial, human, and technical) for statistical production; inadequate infrastructure (physical and statistical) for statistical production; limited institutional capacity; poor quality data; inadequate data management (archiving, analysis, and dissemination) systems; lack of long-term planning for statistical development; and inappropriate profiles of National Statistical Offices (NSOs) in government hierarchy.

In this context, those delivering NSSs in developing countries face specific challenges including: creating greater awareness among data users and especially planners, policy makers and decision makers about the strategic importance of statistics in their work, particularly in evidence-based macro-economic management, policy formulation and poverty measurement and monitoring; playing an advocacy role to ensure that statistical production and use are given high priority by national governments; building ample capacity to make user needs assessments for data of improved quality and keep abreast of the data needs of policy makers, the private sector and civil society;

building capacity to harness technology and to improve the way data are collected and disseminated to users; building competent user groups (policy makers, researchers, media) to properly understand and interpret available statistical data; building competence in Survey Management in NSOs; and promoting co-ordination and synergy among institutions involved in statistical activities.

Conclusion

Several efforts are being made at international, regional, sub-regional and national levels to support NSS of developing countries. In spite of these efforts, the majority of developing countries still do not have statistical systems that are capable of providing, in a sustainable manner, good statistical data and information required for evidence-based planning and policy formulation, democratic governance and accountability, political and personal decisions. It is therefore imperative that those supporting statistical development efforts in developing countries address the above-mentioned challenges to help statistics play their role of enablers of development.

Acknowledgments

The views expressed in this paper are personal to the author and do not necessarily represent those of the United Nations Economic Commission for Africa or its subsidiary organs.

About the Author

Dr. Dimitri Sanga is currently the Acting Director of the African Centre for Statistics (ACS) at the United Nations Economic Commission for Africa (ECA). Until end of July 2009, he was Chief of the Statistical Development and Data Management Section of the ACS. In this capacity, and formerly, he contributed to the revamping of the statistical function at ECA and most notably the inception of ACS. Before joining the United Nations, he served as Senior Economic Statistician at Statistics Canada, occupying several posts in areas such as price statistics, national accounts and household surveys undertaking and analysis. He was also part time Professor of economics, econometrics, and statistics in a number of Canadian universities namely Laval, Sherbrooke, and Ottawa. He has substantively published in refereed journals and produced a number of textbooks in economics with special interest in index number theory and practices. An elected member of the International Statistical Institute (ISI), he currently serves on the Editorial Board of the *African Statistical Journal* and the *African Statistical Newsletter*. He is also member of several international expert groups including the 2010 United Nations Expert Group on Population and Housing Censuses, the Inter Agency and Expert Group on the Millennium Development Goals Indicators, and the Inter Agency and Expert Group on Gender Statistics. He received the joint Natural Resources of Canada and Groupe de recherche en économie de l'énergie, de l'environnement et des ressources naturelles de l'Université Laval (GREEN) awards for 1993, 1994, 1995, 1996, 1997, and 1998 successively.

Cross References

►African Population Censuses
►Aggregation Schemes
►Careers in Statistics
►Economic Growth and Well-Being: Statistical Perspective
►Measurement of Economic Progress
►Promoting, Fostering and Development of Statistics in Developing Countries
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Role of Statistics in Advancing Quantitative Education
►Selection of Appropriate Statistical Methods in Developing Countries
►Statistics and Climate Change

Rubin Causal Model

Donald B. Rubin
John L. Loeb Professor of Statistics
Harvard University, Cambridge, MA, USA

The Rubin Causal Model (RCM) is a formal mathematical framework for causal inference, first given that name by Holland (1986) for a series of previous articles developing the perspective (Rubin 1974, 1975, 1976, 1977, 1978, 1979, 1980). This framework, as formulated in these articles, has two essential parts (the definition of causal effects using the concept of potential outcomes and the definition of a model for the assignment mechanism) and one optional part (which involves the specification of a model for quantities treated as fixed by the first two parts). These three parts are briefly described, emphasizing the implications for practice. A longer encyclopedic entry on the RCM is Imbens and Rubin (2008), chapter length summaries are included in Rubin (2006, 2008) and a full-length text from this perspective is Imbens and Rubin (2010).

The first part is conceptual and implies that we should always start an inquiry into a causal question by carefully defining all causal estimands (quantities to be estimated)

in terms of potential outcomes, which are all values that could be observed in some real or hypothetical experiment comparing the results under a well-defined active treatment versus the results under a well-defined control treatment in a population of units, each of which can be exposed, in principle, to either treatment. That is, causal effects are defined by comparisons of (a) the values that would be observed if the active treatment were applied and (b) the values that would be observed if instead the control treatment were applied to this population of units. This step contrasts with the common practice of defining causal effects in observational studies in terms of parameters in some regression model, where the manipulations defining the active versus control treatments are often left implicit and ill-defined, with the resulting causal inferences correspondingly ambiguous. See, for example, the discussions by Mealli and Rubin (2003) and Angrist et al. (1996). This first step of the RCM can take place before any data are observed or even collected. The set of potential outcomes under the active treatment and the control treatment defines the "science" – the scientific objective of causal inference in all studies, whether randomized [see the entries on experiments by Hinkelman (2010) and Cox (2010)], observational or entirely hypothetical. It appears that the first use of the formal concept of potential outcomes to define causal effects was Neyman (1923) in the context of randomization-based inference in randomized experiments, but this notation was not extended to nonrandomized studies until Rubin (1974); Rubin (2010) provides some historical perspective. The science also includes covariates (background variables) that describe the units in the population.

The second part of the RCM, the formulation of the assignment mechanism, implies that after having defined the science, we should continue by explicating the design of the real or hypothetical study being used to estimate that science. The assignment mechanism mathematically describes why some study units will be (or were) exposed to the active treatment and why other study units will be (or were) exposed to the control treatment. Sometimes the assignment mechanism involves the consideration of background (i.e., pre-treatment) variables for the purpose of creating strata of similar units to be exposed to the active treatment and the control treatment, thereby improving the balance of treatment and control groups with respect to these background variables (i.e., covariates). A true experiment automatically cannot use any outcome (post-assignment) variables to influence design because they are not yet observed. If the observed data were not generated by a true experiment, but rather by an assignment mechanism corresponding to a nonrandomized observational data set, there still should be an explicit design phase. That is, in an observational study, the same guidelines as in an experiment should be followed.

More explicitly, the design step in the analysis of an observational data set for causal inference should structure the data to approximate (or reconstruct or replicate) a true randomized experiment as closely as possible. In this design step, the researcher never uses or even examines any final outcome data, but rather, identifies subsets of units such that the active and control treatments can be thought of as being randomly assigned within the subsets. This assumed randomness of treatment assignment is assessed by examining, within these subsets of units, the similarity of the distributions of the observed covariates in the treatment group and in the control group.

The third part of the RCM is optional; it derives inferences for causal effects from the observed data by conceptualizing the problem as one of imputing the missing potential outcomes. That is, once some outcome data are observed (i.e., observations of the potential outcomes corresponding to the treatments actually received by the various units), then the modeling of the outcome data given the covariates should be structured to derive predictive distributions of those potential outcomes that would have been observed if the treatment assignments had been different. This modeling generates stochastic predictions (i.e., imputations) for all missing potential outcomes in the study, which, when combined with the actually observed potential outcomes, allows the calculation of causal–effect estimands. Because the imputations of the missing potential outcomes are stochastic, repeating the process results in different values for the causal–effect estimands. This variation across the multiple imputations (Rubin 1987, 2004) generates interval estimates and tests for the causal estimands. Typically in practice, this third part is implemented using simulation-based methods, such as Markov chain Monte Carlo computation (see ►Markov Chain Monte Carlo) applied to calculate posterior distributions under Bayesian models.

The conceptual clarity in the first two parts of the RCM often allows previously difficult causal inference situations to be easily formulated. The optional third part often extends this successful formulating by relying on modern computational power to handle analytically intractable problems. Recent, albeit somewhat idiosyncratic, examples include Hirano et al. (2000), Jin and Rubin (2008), and Rubin and Zell (2010).

About the Author

Professor Donald B. Rubin is John L. Loeb Professor of Statistics, Department of Statistics, Harvard University. He was Chair of the department during 1985–1994 and 2000–2004. He is an Elected Fellow/Member of: the American

Statistical Association (1977), the Institute of Mathematical Statistics (1977), the International Statistical Institute (1984), the American Association for the Advancement of Science (1984), the American Academy of Arts and Sciences (1993), and the National Academy of Sciences (2010). He has authored/coauthored over 350 publications (including 10 books) and has made important contributions to statistical theory and methodology, particularly in causal inference, design and analysis of experiments and sample surveys, treatment of missing data, and Bayesian data analysis. Among his many awards, Professor Rubin has received the Samuel S. Wilks Medal (American Statistical Association, 1995), the Parzen Prize for Statistical Innovation (1996), the Fisher Lectureship (2004), and the George W. Snedecor Award of the Committee of Presidents of Statistical Societies (2007). He was named Statistician of the Year, American Statistical Association, Boston Chapter (1995); and Statistician of the Year, American Statistical Association, Chicago Chapter (2001). He was Associate Editor for: *Journal of Educational Statistics* (1976–1979), *Theory and Methods, The Journal of American Statistical Association* (1975–1979), Editor Elect, *The Journal of American Statistical Association* (1979), Coordinating Editor and Applications Editor, *The Journal of American Statistical Association* (1980–1982), *Biometrika* (1992–1995), *Survey Methodology* (1988–1993), *Statistica Sinica* (1993-2004). Professor Rubin has been, for many years, one of the mostly cited authors in mathematics in the world (according to ISI Science Watch); in 2002 he was ranked Seventh in the World for the Decade 1991–2000. His biography is included in many places including Who's Who in The World. In 2008 Professor Rubin was elected a Honorary Member of the European Association of Methodology, and in 2009 he was elected a Corresponding (foreign) Fellow of the British Academy.

Cross References

►Causal Diagrams
►Causation and Causal Inference
►Design of Experiments: A Pattern of Progress
►Experimental Design: An Introduction
►Imputation
►Markov Chain Monte Carlo
►Multiple Imputation
►Randomization

References and Further Reading

Angrist J, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables (with discussion). *J Am Stat Assoc*, Applications Invited Discussion Article with discussion and rejoinder 91(434):444–472

Cox DR (2010) Design of experiments: a pattern of progress (this volume)

Hinkelman K (2010) Introduction to experimental design. (this volume)

Hirano K, Imbens G, Rubin DB, Zhou X (2000) Estimating the effect of an influenza vaccine in an encouragement design. Biostatistics 1:69–88

Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71:1161–1189

Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81:945–970

Imbens GW, Rubin DB (2010) Causal Inference in Statistics and the Medical and Social Sciences. Cambridge University Press, Cambridge, U.K

Jin H, Rubin DB (2008) Principal stratification for causal inference with extended partial compliance: application to Efron-Feldman data. J Am Stat Assoc 103:101–111

Mealli F, Rubin DB (2003) Commentary: assumptions allowing the estimation of direct causal effects, J Economet 112: 79–87

Neyman J (1923) On the application of probability theory to agricultural experiments: essay on principles, section 9. Translated in Statistical Science 5(465–480):1990

Neyman J (1935) Statistical problems in agricultural experimentation. Supplement to J R Stat Soc B 2:107–108 (with discussion). (With cooperation of K. Kwaskiewicz and St. Kolodziejczyk)

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66: 688–701

Rubin DB (1975) Bayesian inference for causality: the importance of randomization. Proceedings of the Social Statistics Section of the American Statistical Association, pp 233–239

Rubin DB (1976) Inference and missing data. Biometrika 63: 581–592

Rubin DB (1977) Assignment of treatment group on the basis of a covariate. J Educ Stat 2:1–26

Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6:34–58

Rubin DB (1979) Discussion of "conditional independence in statistical theory," by A.P. Dawid. J R Stat Soc B 41:27–28

Rubin DB (1980) Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. J Am Stat Assoc 75:591–593

Rubin DB (1987) (2004) Multiple Imputation for Nonresponse in Surveys. 1st edn. and Wiley, Classic edn. Wiley, New York

Rubin DB (2006) Statistical inference for causal effects, with emphasis on applications in psychometrics and education. In: Rao CR, Sinharay S (eds) Handbook of Statistics, Vol. 26: Psychometrics. Elsevier, The Netherlands, Chapter 24, pp 769–800

Rubin DB (2008) Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. In: Rao CR, Miller JP, Rao DC (eds) Handbook of statistics: epidemiology and medical statistics, Chapter 2. Elsevier, The Netherlands, pp 28–63

Rubin DB (2010) Reflections stimulated by the comments of Shadish (2009) and West and Thoemmes. Psychol Methods 15(1): 38–46

Rubin DB and Zell ER (2010) Dealing with noncompliance and missing outcomes in a randomized trial using Bayesian technology: prevention of perinatal sepsis clinical trial, Soweto, South Africa. Stat Methodol 7(3):338–350

S

Saddlepoint Approximations

Juan Carlos Abril
President of the Argentinean Statistical Society, Professor Universidad Nacional de Tucumán and Consejo Nacional de Investigaciones Científicas y Técnicas, San Miguel de Tucumán, Argentina

Introduction

It is often required to approximate to the distribution of some statistics whose exact distribution cannot be conveniently obtained. When the first few moments are known, a common procedure is to fit a law of the Edgeworth type having the same moments as far as they are given. This method is often satisfactory in practice, but has the drawback that error in the "tail" regions of the distribution are sometimes comparable with the frequencies themselves. Notoriously, the Edgeworth approximation can assume negative values in such regions.

The characteristic function of the statistic may be known, and the difficulty is then the analytical one of inverting a Fourier transform explicitly. It is possible to show that for some statistics a satisfactory approximation to its probability density, when it exists, can be obtained nearly always by the method of steepest descents. This gives an asymptotic expansion in powers of n^{-1}, where n is the sample size, whose dominant term, called the saddlepoint approximation, has a number of desirable features. The error incurred by its use is $O(n^{-1})$ as against the more usual $O(n^{-1/2})$ associated with the normal approximation.

The Saddlepoint Approximation

Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be a vector of observations of n random variables with joint density $f(\mathbf{y})$. Suppose that the real random variable $S_n = S_n(\mathbf{y})$ has a density with respect to Lebesgue measure which depends on integer $n > N$ for some positive N. Let $\phi_n(z) = E(e^{izS_n})$ be the characteristic function of S_n where i is the imaginary unit. The cumulant generating function of S_n is $\psi_n(z) = \log \phi_n(z) = K_n(T)$ with $T = iz$. Whenever the appropriate derivatives exist, let $\partial^j \psi_n(\widetilde{z})/\partial z^j$ denote the jth order derivative evaluated at $z = \widetilde{z}$. The jth cumulant κ_{nj} of S_n, where it exists, satisfies the relation

$$i^j \kappa_{nj} = \frac{\partial^j \psi_n(0)}{\partial z^j}. \tag{1}$$

It is assumed that the derivatives $\partial^j \psi_n(z)/\partial z^j$ exist and are $O(n)$ for all z and $j = 1, 2, \ldots, r$ with $r \geq 4$. We use here partial derivatives because the functions involved may depend on something else, a parameter vector for example.

Let $h_n(x)$ be the density of the statistics $X_n = n^{-1/2}\{S_n - E(S_n)\}$. The characteristic function of X_n is

$$\begin{aligned}\phi_n^*(z) = E(e^{izX_n}) &= E\left(\exp\left\{i\frac{z}{\sqrt{n}}\{S_n - E(S_n)\}\right\}\right) \\ &= e^{-i\frac{z}{\sqrt{n}}E(S_n)} E\left\{e^{i\frac{z}{\sqrt{n}}S_n}\right\} \\ &= e^{-i\frac{z}{\sqrt{n}}E(S_n)} \phi_n\left(\frac{z}{\sqrt{n}}\right),\end{aligned} \tag{2}$$

where ϕ_n is the characteristic function of S_n.

Without loss of generality assume that $E(S_n) = 0$, therefore

$$\phi_n^*(z) = E(e^{izX_n}) = \phi_n\left(\frac{z}{\sqrt{n}}\right). \tag{3}$$

The cumulant generating function of X_n is

$$\psi_n^*(z) = \log \phi_n^*(z) = K_n^*(T), \tag{4}$$

with $T = iz$.

Let $\widehat{T} = i\widehat{z}$ be the root of the equation

$$\frac{\partial K_n^*(T)}{\partial T} = X_n. \tag{5}$$

The density function $h_n(x)$ of the statistics X_n is given by the usual Fourier inversion formula

$$\begin{aligned}h_n(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_n^*(z) e^{-izX_n} dz \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left\{\psi_n^*(z) - izX_n\right\} dz.\end{aligned} \tag{6}$$

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

where $\psi_n^*(z)$ was given in (4). It is convenient here to employ the equivalent inversion formula

$$h_n(x) = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} \exp\left\{K_n^*(T) - TX_n\right\}dT, \qquad (7)$$

where $-c_1 < a < c_2, 0 \le c_1 < \infty, 0 \le c_2 < \infty$, but $c_1 + c_2 > 0$, thus either c_1 or c_2 may be zero, though not both, and $K_n^*(T)$ was defined in (4).

Let us write $T = \widehat{T} + iw$, where $\widehat{T}$ is the root of the Eq. (5). The argument then proceeds formally as follows. On the contour near $\widehat{T}$, the exponent of (7) can be written as

$$\begin{aligned} K_n^*(T) - TX_n &= K_n^*(\widehat{T}) - \widehat{T}X_n + iw\frac{\partial}{\partial T}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad + \frac{1}{2}(iw)^2\frac{\partial^2}{\partial T^2}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad + \frac{1}{6}(iw)^3\frac{\partial^3}{\partial T^3}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad + \frac{1}{24}(iw)^4\frac{\partial^4}{\partial T^4}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} + \cdots \\ &= K_n^*(\widehat{T}) - \widehat{T}X_n - \frac{1}{2}w^2\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2} \\ &\quad - \frac{i}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3} \\ &\quad + \frac{1}{24}w^4\frac{\partial^4 K_n^*(\widehat{T})}{\partial T^4} + \cdots, \end{aligned} \qquad (8)$$

where $\frac{\partial}{\partial T}\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} = 0$ because $\widehat{T}$ is the root of (5).

Because of (8), the integrand of (7) can be written as

$$\begin{aligned} &\exp\left\{K_n^*(T) - TX_n\right\} \\ &= \exp\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\}\exp\left\{-\frac{1}{2}w^2\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\} \\ &\quad \times\left\{1 - \frac{i}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3} + \frac{1}{24}w^4\frac{\partial^4 K_n^*(\widehat{T})}{\partial T^4}\right. \\ &\quad \left. -\frac{1}{2}\left\{\frac{1}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3}\right\}^2 + \cdots\right\}. \end{aligned} \qquad (9)$$

Using $T = \widehat{T} + iw$, we can transform from T to w in (7) resulting that

$$\begin{aligned} h_n(x) &= \frac{1}{2\pi}\exp\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad \times\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}w^2\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}\left\{1 - \frac{i}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3}\right. \\ &\quad + \frac{1}{24}w^4\frac{\partial^4 K_n^*(\widehat{T})}{\partial T^4} \\ &\quad \left. -\frac{1}{2}\left\{\frac{1}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3}\right\}^2 + \cdots\right\}dw. \end{aligned} \qquad (10)$$

The odd powers of w vanish on integration. On the other hand, for $j = 2, 3, \ldots$ and since $\frac{\partial^j}{\partial T^j}K_n(T)$ is $O(n)$

$$\begin{aligned} \frac{\partial^j K_n^*(T)}{\partial T^j} &= \frac{\partial^j}{\partial T^j}K_n\left(\frac{T}{\sqrt{n}}\right) = \frac{\partial^j}{\partial T^{*j}}K_n(T^*)\left(\frac{1}{\sqrt{n}}\right)^j \\ &= O\left(n^{-\frac{j}{2}+1}\right), \end{aligned} \qquad (11)$$

where $\mathbf{T}^* = \frac{\mathbf{T}}{\sqrt{n}}$. Therefore

$$\begin{aligned} h_n(x) &= \frac{1}{\sqrt{2\pi}}\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-\frac{1}{2}}\exp\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\} \\ &\quad \times\left\{1 + \frac{1}{n}Q_4\left(\widehat{T}\right) + \cdots\right\}, \end{aligned} \qquad (12)$$

where

$$\begin{aligned} Q_4\left(\widehat{T}\right) &= \frac{n\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-\frac{1}{2}}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}w^2\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\} \\ &\quad \times\left\{\frac{1}{24}w^4\frac{\partial^4 K_n^*(\widehat{T})}{\partial T^4}\right. \\ &\quad \left. -\frac{1}{2}\left\{\frac{1}{6}w^3\frac{\partial^3 K_n^*(\widehat{T})}{\partial T^3}\right\}^2\right\}dw. \end{aligned} \qquad (13)$$

Clearly, $Q_4\left(\widehat{T}\right)$ defined in (13) is n times the sum of two terms. The first of these terms is, apart from a multiplicative constant, $\frac{\partial^4 K_n^*(T)}{\partial T^4}$ times fourth order moments of a normal random variable with zero mean and variance $\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-1}$; and the second term is also a constant times $\left(\frac{\partial^3 K_n^*(T)}{\partial T^3}\right)^2$ and sixth order moments of a normal random variable with zero mean and variance $\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-1}$. Thus, because of (11), $Q_4\left(\widehat{T}\right) = O(1)$. Consequently, we write (12) as

$$h_n(x) = \widehat{h}_n(x)\left\{1 + O(n^{-1})\right\}, \qquad (14)$$

where

$$\widehat{h}_n(x) = \frac{1}{\sqrt{2\pi}}\left\{\frac{\partial^2 K_n^*(\widehat{T})}{\partial T^2}\right\}^{-\frac{1}{2}}\exp\left\{K_n^*(\widehat{T}) - \widehat{T}X_n\right\}. \qquad (15)$$

The expression (15) receives the name of saddlepoint approximation to $h_n(x)$, been the error of approximation of order n^{-1}.

Daniels (1956) pointed out that when the constant term in the saddlepoint approximation is adjusted to make the integral over the whole sample space equal to unity, the order of magnitude of the error is reduced in a certain sense from n^{-1} to $n^{-3/2}$. He called this process renormalization.

About the Author

Professor Abril is co-editor of the *Revista de la Sociedad Argentina de Estadística* (Journal of the Argentinean Statistical Society).

Cross References

►Approximations to Distributions
►Dispersion Models
►Edgeworth Expansion
►Exponential Family Models
►Inverse Sampling

References and Further Reading

Abril JC (1985) Asymptotic expansions for time series problems with applications to moving average models. PhD Thesis, The London School of Economics and Political Science, University of London, England

Barndorff-Nielsen O, Cox DR (1979) Edgeworth and saddle-point approximations with statistical applications. J R Stat Soc B 41:279–312

Daniels HE (1954) Saddlepoint approximations in statistics. Ann Math Stat 25:631–650

Daniels HE (1956) The approximate distribution of serial correlation coefficients. Biometrika 43:169–185

Durbin J (1980) Approximations for the densities of sufficient estimates. Biometrika 67:311–333

Feller W (1971) An introduction to probability theory and its applications, vol 2, 2nd edn. Wiley, New York

Phillips PCB (1978) Edgeworth and saddlepoint approximations in a first order autoregression. Biometrika 65:91–98

Wallace DL (1958) Asymptotic approximations to distributions. Ann Math Stat 29:635–654

Sample Size Determination

Michael P. Cohen
Adjunct Professor
NORC at the University of Chicago, Chicago, IL, USA
Adjunct Professor
George Mason University, Fairfax, VA, USA

A common problem arising in statistics is to determine the smallest sample size needed to achieve a specified inference goal. Examples of inference goals include finding a 95% confidence interval for a given statistic of width no larger than a specified amount, or performing a hypothesis test at the 5% significance level with power no smaller than a specified amount. These examples and others are discussed more fully below.

Sample Size to Achieve a Given Variance or Relative Variance

One may want to estimate a parameter θ by an estimator $\hat{\theta}$ based on a sample of size n. Often the variance of $\hat{\theta}$, $\mathrm{var}(\hat{\theta})$, will have the form $\mathrm{var}(\hat{\theta}) = b/n$ for some known constant b. To achieve a variance of $\hat{\theta}$ no larger than a specified amount A, one simply sets $A = b/n$ and solves for n: $n = b/A$. The value of n must be an integer, so one takes n to be the smallest integer no smaller than b/A. Note that n is inversely related to the desired precision A.

It is more typically the case that b will depend on unknown parameters, usually including θ. Because the sample has not been selected yet, one must estimate the parameters from a previous sample or from other outside information. Precise values are not needed as one is usually satisfied with a conservative (that is, high) estimate for the required sample size n.

It is common to be interested in the *relative variance* $\frac{\mathrm{var}(\hat{\theta})}{\theta^2}$, also known as the square of the coefficient of variation or CV^2. In this case, one has

$$\frac{\mathrm{var}(\hat{\theta})}{\theta^2} = \frac{b}{\theta^2 n}$$

so to keep CV^2 less than a desired amount A, one sets $n = \frac{b}{\theta^2 A}$. Again, b and θ may need to be estimated from a previous sample or some outside source.

The variance of an estimated proportion $\hat{p}$ from a ►simple random sample of size n (from an infinite population) is

$$\mathrm{var}(\hat{p}) = \frac{p(1-p)}{n} = \frac{1}{4n} - \frac{(1/2-p)^2}{n} \le \frac{1}{4n}.$$

Therefore, to achieve a variance of $\hat{p}$ of at most A, it suffices that n be at least $\frac{1}{4A}$. For this conservation determination of the sample size, no estimation of unknown parameters is needed.

One can also consider the estimation of an estimated proportion $\hat{p}$ from a simple random sample of size n from a finite population of size N. In this case,

$$\mathrm{var}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n} \le \left(1 - \frac{n}{N}\right)\frac{1}{4n}.$$

To achieve a variance of $\hat{p}$ of at most A as a conservative estimate, n must be at least

$$\frac{1}{4A + 1/N}.$$

Sample Size to Achieve a Given Power in a Hypothesis Test

In hypothesis testing, the probability of *type I error* (the probability of rejecting a null hypothesis when it, in fact, holds) is typically fixed at a predetermined level, called alpha (α). The value $\alpha = 5\%$ is very common. A sample size n is sought so that the test achieves a certain *type II error* rate (the probability of not rejecting the null hypothesis when a specific alternative actually holds), called beta (β). The *power* of a test is $1 - \beta$, the probability of rejecting the null hypothesis when a specific alternative holds. So sample size determination can be described as finding the smallest value of n so that for the predetermined α the power achieves some desired level for a fixed alternative. The term *statistical* ▶*power analysis* is frequently used as a synonym for sample size determination.

To be specific, suppose one wants to test that the mean μ of independent, identically normally distributed data is equal to μ_0 versus the alternative that the mean is greater than μ_0. One can write this as $H_0 : \mu = \mu_0$ versus $H_\alpha : \mu > \mu_0$. Suppose also that $\mu' > \mu_0$ is sufficiently far from μ_0 that the difference is deemed to be of practical significance in the subject-matter area of the test. Let Z be a standard normal random variable, Φ be its cumulative distribution function, and z_α be defined by $P(Z \geq z_\alpha) = \alpha$. Then it can be calculated that the type II error at $\mu', \beta(\mu')$, is

$$\begin{aligned}\beta(\mu') &= P(H_0 \text{ is not rejected when } \mu = \mu') \\ &= \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)\end{aligned}$$

where σ^2 is the known variance of the data and n is the sample size. It follows from this that

$$-z_\beta = z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}.$$

Solving for n, one gets

$$n = \left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2.$$

This sample size (adjusted upward to an integer value, if necessary) is needed to achieve a significance level of α and power of $1 - \beta(\mu')$ at μ'. The same sample size n applies when the alternative hypothesis is $H_\alpha : \mu < \mu_0$. For the two-sided alternative hypothesis $H_\alpha : \mu \neq \mu_0$, one has by a similar argument (involving an approximation) that

$$n = \left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'}\right]^2.$$

For this testing problem, one is able to get explicit solutions. It is typical, however, to have to resort to complicated tables or, more recently, software, to get a solution.

Sample Size to Achieve a Given Width for a Confidence Interval

A $100(1-\alpha)\%$ ▶confidence interval for the mean μ of a normal population with known variance σ^2 is

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean. When n is reasonably large, say 30 or greater, this interval with σ replaced by $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ holds approximately when σ^2 is unknown.

The width of the interval is $w = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. So, solving for n, the sample size needed to achieve an interval of width w and confidence level $100(1-\alpha)\%$ is $n = 4\sigma^2\left(\frac{z_{\alpha/2}}{w}\right)^2$ (or $n = 4S^2\left(\frac{z_{\alpha/2}}{w}\right)^2$ when σ^2 unknown and $n \geq 30$).

As with hypothesis testing, the sample size problem for confidence intervals more typically requires tables or software to solve.

The Scope of Statistical Procedures for Sample Size Determination

Sample size determination arises in one sample problems, two sample problems, ▶analysis of variance, regression analysis, ▶analysis of covariance, multilevel models, survey sampling, nonparametric testing, ▶logistic regression, survival analysis, and just about every area of modern statistics. In the case of multilevel models (e.g., hierarchical linear models), one must determine the sample size at each level in addition to the overall sample size (Cohen 2005). A similar situation arises in sample size determination for complex sample surveys.

Software for Sample Size Determination

The use of software for sample size determination is highly recommended. Direct calculation is difficult (or impossible) in all but the simplest cases. Tables are cumbersome and often incomplete. Specific software products will not be recommended here, but we mention some to indicate the wide range of products available.

Statisticians who use SAS® should be aware that versions 9.1 and later include releases of PROC POWER and PROC GLMPOWER (PROC means "procedure" in SAS® and GLM stands for "general linear model") that are full featured.

SPSS has a product called SamplePower® that also has many features. Other commercial products include nQuery Advisor and PASS. G*Power is a free product. Sampsize is also free with an emphasis on survey sampling sample size calculations. A Web search will reveal many other products that should suit particular needs.

About the Author

Biography of Cohen is in ►Stratified Sampling.

Cross References

►Confidence Interval
►Power Analysis
►Significance Testing: An Overview
►Statistical Evidence

References and Further Reading

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Erlbaum, Hillsdale
Cohen MP (2005) Sample size considerations for multilevel surveys. Int Stat Rev 73:279–287
Dattalo P (2008) Determining sample size. Oxford University Press, New York

Sample Survey Methods

Peter Lynn
Professor of Survey Methodology
University of Essex, Colchester, UK

A sample survey can be broadly defined as an exercise that involves collecting standardised data from a sample of *study units* (e.g., persons, households, businesses) designed to represent a larger population of units, in order to make quantitative inferences about the population. Within this broad definition there is a large variety of different types of survey. Surveys can differ in terms of the type of data collected, the methods used to collect the data, the design of the sample, and whether data is collected repeatedly, either on the same sample or on different samples. Key features of a sample survey are:

Survey objectives must be clear and agreed at the outset, so that all features of the survey can be designed with these objectives in mind;
The target population – about which knowledge is required – must be defined. For example, it might be all persons usually resident in a particular town, or all farms within a national boundary;
The survey sample must be designed to represent the target population;
Relevant concepts must be addressed by the survey measures, so that the survey data can be used to answer important research questions;
The survey measures – which typically include questions, but could also include anthropometric measures, soil samples, etc – must be designed to provide accurate indicators of the concepts of interest;
Survey implementation should achieve the co-operation of a high proportion of sample members in a cost-efficient and timely manner.

The aim is to obtain relevant data that are representative, reliable and valid.

Representation concerns the extent to which the units in the data set represent the units in the target population and therefore share the pertinent characteristics of the population as a whole. This will depend on the identification of a sampling frame, the selection of a sample from that frame, and the attempts made to obtain data for the units in the sample.

Sampling frame. Ideally, this is a list of all units in the population, from which a sample can be selected. Sometimes the list pre-exists, sometimes it must be constructed especially for the survey, and sometimes a sampling method can be devised that does not involve the creation of an explicit list but is equivalent (Lynn 2002).

Sample design. In 1895 Anders Kiaer, founding Director of Statistics Norway, proposed sampling as a way of learning about a population without having to study every unit in the population. The basic statistical theory of probability sampling developed rapidly in the first half of the twentieth century and underpinned the growth of surveys. The essence is that units must be selected at random with known and non-zero selection probabilities. This enables unbiased estimation of population parameters and estimation of the precision (standard errors) of estimates (Groves et al. 2004, Chap. 4). Design features such as stratified sampling and multi-stage (clustered) sampling are commonly used within a probability sampling framework. Some surveys, particularly in the commercial sector, use non-probability methods such as quota sampling.

Non-response. Once a representative sample has been selected, considerable efforts are usually made to achieve the highest possible *response rate* (Lynn 2008). In many countries, high quality surveys of the general population typically achieve response rates in the range 60–80%, with rates above 80% being considered outstanding. The main reasons for non-response are usually *refusal* (unwillingness of sample member to take part) and *non-contact* (inability of the survey organisation to reach the sample member). Other reasons include an *inability* to take part, for example

due to language or ill health. Different strategies are used by survey organizations to minimize each of these types of non-response. Ultimately, non-response can introduce *bias* to survey estimates if the non-respondents differ from respondents in terms of the survey measures. Adjustment techniques such as *weighting* (Lynn 2004) can reduce the bias caused by non-response.

Obtaining reliable and valid data from respondents depends upon the measurement process. This includes development of *concepts* to be measured, development of *measures* of those concepts (e.g., survey questions), obtaining *responses* to the measures, and post-fieldwork *processing* (such as editing, coding, and combining the answers to a number of questions to produce derived variables). Failure of the obtained responses to correctly reflect the concept of interest is referred to as *measurement error* (Biemer et al. 1991). To minimise measurement error, survey researchers pay attention to cognitive response theory (Tourangeau et al. 2000), which describes four steps in the process of answering a survey question:

Understanding. The sample member must understand the question as intended by the researcher. This requires the question and the required response to be clear, simple, unambiguous and clearly communicated.

Recall. The sample member must be able to recall all the information that is required in order to answer the question. Question designers must be realistic regarding what respondents can remember and should provide tools to aid memory, if appropriate.

Evaluation. The sample member must process the recalled information in order to form an answer to the question.

Reporting. The sample member must be willing and able to communicate the answer. Various special techniques are used by survey researchers to elicit responses to questions on sensitive or embarrassing issues.

Two fundamental survey design issues with considerable implications are the following:

Data collection modes. There are several available methods to collect survey data (Groves et al. 2004, Chap. 5). An important distinction is between interviewer-administered methods (face-to-face personal interviewing, telephone interviewing) and self-completion methods (paper self-completion ►questionnaires, web surveys). With self-completion methods, the researcher usually has less control over factors such as who is providing the data and the order in which questions are answered, as well as having a limited ability to address respondent concerns and to provide help. Self-completion methods also require a higher degree of literacy and cognitive ability than interviews and so may be inappropriate for certain study populations. On the other hand, respondents may be more willing to reveal sensitive or embarrassing answers if there is no interviewer involved. There are often large differences in survey costs between the possible modes. This consideration often leads to surveys being carried out in a mode which might otherwise be thought sub-optimal.

Longitudinal designs. It is often beneficial to collect repeated measures from the same sample over time. This allows the measurement of change and identification of the ordering of events, which can shed light on causality. Surveys which collect data from the same units on multiple occasions are known as longitudinal surveys (Lynn 2009) and involve additional organisation and complexity. Some longitudinal social surveys have been running for several decades and are highly valued data sources.

About the Author

Dr. Peter Lynn is Professor of Survey Methodology, Institute for Social and Economic Research, University of Essex. He is Vice-President of the International Association of Survey Statisticians (2009–2011). He is Editor-in-Chief, *Survey Research Methods* (since 2005), and Director, UK Survey Resources Network (since 2008). He was Joint Editor, *Journal of the Royal Statistical Society Series A* (Statistics in Society) (2002–2005), and Editor of *Survey Methods Newsletter* (1996–2001). Dr Lynn is founding board member (since 2005) of the European Survey Research Association, Elected full member of the International Statistical Institute (since 2002) and Fellow of the Royal Statistical Society (since 1986). He has published widely on topics including survey non-response, weighting, data collection mode effects, respondent incentives, dependent interviewing, sample design, and survey quality. His recent publications include the book *Methodology of Longitudinal Surveys* (Editor, Wiley, 2009) and a chapter, *The Problem of Nonresponse*, in the *International Handbook of Survey Methodology* (Erlbaum, 2008). He was awarded the 2004 Royal Statistical Society Guy Medal in Bronze.

Cross References

►Balanced Sampling
►Business Surveys
►Census
►Cluster Sampling
►Empirical Likelihood Approach to Inference from Sample Survey Data
►Inference Under Informative Probability Sampling
►Internet Survey Methodology: Recent Trends and Developments

►Multistage Sampling
►Non-probability Sampling Survey Methods
►Panel Data
►Questionnaire
►Repeated Measures
►Representative Samples
►Sampling From Finite Populations
►Superpopulation Models in Survey Sampling
►Telephone Sampling: Frames and Selection Techniques
►Total Survey Error

References and Further Reading

Biemer P, Groves RM et al (1991) Measurement errors in surveys. Wiley, New York
Groves RM, Fowler FJ et al (2004) Survey methodology. Wiley, New York
Lynn P (2002) Sampling in human studies. In: Greenfield T (ed) Research methods for postgraduates, 2nd edn. Arnold, London, pp 195–202
Lynn P (2004) Weighting. In: Kempf-Leonard K (ed) Encyclopedia of social measurement. Academic, New York, NY, pp 967–974
Lynn P (2008) The problem of nonresponse. In: deLeeuw E, Hox J, Dillman D (eds) The international handbook of survey methodology. Lawrence Erlbaum Associates, Mahwah, NJ, pp 35–55
Lynn P (ed) (2009) Methodology of longitudinal surveys. Wiley, New York
Tourangeau R, Rips LJ, Rasinski K (2000) The psychology of survey response. Cambridge University Press, Cambridge

Sampling Algorithms

Yves Tillé
Professor
University of Neuchâtel, Neuchâtel, Switzerland

A sampling algorithm is a procedure that allows us to select randomly a subset of units (a sample) from a population without enumerating all the possible samples of the population.

More precisely, let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population and $s \subset U$ a sample or a subset of U. A sampling design $p(s)$ is a probability distribution on the set of all the subsets $s \subset U$, i.e., $p(s) \geq 0$ and

$$\sum_{s \subset U} p(s) = 1.$$

The inclusion probability $\pi_k = pr(k \in s)$ of a unit k is its probability of being selected in the sample s. The sum of the inclusion probabilities is equal to the expectation of the sample size n.

In many sampling problem, the number of possible samples is generally very large. For instance, if $N = 50$ and $n = 10$, the number of possible samples already equals 10,272,278,170. The selection of a sample by enumerating all the possible samples is generally impossible. A sampling algorithm is a method that allows bypassing this enumeration. There exists several class of methods:

- *Sequential algorithms*. In this case, there is only one reading of the population file. Each unit is successively examined and the decision of selection is irremediably taken.
- *One by one algorithms*. At each step, a unit is selected from the population until obtaining the fixed sample size.
- *Eliminatory algorithms*. At each step, a unit is removed from the population until obtaining the fixed sample size.
- *Rejective methods*. For instance, sample with replacement are generated until obtaining a sample without replacement. Rejective methods can be interesting if there exists a more general sampling design that is simpler than the design we want to implement.
- *Splitting methods*. This method described in Deville and Tillé (1998) starts with a vector of inclusion probability. At each step, this vector is randomly replaced by another vector until obtaining a vector containing only zeros and ones i.e., a sample.

The same sampling design can generally be implemented by using different methods. For instance, Tillé (2006) gives sequential, one by one, eliminatory algorithms for several sampling designs like simple random sampling with and without replacement and multinomial sampling.

The main difficulties however appears when the sample is selected with unequal inclusion probabilities without replacement and fixed sample size. The first proposed method was systematic sampling with unequal inclusion probabilities (Madow 1949). For this sequential algorithm, first compute the cumulated inclusion probabilities V_k. Next units such that

$$V_{k-1} \leq u + i - 1 < V_k, \qquad i = 1, 2, \ldots, n,$$

are selected, where u is a uniform continuous random variable in [0,1) and n is the sample size.

An interesting rejective procedure was proposed by Sampford (1967). Samples are selected with replacement. The first unit is selected with probability π_k/n, the $n-1$ other units are selected with probability

$$\frac{\pi_k}{n(1-\pi_k)} \left\{ \sum_{\ell=1}^{N} \frac{\pi_\ell}{n(1-\pi_\ell)} \right\}^{-1}.$$

The sample is accepted if n distinct units are selected, otherwise another sample is selected.

Chen et al. (1994) discussed the sampling design without replacement and fixed sample size that maximizes the ►entropy given by

$$I(p) = -\sum_{s \in U} p(s) \log p(s).$$

They gave a procedure for selecting a sample according this sampling design. Several other efficient algorithms that implement this sampling design are given in Tillé (2006).

Other methods have been proposed by Brewer (1975), Deville and Tillé (1998). A review is given in Brewer and Hanif (1983) and Tillé (2006). Other sampling algorithms allows us to solve more complex problems. For instance, the cube method (Deville and Tillé 2004) allows selecting balanced samples (see ►Balanced Sampling) in the sense that the ►Horvitz-Thompson estimator are equal or approximately equal to the population totals for a set of control variables.

About the Author

Yves Tillé is a professor and Director of the Institute of Statistics of the University of Neuchâtel. He was Associate editor of the *Journal of the Royal Statistical Society B* (2008–2009), *Survey Methodology* (2002–2009) and of *Metron* (2008–). He is author of several books in French and English, including *Sampling Algorithms*, Springer, 2006.

Cross References

►Balanced Sampling
►Entropy
►Horvitz–Thompson Estimator
►Randomization
►Sample Survey Methods
►Sequential Sampling

References and Further Reading

Brewer K (1975) A simple procedure for πpswor. Aust J Stat 17: 166–172
Brewer K, Hanif M (1983) Sampling with unequal probabilities. Springer-Verlag, New York
Chen S, Dempster A, Liu J (1994) Weighted finite population sampling to maximize entropy. Biometrika 81:457–469
Deville J-C, Tillé Y (1998) Unequal probability sampling without replacement through a splitting method. Biometrika 85:89–101
Deville J-C, Tillé Y (2004) Efficient balanced sampling: the cube method. Biometrika 91:893–912
Madow W (1949) On the theory of systematic sampling, II. Ann Math Stats 20:333–354
Sampford M (1967) On sampling without replacement with unequal probabilities of selection. Biometrika 54:499–513
Tillé Y (2006) Sampling algorithms. Springer-Verlag, New York

Sampling Distribution

David W. Stockburger
Deputy Director of Academic Assessment
US Air Force Academy
Emeritus Professor of Psychology
Missouri State University, Springfield, MO, USA

What is it?

The sampling distribution is a distribution of a sample statistic. When using a procedure that repeatedly samples from a population and each time computes the same sample statistic, the resulting distribution of sample statistics is a sampling distribution of that statistic. To more clearly define the distribution, the name of the computed statistic is added as part of the title. For example, if the computed statistic was the sample mean, the sampling distribution would be titled "the sampling distribution of the sample mean."

For the sake of simplicity let us consider a simple example when we are dealing with a small *discrete* population consisting of the first ten integers $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Let us now repeatedly take random samples without replacement of size $n = 3$ from this population. The random sampling might generate sets that look like $\{8, 3, 7\}, \{2, 1, 5\}, \{6, 3, 5\}, \{10, 7, 5\} \ldots$ If the mean $(\overline{X})$ of each sample is found, the means of the above samples would appear as follows: 6, 2.67, 4.67, 7.33 ... How many different samples can we take, or put it differently, how many different sample means can we obtain? In our artificial example only 720, but in reality when we analyze very large populations, the number of possible different samples (of the same size) can be for all practical purposes treated as countless.

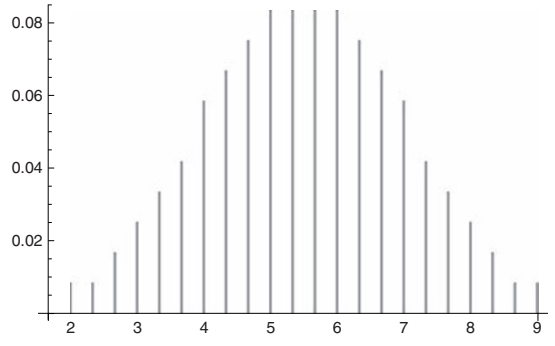
Once we have obtained sample means for all samples, we have to list all their different values and number of their occurrences (frequencies). Finally, we will divide each frequency with the total number of samples to obtain *relative frequencies* (empirical probabilities). In this way we will come up to a list of all possible sample means and their relative frequencies. When the population is discrete, that list is called the *sampling distribution* of that statistic. Generally, the sampling distribution of a statistic is a probability distribution of that statistic derived from all possible samples having the same size from the population.

When we are dealing with a *continuous* population it is impossible to enumerate all possible outcomes, so we have to rely on the results obtained in mathematical statistics (see section "►How Can Sampling Distributions be

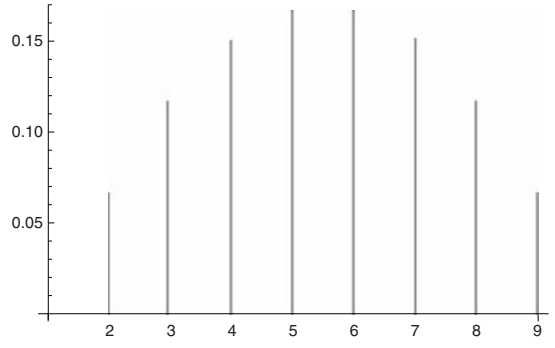
Constructed Mathematically?" of this paper for an example). Still, we can imagine a process that is similar to the one in the case of a discrete population. In that process we will take repeatedly thousands of different samples (of the same size) and calculate their statistic. In that way we will come to the relative frequency distribution of that statistic. The more samples we take, the closer this relative frequency distribution will come to the sampling distribution. Theoretically, as the number of samples approaches infinity our frequency distribution will approach the sampling distribution.

Sampling distribution should not be confused with a *sample* distribution: the latter describes the distribution of values (elements) in a *single* sample.

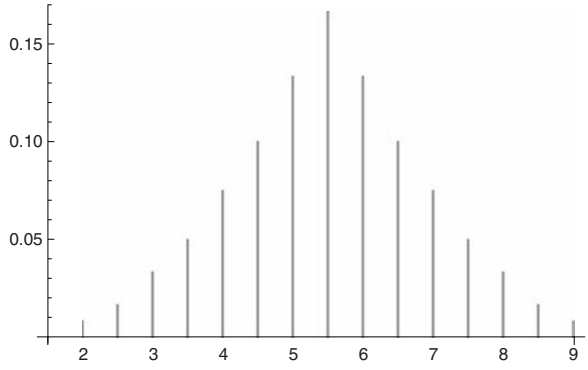
Referring back to our example, we can graphically display the sampling distribution of the mean as follows:



Every statistic has a sampling distribution. For example, suppose that instead of the mean, *medians* (M_d) were computed for each sample. That is, within each sample the scores would be rank ordered and the middle score would be selected as the median. Using the samples above, the medians would be: $7, 2, 5, 7 \ldots$ The distribution of the medians calculated from all possible different samples of the same size is called the sampling distribution of the median and could be graphically shown as follows:



It is possible to make up a new statistic and construct a sampling distribution for that new statistic. For example, by rank ordering the three scores within each sample and finding the mean of the highest and lowest scores a new statistic could be created. Let this statistic be called the mid-mean and be symbolized by $\overline{M}$. For the above samples the values for this statistic would be: $5.5, 3, 4.5, 7.5 \ldots$ and the sampling distribution of the mid-mean could be graphically displayed as follows:



Just as the population distributions can be described with parameters, so can the sampling distribution. The expected value and variance of any distribution can be represented by the symbols μ (mu) and σ^2 (Sigma squared), respectively. In the case of the sampling distribution, the μ symbol is often written with a subscript to indicate which sampling distribution is being described. For example, the expected value of the sampling distribution of the mean is represented by the symbol $\mu_{\overline{X}}$, that of the median by μ_{M_d}, and so on. The value of $\mu_{\overline{X}}$ can be thought of as the theoretical mean of the distribution of means. In a similar manner the value of μ_{M_d}, is the theoretical mean of a distribution of medians.

The square root of the variance of a sampling distribution is given a special name, the *standard error*. In order to distinguish different sampling distributions, each has a name tagged on the end of "standard error" and a subscript on the σ symbol. The theoretical *standard deviation* of the sampling distribution of the mean is called the standard error of the mean and is symbolized by $\sigma_{\overline{X}}$. Similarly, the theoretical standard deviation of the sampling distribution of the median is called the standard error of the median and is symbolized by σ_{M_d}.

In each case the standard error of the sampling distribution of a statistic describes the degree to which the computed statistics may be expected to differ from one another when calculated from a sample of similar size and selected from similar population models. The larger

the standard error of a given statistic, the greater the differences between the computed statistics for the different samples. From the example population, sampling method, and statistics described earlier, we would find $\mu_{\overline{X}} = \mu_{M_d} = \mu_{\overline{M}} = 5.5$ and $\sigma_{\overline{X}} = 1.46$, $\sigma_{M_d} = 1.96$, and $\sigma_{\overline{M}} = 1.39$.

Why is the Sampling Distribution Important – Properties of Statistics

Statistics have different properties as estimators of a population parameters. The sampling distribution of a statistic provides a window into some of the important properties. For example if the expected value of a statistic is equal to the expected value of the corresponding population parameter, the statistic is said to be unbiased. In the example above, all three statistics would be unbiased estimators of the population parameter μ_X.

Consistency is another valuable property to have in the estimation of a population parameter, as the *statistic* with the smallest standard error is preferred as an *estimator* of the corresponding population parameter, everything else being equal. Statisticians have proven that the standard error of the mean is smaller than the standard error of the median. Because of this property, the mean is generally preferred over the median as an estimator of μ_X.

Hypothesis Testing

The sampling distribution is integral to the hypothesis testing procedure. The sampling distribution is used in hypothesis testing to create a model of what the world would look like given the null hypothesis was true and a statistic was collected an infinite number of times. A single sample is taken, the sample statistic is calculated, and then it is compared to the model created by the sampling distribution of that statistic when the null hypothesis is true. If the sample statistic is unlikely given the model, then the model is rejected and a model with real effects is more likely. In the example process described earlier, if the sample $\{3, 1, 4\}$ was taken from the population described above, the sample mean (2.67), median (3), or mid-mean (2.5) can be found and compared to the corresponding sampling distribution of that statistic. The probability of finding a sample statistic of that size or smaller could be found for each e.g. mean ($p < .033$), median ($p < .18$), and mid-mean ($p < .025$) and compared to the selected value of alpha (α). If alpha was set to .05, then the selected sample would be unlikely given the mean and mid-mean, but not the median.

How Can Sampling Distributions be Constructed Mathematically?

Using advanced mathematics statisticians can prove that under given conditions a sampling distribution of some statistic must be a specific distribution. Let us illustrate this with the following theorem (for the proof see for example Hogg and Tanis (1997, p. 256)):

If $X_1, X_2, \ldots, X_n$ are observations of a random sample of size n from the normal distribution $N(\mu, \sigma^2)$,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

then

$$\frac{(n-1)S^2}{\sigma^2} \text{ is } \chi^2(n-1).$$

The given conditions describe the assumptions that must be made in order for the distribution of the given sampling distribution to be true. For example, in the above theorem, assumptions about the sampling process (random sampling) and distribution of X (a normal distribution) are necessary for the proof.

Of considerable importance to statistical thinking is the sampling distribution of the mean, a theoretical distribution of sample means. A mathematical theorem, called the Central Limit Theorem, describes the relationship of the parameters of the sampling distribution of the mean to the parameters of the probability model and sample size. The Central Limit Theorem also specifies the form of the sampling distribution (Gaussian) in the limiting case.

Selection of Distribution Type to Model Scores

The sampling distribution provides the theoretical foundation to select a distribution for many useful measures. For example, the central limit theorem describes why a measure, such as intelligence, that may be considered a summation of a number of independent quantities would necessarily be (approximately) distributed as a normal (Gaussian) curve.

Monte Carlo Simulations

It is not always easy or even possible to derive the exact nature of a given sampling distribution using mathematical derivations. In such cases it is often possible to use Monte Carlo simulations to generate a close approximation to the true sampling distribution of the statistic. For example, a non-random sampling method, a non-standard

distribution, or may be used with the resulting distribution not converging to a known type of probability distribution. When much of the current formulation of statistics was developed, Monte Carlo techniques, while available, were very inconvenient to apply. With current computers and programming languages such as Wolfram Mathematica (Kinney 2009), Monte Carlo simulations are likely to become much more popular in creating sampling distributions.

Summary

The sampling distribution, a theoretical distribution of a sample statistic, is a critical concept in statistical thinking. The sampling distribution allows the statistician to hypothesize about what the world would look like if a statistic was calculated an infinite number of times.

About the Author

Dr. David W. Stockburger is currently the Deputy Director of Academic Assessment at the US Air Force Academy. He is an emeritus professor of psychology at Missouri State University where he taught from 1973 to 2002. His online introductory statistics text http://www.psychstat.missouristate.edu/IntroBook2/sbk.htm has been continuously available since 1996 and an intermediate text http://www.psychstat.missouristate.edu/multibook2/mlt.htm appeared in 1997. His online probability calculator (2001) replaced statistical tables and provided a visual representation of probability distributions, saving students countless hours of learning how to use statistical tables and providing an exact significance level. He has entries in "Encyclopedia of Measurement and Statistics" and "Encyclopedia of Research Design."

Cross References

►Bootstrap Methods
►Central Limit Theorems
►Cornish-Fisher Expansions
►Mean Median and Mode
►Monte Carlo Methods in Statistics
►Nonparametric Statistical Inference
►Significance Testing: An Overview
►Statistical Inference: An Overview

References and Further Reading

Hogg RV, Tanis EA (1997) Probability and statistical inference. 5th edn. Prentice Hall, Upper Saddle River, NJ

Kinney JJ (2009) A probability and statistics companion. Wiley, Hoboken, NJ

Sampling From Finite Populations

Jill M. Montaquila, Graham Kalton
Westat, Rockville, MD, USA

Introduction

The statistical objective in survey research and in a number of other applications is generally to estimate the parameters of a finite population rather than to estimate the parameters of a statistical model. As an example, the finite population for a survey conducted to estimate the unemployment rate might be all adults aged 18 or older living in a country at a given date. If valid estimates of the parameters of a finite population are to be produced, the finite population needs to be defined very precisely and the sampling method needs to be carefully designed and implemented. This entry focuses on the estimation of such finite population parameters using what is known as the *randomization* or *design-based approach*. Another approach that is particularly relevant when survey data are used for analytical purposes, such as for regression analysis, is known as the *superpopulation approach* (*see* ►Superpopulation Models in Survey Sampling).

This entry considers only methods for drawing probability samples from a finite population; *Nonprobability Sampling Methods* are reviewed in another entry. The basic theory and methods of probability sampling from finite populations were largely developed during the first half of the twentieth century, motivated by the desire to use samples rather than censuses (see ►Census) to characterize human, business, and agricultural populations. The paper by Neyman (1934) is widely recognized as a seminal contribution because it spells out the merits of *probability sampling* relative to purposive selection. A number of full-length texts on survey sampling theory and methods were published in the 1950's and 1960's including the first editions of Cochran (1977), Deming (1960), Hansen et al. (1953), Kish (1965), Murthy (1967), Raj (1968), Sukhatme et al. (1984), and Yates (1981). Several of these are still widely used as textbooks and references. Recent texts on survey sampling theory and methods include Fuller (2009), Lohr (2010), Pfeffermann and Rao (2009), Särndal et al. (1992), Thompson (1997), and Valliant et al. (2000).

Let the size of a finite population be denoted by N and let Y_i $(i = 1, 2, \ldots, N)$ denote the individual values of a variable of interest for the study. To carry forward the example given above, in a survey to estimate the unemployment rate, Y_i might be the labor force status of person (element) i. Consider the estimation of the population total

$Y = \sum_i^N Y_i$ based on a probability sample of n elements drawn from the population by sampling without replacement so that elements cannot be selected more than once. Let π_i denote the probability that element i is selected for the sample, with $\pi_i > 0$ for all i, and let π_{ij} denote the probability that elements i and j are jointly included in the sample. The sample estimator of Y can be represented as $\hat{Y} = \sum_i^N w_i Y_i$ where w_i is a random variable reflecting the sample selection, with $w_i = 0$ for elements that were not selected. The condition for $\hat{Y}$ to be an unbiased estimator of Y is that $E(w_i) = 1$. Now $E(w_i) = \pi_i w_i + (1-\pi_i)0$ so that for $\hat{Y}$ to be unbiased $w_i = \pi_i^{-1}$. The reciprocal of the selection probability, $w_i = \pi_i^{-1}$, is referred to as the *base weight*. The unbiased estimator for Y, $\hat{Y} = \sum_i^n w_i Y_i$, is widely known as the ▶Horvitz-Thompson estimator. The variance of $\hat{Y}$ is given by

$$
\begin{aligned}
V(\hat{Y}) &= \sum_i^N V(w_i) Y_i^2 + 2 \sum_i^N \sum_{j>i}^N Cov(w_i, w_j) Y_i Y_j \\
&= \sum_i^N \pi_i^{-1} (1 - \pi_i) Y_i^2 \\
&\quad + 2 \sum_i^N \sum_{j>i}^N \pi_i^{-1} \pi_j^{-1} (\pi_{ij} - \pi_i \pi_j) Y_i Y_j
\end{aligned}
$$

These general results cover a range of the different sample designs described below depending on the values of π_j and π_{ij}. The selection probabilities π_i appear in the estimator and, in addition, the joint selection probabilities π_{ij} appear in the variance. Note that when estimating the parameters of a finite population using the design-based approach for inference, the Y_i values are considered fixed; it is the w_i's that are the random variables.

The selection of a probability sample from a finite population requires the existence of a *sampling frame* for that population. The simplest form of sampling frame is a list of the individual population elements, such as a list of business establishments (when they are the units of analysis). The frame may alternatively be a list of clusters of elements, such as a list of households when the elements are persons. The initial frame may be a list of geographical areas that are sampled at the first stage of selection. These areas are termed *primary sampling units* (PSUs). At the second stage, subareas, or *second stage units*, may be selected within the sampled PSUs, etc. This design, which is known as an *area sample*, is a form of multistage sampling (see below).

The quality of the sampling frame has an important bearing on the quality of the final sample. An ideal sampling frame would contain exactly one listing for each element of the target population and nothing else. Sampling frames used in practice often contain departures from this ideal, in the form of noncoverage, duplicates, clusters, and ineligible units (see Kish 1965, Section 2.7, for a discussion of each of these frame problems). Issues with the sampling frames used in telephone surveys are discussed in the entry ▶Telephone Sampling: Frames and Selection Techniques. Sometimes, two or more sampling frames are used, leading to dual- or multiple-frame designs.

Sampling frames often contain auxiliary information that can be used to improve the efficiency of the survey estimators at the sample design stage, at the estimation stage, or at both stages. Examples are provided below.

Simple Random Sampling

A *simple random sample* is a sample design in which every possible sample of size n from the population of N elements has an equal probability of selection (*see* ▶Simple Random Sample). It may be selected by taking random draws from the set of numbers $\{1, 2, \ldots, N\}$. With simple random sampling, elements have equal probabilities of selection and simple random sampling is therefore *an equal probability selection method* (*epsem*).

Simple random sampling with replacement (SRSWR), also known as *unrestricted sampling*, allows population elements to be selected at any draw regardless of their selection on previous draws. Since elements are selected independently with this design, $\pi_{ij} = \pi_i \pi_j$ for all i, j. Standard statistical theory and analysis generally assumes SRSWR; this is discussed further in the entry ▶Superpopulation Models in Survey Sampling.

In *simple random sampling without replacement* (SRSWOR), also simply known as simple random sampling, once an element has been drawn, it is removed from the set of elements eligible for selection on subsequent draws. Since SRSWOR cannot select any element more than once (so that there are n distinct sampled elements), it is more efficient than SRSWR (i.e., the variances of the estimators are lower under SRSWOR than under SRSWR).

Systematic Sampling

In the simple case where the *sampling interval* $k = N/n$ is an integer, a *systematic sample* starts with a random selection of one of the first k elements on a list frame, and then selects every kth element thereafter. By randomly sorting the sampling frame, systematic sampling provides a convenient way to select a SRSWOR. Kish (1965, Section 4.1B)

describes various techniques for selecting a systematic sample when the sampling interval is not an integer.

If the sampling frame is sorted to place elements that are similar in terms of the survey variables near to each other in the sorted list, then systematic sampling may reduce the variances of the estimates in much the same way as proportionate stratified sampling does. Systematic sampling from such an ordered list is often described as *implicit stratification*. A general drawback to systematic sampling is that the estimation of the variances of survey estimates requires some form of model assumption.

Stratified Sampling

Often, the sampling frame contains information that may be used to improve the efficiency of the sample design (i.e., reduce the variances of estimators for a given sample size). *Stratification* involves using information available on the sampling frame to partition the population into L classes, or *strata*, and selecting a sample from each stratum. (*See* ►Stratified Sampling).

With *proportionate stratification*, the same sampling fraction (i.e., the ratio of sample size to population size) is used in all the strata, producing an epsem sample design. Proportionate stratification reduces the variances of the survey estimators to the extent that elements within the strata are homogeneous with respect to the survey variables.

With *disproportionate stratification*, different sampling fractions are used in the various strata, leading to a design in which selection probabilities vary. The unequal selection probabilities are redressed by the use of the base weights in the analysis. One reason for using a disproportionate stratified design is to improve the precision of survey estimates when the element standard deviations differ across the strata. Disproportionate stratified samples are widely used in business surveys for this reason, sampling the larger businesses with greater probabilities, and even taking all of the largest businesses into the sample (*see* ►Business Surveys). The allocation of a given overall sample size across strata that minimizes the variance of an overall survey estimate is known as *Neyman allocation*. If data collection costs per sampled element differ across strata, it is more efficient to allocate more of the sample to the strata where data collection costs are lower. The sample allocation that maximizes the precision of an overall survey estimate for a given total data collection cost is termed an *optimum allocation*.

A second common reason for using a disproportionate allocation is to produce stratum-level estimates of adequate precision. In this case, smaller strata are often sampled at above average sampling rates in order to generate sufficiently large sample sizes to support the production of separate survey estimates for them.

Cluster and Multistage Sampling

In many surveys, it is operationally efficient to sample clusters of population elements rather than to sample the elements directly. One reason is that the sampling frame may be a list that comprises clusters of elements, such as a list of households for a survey of persons (the elements). Another reason is that the population may cover a large geographical area; when the survey data are to be collected by face-to-face interviewing, it is then cost-effective to concentrate the interviews in a sample of areas in order to reduce interviewers' travel. The selection of more than one element in a sampled cluster affects the precision of the survey estimates because elements within the same cluster tend to be similar with respect to many of the variables studied in surveys. The homogeneity of elements within clusters is measured by the *intraclass correlation* (*see* ►Intraclass Correlation Coefficient). A positive intraclass correlation decreases the precision of the survey estimates from a cluster sample relative to a SRS with the same number of elements.

When the clusters are small, it is often efficient to include all the population elements in selected clusters, for example, to collect survey data for all persons in sampled households. Such a design is termed a *cluster sample* or more precisely a *single-stage cluster sample* (*see* ►Cluster Sampling).

Subsampling, or the random selection of elements within clusters, may be used to limit the effect of clustering on the precision of survey estimates. Subsampling is widely used when the clusters are large as, for example, is the case with areal units such as counties or census enumeration districts, schools, and hospitals. A sample design in which a sample of clusters is selected, followed by the selection of a subsample of elements within each sampled cluster is referred to as a *two-stage sample*. *Multistage sampling* is an extension of two-stage sampling, in which there are one or more stages of subsampling of clusters within the *first-stage units* (or primary sampling units, PSUs) prior to the selection of elements. In multistage sample designs, a key consideration is the determination of the sample size at each stage of selection. This determination is generally based on cost considerations and the contribution of each stage of selection to the variance of the estimator (*See* ►Multistage Sampling).

In general, large clusters vary considerably in the number of elements they contain. Sampling unequal-sized clusters with equal probabilities is inefficient and, with an overall epsem design, it fails to provide control on the

sample size. These drawbacks may be addressed by sampling the clusters with *probability proportional to size* (PPS) *sampling*. By way of illustration, consider a two-stage sample design. At the first stage, clusters are sampled with probabilities proportional to size, where size refers to the number of elements in a cluster. Then, at the second stage, an equal number of population elements is selected within each PSU. The resulting sample is an epsem sample of elements. This approach extends to multi-stage sampling by selecting a PPS sample of clusters at each stage through to the penultimate stage. At the last stage of selection, an equal number of population elements is selected within each cluster sampled at the prior stage of selection. In practice, the exact cluster sizes are rarely known and the procedure is applied with estimated sizes, leading to what is sometimes called *sampling with probability proportional to estimated size* (PPES).

Two-Phase Sampling

It would be highly beneficial in some surveys to use certain auxiliary variables for sample design, but those variables are not available on the sampling frame. Similarly, it may be beneficial to use certain auxiliary variables at the estimation stage, but the requisite data for the population are not available. In these cases, *two-phase sampling* (also known as *double sampling*) may be useful. As an example, consider the case where, if frame data were available for certain auxiliary variables, stratification based on these variables with a disproportionate allocation would greatly improve the efficiency of the sample design. Under the two-phase sampling approach, at the first phase, data are collected on the auxiliary variables for a larger preliminary sample. The first-phase sample is then stratified based on the auxiliary variables, and a second phase subsample is selected to obtain the final sample. To be effective, two-phase sampling requires that the first phase data collection can be carried out with little effort or resource requirements.

Estimation

As noted above, differential selection probabilities must be accounted for by the use of base weights in estimating the parameters of a finite population. In practice, adjustments are usually made to the base weights to compensate for sample deficiencies and to improve the precision of the survey estimates.

One type of sample deficiency is *unit nonresponse*, or complete lack of response from a sampled element. Compensation for unit nonresponse is typically made by inflating the base weights of similar responding elements in order to also represent the base weights of nonresponding eligible elements (see ►Nonresponse in Surveys, Groves et al. 2001, and Särndal and Lundström 2005).

A second type of deficiency is *noncoverage*, or a failure of the sampling frame to cover some of the elements in the population. Compensation for noncoverage requires population information from an external source. Noncoverage is generally handled through a weighting adjustment using some form of *calibration* adjustment, such as post-stratification (see Särndal 2007). Calibration adjustments also serve to improve the precision of survey estimates that are related to the variables used in calibration.

A third type of deficiency is *item nonresponse*, or the failure to obtain a response to a particular item from a responding element. Item nonresponses are generally accounted for through *imputation*, that is, assigning values for the missing responses (see ►Imputation and Brick and Kalton 1996).

In practice, samples from finite populations are often based on complex designs incorporating stratification, clustering, unequal selection probabilities, systematic sampling, and sometimes, two-phase sampling. The estimation of the variances of the survey estimates needs to take the complex sample design into account. There are two general methods for estimating variances from complex designs, known as the *Taylor Series* or *linearization* method and the *replication* method (including balanced repeated replications, jackknife repeated replications, and the bootstrap). See Wolter (2007) and Rust and Rao (1996). There are several software programs available for analyzing complex sample survey data using each method.

About the Authors

Jill Montaquila is an Associate Director of the statistical staff and a senior statistician at Westat. Dr. Montaquila is also a Research Associate Professor in the Joint Program in Survey Methodology at the University of Maryland. Her statistical interests cover various aspects of complex sample survey methodology, including random digit dialing survey methodology, address based sampling, and evaluations of nonsampling error. Dr. Montaquila has served as President of the Washington Statistical Society and is a Fellow of the American Statistical Association.

Graham Kalton is Chairman of the Board and Senior Vice President at Westat. He is also a Research Professor in the Joint Program in Survey Methodology at the University of Maryland. He has been at Westat since 1992. Earlier positions include: research scientist at the Survey Research Center of the University of Michigan, where he also held titles of Professor of Biostatistics and Professor of Statistics; Leverhulme Professor of Social Statistics at the University of Southampton; and Reader in Social Statistics

at the London School of Economics. Dr. Kalton has wide ranging interests in survey methodology and has published on several aspects of the subject. He has served as President of the International Association of Survey Statisticians and President of the Washington Statistical Society. He is a Fellow of the American Association for the Advancement of Science, a Fellow of the American Statistical Association, a National Associate of the National Academies, and an elected member of the International Statistical Institute.

Cross References

►Cluster Sampling
►Estimation
►Estimation: An Overview
►Horvitz–Thompson Estimator
►Imputation
►Intraclass Correlation Coefficient
►Multiple Imputation
►Multistage Sampling
►Non-probability Sampling Survey Methods
►Nonresponse in Surveys
►Sample Survey Methods
►Simple Random Sample
►Stratified Sampling
►Superpopulation Models in Survey Sampling
►Telephone Sampling: Frames and Selection Techniques
►Total Survey Error

References and Further Reading

Brick JM, Kalton G (1996) Handling missing data in survey research. Stat Methods Med Res 5:215–238

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Deming WE (1960) Sample design in business research. Wiley, New York

Fuller WA (2009) Sampling statistics. Wiley, New York

Groves RM, Dillman DA, Eltinge JA, Little RJA (eds) (2001) Survey nonresponse. Wiley, New York

Hansen MH, Hurwitz WN, Madow WG (1953) Sample survey methods and theory, vols I and II. Wiley, New York

Kish L (1965) Survey sampling. Wiley, New York

Lohr S (2010) Sampling: design and analysis, 2nd edn. Brooks/Cole, Pacific Grove, CA

Murthy MN (1967) Sampling theory and methods. Statistical Publishing Society, Calcutta, India

Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J R Stat Soc 97(4):558–625

Pfeffermann D, Rao CR (eds) (2009) Handbook of statistics. Volume 29A, sample surveys: design, methods and application and volume 29B, sample surveys: inference and analysis. Elsevier, New York

Raj D (1968) Sampling theory. McGraw-Hill, New York

Rust KF, Rao JNK (1996) Variance estimation for complex surveys using replication techniques. Stat Methods Med Res 5:283–310

Särndal CE (2007) The calibration approach in survey theory and practice. Surv Methodol 33:99–119

Särndal CE, Lundström S (2005) Estimation in surveys with nonresponse. Wiley, New York

Särndal CE, Swensson B, Wretman J (1992) Model-assisted survey sampling. Springer-Verlag, New York

Sukhatme PV, Sukhatme BV, Sukhatme S, Asok C (1984) Sampling theory of surveys with applications, 3rd Rev. edn. Iowa State University Press and Indian Society of Agricultural Statistics, Ames, Iowa and New Delhi

Thompson ME (1997) Theory of sample surveys. Chapman and Hall, London

Valliant R, Dorfman AH, Royall RM (2000) Finite population sampling and inference: a prediction approach. Wiley, New York

Wolter K (2007) Introduction to variance estimation, 2nd edn. Springer, New York

Yates F (1981) Sampling methods for censuses and surveys, 4th edn. Charles Griffin, London

Sampling Problems for Stochastic Processes

Masayuki Uchida[1], Nakahiro Yoshida[2]
[1]Professor
Osaka University, Osaka, Japan
[2]Professor
University of Tokyo, Tokyo, Japan

Let $X = (X_t)_{t\in[0,T]}$ be a d-dimensional diffusion process defined by the following stochastic differential equation

$$dX_t = b(X_t, \alpha)dt + \sigma(X_t, \beta)dw_t, \qquad t \in [0, T], \quad X_0 = x_0,$$

where w is an r-dimensional Wiener process, $(\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta$, Θ_α and Θ_β are subsets of $\mathbf{R}^p$ and $\mathbf{R}^q$, respectively. Furthermore, b is an $\mathbf{R}^d$-valued function on $\mathbf{R}^d \times \Theta_\alpha$ and σ is an $\mathbf{R}^d \otimes \mathbf{R}^r$-valued function on $\mathbf{R}^d \times \Theta_\beta$. The drift function b and the diffusion coefficient function σ are known apart from the parameters α and β.

In the asymptotic theory of diffusion processes, the following two types of data are treated: (1) the continuously observed data and (2) the discretely observed data of diffusion processes. Concerning the first order asymptotic theory of diffusion processes based on the continuously observed data, Kutoyants extended Ibragimov and Has'minskii's approach (1981) to semimartingales, and many researchers made contributions to establish the asymptotic theory of semimartingales; see Kutoyants (1984, 1994, 2004) and Küchler and Sørensen (1997), Prakasa Rao (1999a, b) and references therein.

On the other hand, parametric estimation for discretely observed diffusion processes is highly important for

practical applications and now developing progressively. The data are discrete observations at regularly spaced time point on the fixed interval $[0, T]$, that is, $(X_{kh_n})_{0\le k\le n}$ with $nh_n = T$ and h_n is called a discretization step. The discretely observed data are roughly classified into the following three types:

(i) decreasing step size on a fixed interval: the observation time $T = nh_n$ is fixed and the discretization step h_n tends to zero as $n \to \infty$.
(ii) constant step size on an increasing interval: the discretization step is fixed ($h_n = \Delta$) and the observation time $T = nh_n = n\Delta$ tends to infinity as $n \to \infty$.
(iii) decreasing step size on an increasing interval: the discretization step h_n tends to zero and the observation time $T = nh_n$ tends to infinity as $n \to \infty$.

For the setting of type (i), Genon-Catalot and Jacod (1993) proposed estimators of the diffusion coefficient parameter β and they showed that the estimators are consistent, asymptotic mixed normal and asymptotic efficient. For the linearly parametrized case of diffusion coefficient, Yoshida (1997) obtained the asymptotic expansion for the estimator by means of the Malliavin calculus. Gobet (2001) proved the local asymptotic mixed normality for likelihoods by using the Malliavin calculus. On the other hand, for the drift parameter α, we can not generally construct even a consistent estimator under the setting of type (i). However, under the situation where diffusion term is very small, which is called a small diffusion process, we can estimate the drift parameter α. Genon-Catalot (1990) and Laredo (1990) proposed estimators of the drift parameter under the assumption that the diffusion coefficient is known, and they proved that the estimators have consistency, ►asymptotic normality and asymptotic efficiency. Uchida (2008) investigated asymptotic efficient estimators under the general asymptotics. Sørensen and Uchida (2003) obtained estimators of both the drift and the diffusion coefficient parameters simultaneously and investigated the asymptotic properties of their estimators. Gloter and Sørensen (2009) developed the result of Sørensen and Uchida (2003) under the general asymptotics.

As concerns the type (ii), Bibby and Sørensen (1995) proposed martingale estimating functions and obtained the estimators of the drift and the diffusion coefficient parameters from the martingale estimating functions. They proved that both estimators have consistency and asymptotic normality under ergodicity. Masuda (2005) showed the asymptotic normality of the moment estimator for a state space model involving jump noise terms.

Under the setting of type (iii), Prakasa-Rao (1983, 1988) are early work. As seen in Yoshida (1992a), the estimators of α and β jointly converge, and they are asymptotically orthogonal, however their convergence rates are different. Those authors' estimators are of maximum likelihood type in their settings. Kessler (1997) improved the condition on the sampling scheme and gave generalization. Gobet (2002) showed local asymptotic normality for the likelihood. A polynomial type large deviation inequality for an abstract statistical random field, which includes likelihood ratios of stochastic processes, enables to obtain the asymptotic behaviors of the Bayes and maximum likelihood type estimators; see Yoshida (2010) for details. For the asymptotic theory of diffusion processes with jumps, see for example Shimizu and Yoshida (2006).

Regarding the higher order asymptotic theory of diffusion processes, the asymptotic expansions have been studied; see Yoshida (1992b, 1997), Sakamoto and Yoshida (2004) and recent papers.

About the Author

Professor Nakahiro Yoshida was awarded the first Research Achievement Award by the Japan Statistical Society, for studies in the theory of statistical inference for stochastic processes and their applications (2007). He has also received the Analysis Prize, Mathematical Society of Japan (2006) and the Japan Statistical Society Award, Japan Statistical Society (2009). Professor Yoshida is Section Editor and Scientific Secretary, Bernoulli Society for Mathematical Statistics and Probability. Professors Nakahiro Yoshida and Masayuki Uchida are Associate editors of the *Annals of the Institute of Statistical Mathematics*.

Cross References

►Asymptotic Normality
►Brownian Motion and Diffusions
►Local Asymptotic Mixed Normal Family
►Stochastic Differential Equations
►Stochastic Processes
►Stochastic Processes: Classification

References and Further Reading

Bibby BM, Sørensen M (1995) Martingale estimating functions for discretely observed diffusion processes. Bernoulli 1:17–39
Genon-Catalot V (1990) Maximum contrast estimation for diffusion processes from discrete observations. Statistics 21:99–116
Genon-Catalot V, Jacod J (1993) On the estimation of the diffusion coefficient for multidimensional diffusion processes. Ann Inst Henri Poincaré Prob Stat 29:119–151

Gloter A, Sørensen M (2009) Estimation for stochastic differential equations with a small diffusion coefficient. Stoch Process Appl 119:679–699
Gobet E (2001) Local asymptotic mixed normality property for elliptic diffusion: a Malliavin calculus approach. Bernoulli 7:899–912
Gobet E (2002) LAN property for ergodic diffusions with discrete observations. Ann Inst Henri Poincaré Prob Stat 38:711–737
Ibragimov IA, Has'minskii RZ (1981) Statistical estimation. Springer Verlag, New York
Kessler M (1997) Estimation of an ergodic diffusion from discrete observations. Scand J Stat 24:211–229
Kutoyants YuA (1984) In: Prakasa Rao BLS (ed) Parameter estimation for stochastic processes. Heldermann, Berlin
Kutoyants YuA (1994) Identification of dynamical systems with small noise. Kluwer Dordrecht
Kutoyants YuA (2004) Statistical inference for ergodic diffusion processes. Springer-Verlag, London
Küchler U, Sørensen M (1997) Exponential families of stochastic processes. Springer, New York
Laredo CF (1990) A sufficient condition for asymptotic sufficiency of incomplete observations of a diffusion process. Ann Stat 18:1158–1171
Masuda H (2005) Classical method of moments for partially and discretely observed ergodic models. Stat Inference Stoch Proc 8:25–50
Prakasa Rao BLS (1983) Asymptotic theory for nonlinear least squares estimator for diffusion processes. Math Oper Forsch Stat Ser Stat 14:195–209
Prakasa Rao BLS (1988) Statistical inference from sampled data for stochastic processes. In: Contemporary mathematics, vol 80. American Mathematical Society, Providence, RI, pp 249–284
Prakasa Rao BLS (1999a) Statistical inference for diffusion type processes. Arnold, London
Prakasa Rao BLS (1999b) Semimartingales and their statistical inference. Boca Rotan, FL, Chapman & Hall/CRC
Sakamoto Y, Yoshida N (2004) Asymptotic expansion formulas for functionals of ϵ-Markov processes with a mixing property. Ann Inst Stat Math, 56:545–597
Shimizu Y, Yoshida N (2006) Estimation of parameters for diffusion processes with jumps from discrete observations. Stat Inference Stoch Process 9:227–277
Sørensen M, Uchida M (2003) Small-diffusion asymptotics for discretely sampled stochastic differential equations. Bernoulli 9:1051–1069
Uchida M (2008) Approximate martingale estimating functions for stochastic differential equations with small noises. Stoch Process Appl 118:1706–1721
Yoshida N (1992a) Estimation for diffusion processes from discrete observation, J Multivar Anal 41:220–242
Yoshida N (1992b) Asymptotic expansions of maximum likelihood estimators for small diffusions via the theory of Malliavin-Watanabe. Probab Theory Relat Field 92:275–311
Yoshida N (1997) Malliavin calculus and asymptotic expansion for martingales. Probab Theory Relat Fields 109:301–342
Yoshida N (2010) Polynomial type large deviation inequalities and quasi-likelihood analysis for stochastic differential equations. Ann Inst Stat Math, doi:10.1007/s10463-009-0263-z

Scales of Measurement

Karl L. Wuensch
Professor
East Carolina University, Greenville, NC, USA

Measurement involves the assignment of scores (numbers or other symbols) to entities (objects or events) in such a way that the scores carry information about some characteristic of the measured entities. With careful consideration of the method by which the scores have been assigned, one can classify the method of measurement as belonging to one or more "scales of measurement." S.S. Stevens (1951) defined four scales of measurement: nominal, ordinal, interval, and ratio. Membership in one or more of these categories depends on the extent to which empirical relationships among the measured entities correspond to numerical relationships among the scores.

If the method of measurement produces scores that allow one to determine whether the measured entities are or are not equivalent on the characteristic of interest, then the scale is referred to as "nominal." For example, I ask the students in my class to take out all of their paper money, write their university identification number on each bill, and deposit all the bills in a bag. I then shake the bag and pull out two bills. From the identification numbers on the bills, I can determine whether or not the same student contributed both bills. The attribute of interest is last ownership of the bill, and the scores allow one to determine whether or not two bills are equivalent on that characteristic – accordingly, the identification number scores represent a nominal scale. "Nominal" derives from the Latin "nomen," name. Nominal scores may be no more than alternative names for entities.

If the scores can be employed to determine whether two entities are equivalent or not on the measured characteristic and, if they are not equivalent, which entity has the greater amount of the measured characteristic, then the scale is "ordinal." The order of the scores is the same as the order of the true amounts of the measured attribute. The identification numbers my students wrote on their bills would not allow one to determine whether "004387" represents more or less of something than does "093752." Imagine that I throw all the money out the window and then invite the students to retrieve the bills. My associate, outside, assigns to the students the ordinal scores shown in Table 1. The measured attribute is time taken to retrieve

Scales of Measurement. Table 1 Relationship between true scores and observed scores

Entity	A	B	C	D	E
True Score	1.0	2.0	4.0	8.0	9.0
Ordinal Score	0.5	0.6	0.7	1.1	1.5
Interval Score	12.0	14	18.0	26.0	28.0
Ratio Score	2.0	4.0	8.0	16.0	18.0

a bill, and the order of the scores is the same as the order of the magnitudes of the measured attribute. If Student A obtains a score of .5 and Student B a score of .6, I am confident that they differ on retrieval time and that Student B took longer than Student A.

Scale of measurement can be inferred from the nature of the relationship between the "observed scores" (the measurements) and the "true scores" (the true amounts of the measured characteristic) (Winkler and Hays 1975, pp. 277–282). If that relationship is positive monotonic, then the scale of measurement is ordinal. Notice that the ordinal scores in Table 1 are related to the true scores in a positive monotonic fashion.

The ordinal scores in Table 1 do not allow one to establish the equivalence of differences or to order differences. Consider the differences between A and B and between D and E. The true scores show that the differences are equivalent, but the ordinal scores might lead one to infer that the difference between D and E is greater than the difference between A and B. Also, the ordinal scores might lead one to infer that the difference between C and D (0.4) is equivalent to the difference between D and E (0.4), but the true scores show that not to be true.

If the relationship between the observed scores and the true scores is not only positive monotonic but also linear, then one will be able to establish the equivalence of differences and will be able to order differences. Such a scale is called "interval." My hypothetical associate used a mechanical device to measure the retrieval times, obtaining the interval scores in Table 1. From these observed scores, one would correctly infer that the difference between A and B is equivalent to the difference between D and E and that the difference between C and D is greater than the difference between D and E.

For the interval scores in Table 1, the function relating the measurements (m) to the true scores (t) is $m = 10 + 2t$. This hypothetical interval scale does not have a "true zero point." That is, it is not true that an entity that has absolutely none of the measured characteristic will obtain a measurement of zero. In this case, it will obtain a measurement of 10. This is problematic if one wishes to establish the equivalences of and orders of ratios of measurements. With the interval data one might infer that the ratio D/C > C/B > B/A, but the true scores show that these ratios are all equivalent. To achieve a ratio scale, the function relating the measurements to the true scores must not only be positive linear but also must have an intercept of zero. For the hypothetical ratio data in Table 1, that function is $m = 0 + 2t$. With the ratio scale the ratios of observed scores are identical to the corresponding ratios of the true scores.

Stevens (1951) argued that scale of measurement is an important consideration when determining the type of statistical analysis to be employed. For example, the mode was considered appropriate for any scale, even a nominal scale. If a fruit basket contains five apples, four oranges, and nine bananas, the modal fruit is a banana. The median was considered appropriate for any scale that was at least ordinal. Imagine that we select five fruits, identified as A, B, C, D, and E. Their true weights are 1.5, 3, 4.5, 9, and 27, and their ordinal scores are 1, 2, 3, 4, and 5. The entity associated with the median is C regardless of whether you use the true scores of the ordinal scores. Interval scores 4, 7, 10, 19, and 55 have a linear relationship with the true scores, $m = 1 + 2t$. The mean true score, 9, is associated with Entity D, and the mean interval score, 19, is also associated with Entity D. With the ordinal scores, however, the mean score, 3, is associated with Entity B.

There has been considerable controversy regarding the role that scale of measurement should play when considering the type of statistical analysis to employ. Most controversial has been the suggestion that parametric statistical analysis is appropriate only with interval or ratio data, but that nonparametric analysis can be employed with ordinal data. This proposition has been attacked by those who opine that the only assumptions required when employing parametric statistics are mathematical, such as homogeneity of variance and normality (Gaito 1980; Velleman and Wilkinson 1993). Defenders of the measurement view have argued that researchers must consider scale of measurement, the relationship between true scores and observed scores, because they are interested in making inferences about the constructs underlying the observed scores (Maxwell and Delaney 1985; Townsend and Ashby 1984). Tests of hypotheses that groups have identical means

on an underlying construct or that the Pearson ρ between two underlying constructs is zero do not require interval level data given the usual assumptions of homogeneity of variance and normality, but with non-interval data the effect size estimates will not apply to the underlying constructs (Davison and Sharma 1988).

When contemplating whether the observed scores to be analyzed represent an interval scale or a non-interval, ordinal scale, one needs makes a decision about the nature of the relationship between the true scores and the observed scores. If one conceives of true scores as part of some concrete reality, the decision regarding scale of measurement may come down to a matter of faith. For example, how could one know with certainty whether or not the relationship between IQ scores and true intelligence is linear? One way to avoid this dilemma is to think of reality as something that we construct rather than something we discover. One can then argue that the results of parametric statistical analysis apply to an abstract reality that is a linear function of our measurements. Conceptually, this is similar to defining a population on a sample rather than the other way around – when we cannot obtain a true random sample from a population, we analyze the data we can obtain and then make inferences about the population for which our data could be considered random.

About the Author

For biography *see* the entry ►Chi-Square Tests.

Cross References

►Rating Scales
►Scales of Measurement and Choice of Statistical Methods
►Variables

References and Further Reading

Davison ML, Sharma AR (1988) Parametric statistics and levels of measurement. Psychol Bull 104:137–144
Gaito J (1980) Measurement scales and statistics: resurgence of an old misconception. Psychol Bull 87:564–567
Maxwell SE, Delaney HD (1985) Measurement and statistics: an examination of construct validity. Psychol Bull 97:85–93
Stevens SS (1951) Mathematics, measurement, and psychophysics. In: Stevens SS (ed) Handbook of experimental psychology. Wiley, New York, pp 1–49
Townsend JT, Ashby FG (1984) Measurement scales and statistics: the misconception misconceived. Psychol Bull 96:394–401
Velleman PF, Wilkinson L (1993) Nominal, ordinal, interval, and ratio typologies are misleading. Am Stat 47:65–72
Winkler RL, Hays WL (1975) Statistics: probability, inference, and decision, 2nd edn. Holt Rinehart and Winston, New York

Scales of Measurement and Choice of Statistical Methods

Donald W. Zimmerman
Professor Emeritus
Carleton University, Ottawa, ON, Canada

During the last century, it was conventional in many disciplines, especially in psychology, education, and social sciences, to associate statistical methods with a hierarchy of levels of measurement. The well-known classification proposed by Stevens (1946) included nominal, ordinal, interval, and ratio scales, defined by increasingly stronger mathematical restrictions. It came to be generally believed that the use of statistical significance tests in practice required choosing a test to match the scale of measurement responsible for the data at hand. Classes of appropriate statistical methods were aligned with the hierarchy of levels of measurement.

In research studies in psychology and education, the most relevant distinction perhaps was the one made between interval scales and ordinal scales. The Student *t* test (see ►Student's *t* Tests), the ANOVA *F* test, and regression methods were deemed appropriate for interval measurements, and nonparametric tests, such as the ►Wilcoxon–Mann–Whitney test and the Kruskal–Wallis test were appropriate for ordinal measurements.

Despite the widespread acceptance of these ideas by many statisticians and researchers, there has been extensive controversy over the years about their validity (see, for example, Cliff and Keats 2003; Maxwell and Delaney 1985; Michell 1986; Rozeboom 1966; Velleman and Wilkinson 1993; Zimmerman and Zumbo 1993). The mathematical theory eventually included more refined definitions of scales of measurement and additional types of scales (Luce 2001; Narens 1981), but the fourfold classification persisted for a long time in textbooks and research articles.

Scales of Measurement and Distributional Assumptions

The derivation of all significance tests is based on assumptions about probability distributions, such as independence, normality, and equality of the variances of separate groups, and some tests involve more restrictive assumptions than others. In many textbooks and research papers, the requirement of a specific level of measurement was placed on the same footing as these

distributional assumptions made in the mathematical derivation of a test statistic. For example, the Student *t* test and ANOVA *F* test were widely believed to assume three things: normality, homogeneity of variance, and interval measurement, while a nonparametric test such as the Wilcoxon–Mann–Whitney test is presumably free from the two distributional assumptions and requires only ordinal measurement. The assumption of within-sample independence is part of the definition of random sampling, and it is typically taken for granted that the data at hand meets that requirement before a test is chosen.

Many researchers believed that the parametric methods are preferable when all assumptions are satisfied, because nonparametric tests discard some information in the data and have less power to detect differences. Furthermore, the parametric methods were considered to be robust in the sense that a slight violation of assumptions does not lessen their usefulness in practical research. Early simulation studies, such as the one by Boneau (1960), were consistent with these ideas.

Some complications arose for the orderly correspondence of scales and statistics when researchers began to investigate how the Type I and Type II errors of both parametric and nonparametric significance tests depend on properties of standard probability densities. It was found that the nonparametric tests were often more powerful than their parametric counterparts for quite a few continuous densities, such as the exponential, lognormal, mixed-normal, Weibull, extreme value, chi-square, and others familiar in theoretical statistics. The power advantage of the nonparametric tests often turned out to be quite large (see, for example, Blair and Higgins 1980; Lehmann 1975; Randles and Wolfe 1979; Sawilowsky and Blair 1992; Zimmerman and Zumbo 1993). The superiority of nonparametric rank methods for many types of non-normal data has been extensively demonstrated by many simulation studies.

It can be argued that samples from one of these continuous densities by definition conform to interval measurement. That is, equal intervals are assumed in defining the parameters of the probability density. For this reason it is legitimate to employ *t* and *F* tests of location on sequences of random variates generated by different computer programs and obtain useful information. Similarly, the scaling criteria imply that calculation of means and variances is appropriate only for interval measurement, but it has become clear that slight violations of "homogeneity of variance" have severe consequences for both parametric and nonparametric tests.

Rank Transformations and Appropriate Statistics

In the controversies surrounding the notion of levels and measurement, theorists have tended to overlook the implications of a procedure known as the *rank transformation*. It was discovered that the *large-sample* normal approximation form of the Wilcoxon–Mann–Whitney test is equivalent to the Student *t* test performed on ranks replacing the original scores and that the Kruskal–Wallis test is equivalent to the ANOVA *F* test on ranks (Conover and Iman 1981). In the Wilcoxon–Mann–Whitney test, two samples of scores of size n_1 and n_2 are combined and converted to a single series of ranks, that is, integers from 1 to n_1+n_2. Similarly, in one-way ANOVA, scores in *k* groups are combined and converted to $n_1 + n_2 + \ldots + n_k$ ranks. Then, the scores in the original samples are replaced by their corresponding ranks in the combined group.

The above equivalence means that this rank transformation followed by the usual Student *t* test on the ranks replacing the initial scores leads to the same statistical decision as calculating and comparing rank sums, as done by a Wilcoxon–Mann–Whitney test. The Type I and Type II error probabilities turn out to be the same in both cases. That is true irrespective of the distributional form of the original data. If a Student *t* test performed on ranks is not appropriate for given data, then the Wilcoxon–Mann–Whitney test is not appropriate either, and vice versa.

Considered together with the power superiority of nonparametric tests for various non-normal densities, these findings imply that the power of *t* and *F* tests often can be increased by transforming interval data to ordinal data. Arguably, the main benefit of converting to ranks is not a change in scale, but rather augmentation of the robustness of the *t* and *F* tests. At first glance it seems paradoxical that statistical power can be increased, often substantially, by discarding information. However, one should bear in mind that conversion to ranks not only replaces real numbers by integers, but also alters the shape of distributions. Whatever the initial form of the data, ranks have a rectangular distribution, and, as noted before, the shape of non-normal distributions, especially those with heavy tails and extreme outlying values, certainly influences the power, or the extent of the loss of power, of significance tests.

Otherwise expressed, changing the distributional form of the data before performing a significance test appears to be the source of the power advantages, not the details of calculating rank-sums and finding quantiles of the resulting test statistic from a unique formula.

The rank transformation concept, together with the fact that unequal variances of scores in several groups is inherited by unequal variances of the corresponding ranks in the same groups, also provides a rationale for the dependence of both parametric and nonparametric tests on homogeneity of variance (Zimmerman 1996).

Another finding that is difficult to reconcile with notions of scaling is the fact that the beneficial properties of rank tests can be maintained despite alteration of the ranks in a way that modifies the scale properties, sometimes substantially. For example, small random numbers can be added to ranks, or the number of ranks can be reduced in number, with little effect on the power of the *t* and *F* tests under a rank transformation. That is, combining ranks 1, 2, 3, and 4 all into the value 1, ranks 5, 6, 7, and 8 into the value 2, and so on, has little influence on the power of the test when sample sizes are moderately large.

A quick illustration of these properties of scores and ranks is provided by Table 1, which gives the probability of rejecting H_0 by three significance tests at the 0.05 level. These computer simulations consisted of 50,000 pairs of independent samples of size 50 from normal and seven non-normal distributions, generated by a *Mathematica* program. The columns, labeled *t* represent the Student *t* test, those labeled *W* are the Wilcoxon–Mann–Whitney test, and those labeled *m* are the *t* test performed on *modified ranks*.

In this modification, all scores from both groups were combined and ranked as usual. Then, instead of transforming to integers, each original score was replaced by the median of all higher scores in the ranking; that is, the lowest score, ranked 1, was replaced by the median of all the higher scores ranked from 2 to $n_1 + n_2$, the score ranked 2 was replaced by the median of scores ranked from 3 to $n_1 + n_2$, and so on. Finally, the scores in the two initial groups were replaced by their corresponding modified ranks, and the significance test was performed.

This procedure resulted in a kind of hybrid ordinal/interval data not too different from ordinary ranks, whereby the real values of the original scores were retained, the distribution shape was compressed, and ▶outliers were eliminated. Table 1 shows that the Type I error rates of the *t* test on these modified ranks were close to those of ordinary ranks for the various distributions. Moreover, the *t* test on the modified values was nearly as powerful as the Wilcoxon–Mann–Whitney test for two distributions where the ordinary *t* test is known to be superior, and it was considerably more powerful than the *t* test and somewhat more powerful than the Wilcoxon–Mann–Whitney test for distributions for which the nonparametric test is known to be superior.

All these facts taken together imply there is not a one-to-one correspondence between the hierarchy of levels of measurement and methods that are appropriate for making correct statistical decisions. Transforming data so that it conforms to the assumptions of a significance test is not itself unusual, because for many years statisticians employed square-root, reciprocal, and logarithmic

Scales of Measurement and Choice of Statistical Methods. Table 1 Type I error rates and power of Student t test, Wilcoxon–Mann–Whitney test, and t test on modified ranks, 50,000 iterations at 0.05 level, samples from normal and seven non-normal distributions

	$\mu_1 - \mu_2 = 0$			$\mu_1 - \mu_2 = 0.3\sigma$			$\mu_1 - \mu_2 = 0.6\sigma$		
Distribution	t	W	m	t	W	m	t	W	m
Normal	0.051	0.051	0.053	0.314	0.298	0.295	0.847	0.831	0.812
Exponential	0.049	0.049	0.050	0.331	0.615	0.689	0.842	0.978	0.995
Mixed-normal	0.052	0.051	0.050	0.336	0.952	0.967	0.842	1.000	1.000
Lognormal	0.041	0.051	0.051	0.393	0.913	0.962	0.841	0.999	1.000
Extreme value	0.048	0.048	0.049	0.329	0.380	0.426	0.842	0.899	0.934
Uniform	0.049	0.048	0.049	0.311	0.294	0.310	0.845	0.798	0.840
Half-normal	0.049	0.050	0.051	0.318	0.385	0.420	0.837	0.890	0.943
Chi-square	0.049	0.049	0.050	0.326	0.489	0.551	0.845	0.958	0.987

transformations. The rank transformation can be regarded as a member of the same broad class of methods as those procedures. Unlike those methods, it is not continuous and has no inverse. That can be an advantage, because, by substituting small integers for intervals of real numbers, it lessens skewness and eliminates outliers.

As we have seen, the rank transformation in several instances is *equivalent* to a corresponding nonparametric test, in the sense that both either reject or fail to reject H_0 for given data. The earlier normalizing transformations do not possess such equivalences with well-known nonparametric methods. Each is best suited to a specific problem, such as stabilizing variances or changing the shape of a particular distribution, whereas conversion to ranks is an omnibus transformation that always brings data into a rectangular form with no outliers. Also, it is possible to reverse the perspective and regard the Wilcoxon–Mann–Whitney test and the Kruskal–Wallis test as having an affinity with those normalizing transformations, because the conversion to ranks, not the specific formula used in calculations, is apparently what makes the difference.

Conclusion

When all is said and done, *the theory of scales of measurement, although interesting and informative in its own right, is not closely related to practical decision-making in applied statistics*. Present evidence suggests that the mathematical property most relevant to choice of statistics in research is the probability distribution of the random variable that accounts for the observed data.

Caution is needed in making choices, and the rationale for a decision is likely to be more subtle and complex than the prescriptions in textbooks and software packages. In practice, the shape of a population distribution is not usually known with certainty. The degree of violation of assumptions fluctuates from sample to sample along with the estimates of the parameters, no matter what the population may be and what measurement procedures are used. Basing the choice of an appropriate test on inspection of samples, or even on preliminary significance tests performed to assess the validity of assumptions, can lead to incorrect statistical decisions with high probability.

About the Author

Dr. Donald W. Zimmerman is Professor Emeritus of Psychology at Carleton University in Ottawa, Ontario, Canada, and is currently living in Vancouver, British Columbia, Canada. He is the author of over 160 papers in peer-reviewed psychological and statistical journals, and he has served as an editorial consultant and reviewed manuscripts for 13 journals. His teaching and research interests have been in the areas of test theory, statistics, learning theory and conditioning, and the philosophy of science.

Cross References

►Analysis of Variance
►Nonparametric Models for ANOVA and ANCOVA Designs
►Parametric Versus Nonparametric Tests
►Rank Transformations
►Scales of Measurement
►Significance Testing: An Overview
►Statistical Fallacies: Misconceptions, and Myths
►Student's *t*-Tests
►Validity of Scales
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Blair RC, Higgins JJ (1980) The power of *t* and Wilcoxon statistics: a comparison. Eval Rev 4:645–655

Boneau CA (1960) The effects of violation of assumptions underlying the *t*- test. Psychol Bull 57:49–64

Cliff N, Keats JA (2003) Ordinal measurement in the behavioral sciences. Mahwah, Erlbaum, NJ

Conover WJ, Iman RL (1981) Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 35:124–129

Lehmann EL (1975) Nonparametrics: statistical methods based on ranks. Holden-Day, San Francisco

Luce RD (2001) Conditions equivalent to unit representations of ordered relational structures. J Math Psychol 45:81–98

Maxwell SE, Delaney HD (1985) Measurement and statistics: an examination of construct validity. Psychol Bull 97:85–93

Michell J (1999) Measurement in psychology – a critical history of a methodological concept. Cambridge University Press, Cambridge

Narens L (1981) On the scales of measurement. J Math Psychol 24:249–275

Randles RH, Wolfe DA (1979) Introduction to the theory of nonparametric statistics. Wiley, New York

Rozeboom WW (1966) Scaling theory and the nature of measurement. Synthese 16:170–233

Sawilowsky SS, Blair RC (1992) A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. Psychol Bull 111:352–360

Stevens SS (1946) On the theory of scales of measurement. Science 103:677–680

Velleman PJ, Wilkinson L (1993) Nominal, ordinal, interval, and ratio typologies are misleading. Am Stat 47:65–72

Zimmerman DW (1996) A note on homogeneity of variance of scores and ranks. J Exp Educ 4:351–362

Zimmerman DW, Zumbo BD (1993) The relative power of parametric and nonparametric statistical methods. In: Keren G, Lewis C (eds) A handbook for data analysis in the behavioral sciences. Erlbaum, Mahwah, NJ

Seasonal Integration and Cointegration in Economic Time Series

Svend Hylleberg
Professor, Dean of the Faculty of Social Sciences
University of Aarhus, Aarhus C, Denmark

Introduction

A simple filter often applied in empirical econometric work is the seasonal difference filter $(1 - L^s)$, where s is the number of observations per year, typically $s = 2, 4, 12$ or 52. The seasonal differencing assumes that there are unit roots at all the seasonal frequencies. The seasonal difference filter can be written as the product of $(1 - L)$ and the seasonal summation filter $S(L)$, which for quarterly data is $S(L) = (1+L+L^2+L^3)$. The quarterly seasonal summation filter has the real root -1 and the two complex conjugate roots $\pm i$.

The existence of seasonal unit roots in the data generating process implies a varying seasonal pattern where "Summer may become Winter." In most cases, such a situation is not feasible and the findings of seasonal unit roots should be interpreted with care and taken as an indication of a varying seasonal pattern, where the unit root model is a parsimonious approximation and not the true DGP.

The idea that the seasonal components of a set of economic time series are driven by a smaller set of common seasonal features seems a natural extension of the idea that the trend components of a set of economic time series are driven by common trends. In fact, the whole business of seasonal adjustment may be interpreted as an indirect approval of such a view.

If the seasonal components are integrated, the idea immediately leads to the concept of seasonal cointegration, introduced in the paper by Hylleberg et al. (1990). In case the seasonal components are stationary, the idea leads to the concept of seasonal common features, see Engle and Hylleberg (1996), while so-called periodic cointegration considers cointegration season by season, introduced by Birchenhal et al. (1989). For a recent survey see Brenstrup et al. (2004).

Seasonal Integration

In general, consider the autoregressive representation $\phi(L)y_t = \varepsilon_t$, $\varepsilon_t \sim iid(0, \sigma^2)$, where $\phi(L)$ is a finite lag polynomial. Suppose $\phi(L)$ has all its roots outside the unit circle except for possible unit roots at the long-run frequency $\omega = 0$ corresponding to $L = 1$, semiannual frequency $\omega = \pi$ corresponding to $L = -1$, and annual frequencies $\omega = \left\{\frac{\pi}{2}, \frac{3\pi}{2}\right\}$ corresponding to $L = \pm i$.

Dickey et al. (1984) suggested a simple test for seasonal unit roots in the spirit of the ▶Dickey – Fuller test for long-run unit roots. They suggested estimating the auxiliary regression $\left(1 - L^4\right) y_t = \pi_0 y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim iid(0, \sigma^2)$. The DHF test statistic is the "t-value" corresponding to π_0, which has a non-standard distributed tabulated in Dickey et al. (1984). This test, however, is a joint test for unit roots at the long-run and all the seasonal frequencies.

In order to construct a test for each individual unit root and overcome the lack of flexibility in the DHF test, Hylleberg et al. (1990) refined this idea. By use of the result that any lag polynomial of order p, $\phi(L)$, with possible unit roots at each of the frequencies $\omega = 0, \pi, [\pi/2, 3\pi/2]$, can be written as $\phi(L) = \sum_{k=1}^{4} \frac{\xi_k \Delta(L)(1-\delta_k(L))}{\delta_k(L)} + \phi^*(L)\Delta(L)$, $\delta_k(L) = 1 - \frac{1}{\varsigma_k}L$, $\varsigma_k = 1, -1, i, -i$, $\Delta(L) = \Pi_{k=1}^{4}\delta_k(L)$, where ξ_k is a constant and $\phi^*(z) = 0$ has all its roots outside the unit circle, it can be shown that the autoregression can be written in the equivalent form

$$\phi^*(L)y_{4t} = \pi_1 y_{1t-1} + \pi_2 y_{2t-1} + \pi_3 y_{3t-2} + \pi_4 y_{3t-1} + \varepsilon_t. \quad (1)$$

where $y_{1t} = \left(1 + L + L^2 + L^3\right) y_t = (1 + L)(1 + L^2)y_t$, $y_{2t} = -\left(1 - L + L^2 - L^3\right) y_t = -(1 - L)(1 + L^2)y_t$, $y_{3t} = -\left(1 - L^2\right) y_t = -(1 - L)(1 + L)y_t$, and $y_{4t} = \left(1 - L^4\right) y_t = (1-L)(1+L)(1+L^2)y_t$. Notice that, in this representation, $\phi^*(L)$ is a stationary and finite polynomial if $\phi(L)$ only has roots outside the unit circle except for possible unit roots at the long-run, semiannual , and annual frequencies.

The HEGY tests of the null hypothesis of a unit root are now conducted by simple "t-value" tests on π_1 for the long-run unit root, π_2 for the semiannual unit root, and "F-value" tests on π_3, π_4 for the annual unit roots. As in the Dickey–Fuller and DHF models, the statistics are not t or F distributed but have non-standard distributions. Critical values for the "t" tests are tabulated in Fuller (1976) while critical values for the "F" test are tabulated in Hylleberg et al. (1990).

Tests for combinations of unit roots at the seasonal frequencies are suggested by Ghysels et al. (1994). See also Ghysels and Osborn (2001), who correctly argue that if the null hypothesis is four unit roots, i.e., the proper transformation is $(1 - L^4)$, the test applied should be an "F-test" of π_i, $i = 1, 2, 3, 4$, all equal to zero.

As in the Dickey–Fuller case the correct lag-augmentation in the auxiliary regression (1) is crucial. The errors need to be rendered white noise in order for the size to be close to the stipulated significance level, but the use of too many lag coefficients reduces the power of the tests.

Obviously, if the data generating process, the DGP, contains a moving average component, the augmentation of the autoregressive part may require long lags, see Hylleberg (1995). As is the case for the Dickey-Fuller test, the HEGY test may be seriously affected by moving average terms with roots close to the unit circle, but also one time jumps in the series, often denoted structural breaks in the seasonal pattern, and noisy data with ▶outliers may cause problems.

A straightforward extension of the HEGY test for quarterly data produces tests for semiannual and monthly data, see Franses (1991) However the extension to weekly or daily data is not possible in practice due to number of regressors in the auxiliary regressions.

The results of a number of studies testing for seasonal unit roots in economic data series suggest the presence of one or more seasonal unit roots, but often not all required for the application of the seasonal difference filter, $(1-L^s)$, or the application of the seasonal summation filter, $S(L)$. Thus, these filters should be modified by applying a filter which removes the unit roots at the frequencies where they were found, and not at the frequencies where no unit roots can be detected. Another and maybe more satisfactory possibility would be to continue the analysis applying the theory of seasonal cointegration.

Seasonal Cointegration

Seasonal cointegration exists at a particular seasonal frequency if at least one linear combination of series, which are seasonally integrated at the particular frequency, is integrated of a lower order. For ease of exposition we will concentrate on quarterly time series integrated of order 1. Quarterly time series may have unit roots at the annual frequency $\pi/2$ with period 4 quarters, at the semiannual frequency π with period 2 quarters, and/or at the long-run frequency 0. The cointegration theory at the semiannual frequency, where the root on the unit circle is real, is a straightforward extension of the cointegration theory at the long run frequency. However, the complex unit roots at the annual frequency leads to the concept of polynomial cointegration, where cointegration exists if one can find at least one linear combination including a lag of the seasonally integrated series which is stationary.

In Hylleberg et al. (1990) seasonal cointegration was analyzed along the path set up in Engle and Granger (1987). Consider the quarterly VAR model $\Pi(L)X_t = \varepsilon_t, t = 1,2,....T$, where $\Pi(L)$ is a $p \times p$ matrix of lag polynomials of finite dimension, X_t is a $p \times 1$ vector of observations on the demeaned variables, while the $p \times 1$ disturbance vector is $\varepsilon_t \sim NID(0,\Omega)$. Under the assumptions that the p variables are integrated at the frequencies $0, \pi/2, 3\pi/2$, and π, and that cointegration exists at these frequencies as well, the VAR model can be rewritten as a seasonal error correction model

$$\begin{aligned}\Phi(L)X_{4t} &= \Pi_1 X_{1,t-1} + \Pi_2 X_{2,t-1} + \Pi_3 X_{3,t-2} + \Pi_4 X_{3,t-1} + \varepsilon_t,\\ \Pi_1 &= \alpha_1\beta_1', \Pi_2 = \alpha_2\beta_2', \Pi_3 = \alpha_4\beta_4' - \alpha_3\beta_3',\\ \Pi_3 &= \alpha_4\beta_3' + \alpha_3\beta_4', \end{aligned} \tag{2}$$

where the transformed $p \times 1$ vectors $X_{j,t}, j = 1,2,3,4$, are defined as in a similar way as $y_{j,t}, j = 1,2,3,4$ above, and where $Z_{1t} = \beta_1' X_{1t}$ and $Z_{2t} = \beta_2' X_{2t}$ contain the cointegrating relations at the long-run and semiannual frequencies, respectively, while $Z_{3t} = \left(\beta_3' + \beta_4' L\right) X_{3t}$ contains the polynomial cointegrating vectors at the annual frequency. In Engle et al. (1993) seasonal and non-seasonal cointegrating relations were analyzed between the Japanese consumption and income, estimating the relations for $Z_{jt}, j = 1,2,3$, in the first step following the Granger-Engle two step procedure.

The well known drawbacks of this method, especially when the number of variables included exceeds two, is partly overcome by Lee (1992) who extended the maximum likelihood based methods of Johansen (1988) for cointegration at the long run frequency, to cointegration at the semiannual frequency π.

To adopt the ML based cointegration analysis at the annual frequency $\pi/2$ with the complex pair of unit roots $\pm i$, is somewhat more complicated, however.

To facilitate the analysis, a slightly different formulation of the seasonal error correction model is given in Johansen and Schaumburg (1999). In our notation the formulation is

$$\begin{aligned}\Phi(L)X_{4t} &= \alpha_1\beta_1' X_{1,t-1} + \alpha_2\beta_2' X_{2,t-1} + \alpha_*\beta_*' X_{*,t}\\ &\quad + \alpha_{**}\beta_{**}' X_{**,t} + \varepsilon_t\\ 2\alpha_* &= \alpha_3 + i\alpha_4, 2\alpha_{**} = \alpha_3 - i\alpha_4, \beta_* = \beta_3 + i\beta_4, \beta_{**}\\ &= \beta_3 - i\beta_4\\ X_{*,t} &= (X_{t-2} - X_{t-4}) + i(X_{t-1} - X_{t-3})\\ &= -X_{3,t-2} - iX_{3,t-1}\\ X_{**,t} &= (X_{t-2} - X_{t-4}) - i(X_{t-1} - X_{t-3})\\ &= -X_{3,t-2} + iX_{3,t-1}. \end{aligned} \tag{3}$$

The formulation in (3), writes the error correction model with two complex cointegrating relations, $Z_{*,t} = \beta_*' X_{*,t}$ and $Z_{**,t} = \beta_{**}' X_{**,t}$, corresponding to the complex pair of roots $\pm i$. Notice that (2)) can be obtained from (3) by inserting the definitions of $\alpha_*, \beta_*, X_{*,t}$, and their complex conjugates $\alpha_{**}, \beta_{**}, X_{**,t}$, and order the terms.

Note that (2) and (3) show the isomorphi between polynomial lags and complex variables. The general results may be found in Johansen and Schaumburg (1999) and Cubbada (2001). The relation between the cointegration vector β_m and polynomial cointegration vector $\beta_m(L)$ is $\beta_m(L) = \beta_m$ for $\omega_m = 0, \pi$, and $\beta_m(L) = [Re(\beta_m) - Im(\beta_m)]\frac{\cos(\omega_m)-L}{\sin(\omega_m)}$ for $\omega_m \in (0, \pi)$.

Based on the extension of the ►canonical correlation analysis to the case of complex variables by Brillinger (1981), Cubbada applies the Johansen ML approach based on canonical correlations to obtain tests for cointegration at all the frequencies of interest, i.e., at the frequencies 0 and π with the real unit roots ±1 and at the frequency $\pi/2$ with the complex roots $\pm i$.

Hence, for each of the frequencies of interest the likelihood function is concentrated by a regression of X_{4t} and $X_{1,t-1}$, $X_{2,t-1}$ or the complex pair $(X_{*,t}, X_{**,t})$ on the other regressors, resulting in the complex residual matrices $U_{*,t}$ and $V_{*,t}$ with complex conjugates $U_{**,t}$ and $V_{**,t}$, respectively. After having purged X_{4t} and $X_{1,t-1}$, $X_{2,t-1}$ or the complex pair $(X_{*,t}, X_{**,t})$ for the effects of the other regressors, the cointegration analysis is based on a canonical correlation analysis of the relations between $U_{*,t}$ and $V_{*,t}$. The product matrices are $S_{UU} = T^{-1}\sum_{t=1}^{T} U_{*,t}U'_{**,t}$, $S_{VV} = T^{-1}\sum_{t=1}^{T} V_{*,t}V'_{**,t}$, and $S_{UV} = T^{-1}\sum_{t=1}^{T} U_{*,t}V'_{**,t}$, and the trace test of r or more cointegrating vectors is found as $TR = -2T\sum_{i=r+1}^{p}\ln(1-\widehat{\lambda}_i)$, where $\widehat{\lambda}_1 > \widehat{\lambda}_2 > .....\widehat{\lambda}_p$ are the ordered eigenvalues of the problem $\left|\lambda S_{VV} - S_{VU}S_{UU}^{-1}S_{UV}\right| = 0$ .The corresponding (possibly complex) eigenvectors properly normalized are v_j, $j = 1, 2...., p$, where the first r vectors form the cointegrating matrix β.

Critical values of the trace tests for the complex roots are supplied by Johansen and Schaumburg (1999) and Cubadda (2001), while the critical values for cointegration at the real root cases are found in Lee (1992) and Osterwald-Lenum (1992).

Furthermore, tests of linear hypotheses on the polynomial cointegration vectors may be executed as χ^2 test, similar to the test applied in the long-run cointegration case.

Although economic time series often exhibit nonstationary behavior, stationary economic variables exist as well, especially when conditioned on some deterministic pattern such as linear trends, seasonal dummies, breaks etc. However, a set of stationary economic times series may also exhibit common behavior, and for instance share a common seasonal pattern. The technique for finding such patterns, known as *Common Seasonal Features* were introduced by Engle and Hylleberg (1996) and further developed by Cubadda (1999).

About the Author

Dr. Svend Hylleberg is Professor of economics at the School of Economics and Management, The Social Science Faculty, Aarhus University. Currently he is Dean of the Social Science Faculty. He has authored or co-authored numerous papers, including some leading papers on the existing standard economic theory of seasonality as well as papers which apply newer statistical tools to the modeling of seasonal phenomena. He is a co-author (with Robert Engle, Clive Granger and B. Sam Yoo) of the seminal paper *Seasonal Integration and Cointegration* (Journal of Econometrics, 44, 215-238, 1990). Professor Hylleberg has written or edited five books including *Seasonality in Regression* (Academic Press, 1986), and *Modelling Seasonality* (Oxford University Press, 1992). He is a member of Econometric Society and Royal Economic Society. He was an Associate editor for *Econometric Review* (1994–2005) and *Scandinavian Journal of Economics* (1995–2006). Currently, he is Associate editor for *Macroeconomic Dynamics* (1997–).

Cross References

►Bayesian Approach of the Unit Root Test
►Dickey-Fuller Tests
►Econometrics
►Seasonality
►Time Series
►Trend Estimation
►Vector Autoregressive Models

References and Further Reading

Birchenhal CR, Bladen-Howell RC, Chui APL, Osborn DR, Smith JP (1989) A seasonal model of consumption. Econ J 99:837–843
Brendstrup B, Hylleberg S, Nielsen MØ, Skipper L, Stentoft L (2004) Seasonality in economic models. Macroeconomic Dyn 8(3):362–394
Brillinger DR (1981) Time series: data analysis and theory. Holden Day, San Francisco
Cubadda G (1999) Common cycles in seasonal non-stationary time series. J Appl Econ 13:273–291
Cubadda G (2001) Complex reduced rank models for seasonally cointegrated time series. Oxf Bull Econ Stat 63:497–511
Dickey DA, Hasza DP, Fuller WA (1984) Testing for unit roots in seasonal time series. J Am Stat Assoc 79:355–367
Engle RF, Granger CWJ (1987) Co-integration and error correction: representation, estimation and testing. Econometrica 55: 251–276
Engle RF, Granger CWJ, Hylleberg S, Lee H (1993) Seasonal cointegration: the Japanese consumption function. J Econom 55: 275–298
Engle RF, Hylleberg S (1996) Common seasonal features: global unemployment. Oxf Bull Econ Stat 58:615–630
Franses PH (1991) Seasonality, nonstationarity and the forecasting of monthly time series. Int J Forecast 7:199–208
Fuller WA (1976) Introduction to statistical time series. Wiley, New York

Ghysels E, Lee HS, Noh J (1994) Testing for unit roots in seasonal time series. Some theoretical extensions and a Monte Carlo investigation. J Econom 62:415–442
Ghysels E, Osborn DR (2001) The econometric analysis of seasonal time series. Cambridge University Press, Cambridge
Hylleberg S, Engle RF, Granger CWJ, Yoo BS (1990) Seasonal integration and cointegration. J Econom 44:215–238
Hylleberg S (1995) Tests for seasonal unit roots. General to Specific or Specific to General. J Econom 69:5–25
Johansen S (1995) Likelihood-based inference in cointegrated vector autoregressive models. Oxford University Press, Oxford
Johansen S, Schaumburg E (1999) Likelihood analysis of seasonal cointegration. J Econom 88:301–339
Lee HS (1992) Maximum likelihood inference on cointegration and seasonal cointegration. J Econom 54:1–47
Osterwald-Lenum M (1992) Recalculated and extended tables of the asymptotic distribution of some important maximum likelihood cointegration test statistics. Oxf Bull Econ Stat 54: 6461–6472

Seasonality

Robert M. Kunst
Professor
University of Vienna, Vienna, Austria

Introduction

Seasonality customarily refers to the annual cycle in time series sampled at intervals that are integer fractions of the annual, such as quarterly or monthly observations. The concept can easily be generalized to analogous features, such as the daily cycle in hourly observations.

The characteristics of seasonality are most easily visualized in the frequency-domain representation of the time series. Denoting the number of observations per year by S, the seasonal cycle is represented by peaks in the spectral density at $2\pi/S$ and at integer multiples of this frequency $2k\pi/S, 1 \le k \le S/2$. Seasonal cycles are distinct from other cycles by their time-constant length, though their shapes often change over time. These shapes often differ strongly from pure sine waves, and two peaks and troughs over the year are not uncommon.

The occurrence of seasonal cycles in time series has generated two related but distinct strands of literature, which can be roughly labeled as *seasonal modeling* and *seasonal adjustment*.

Seasonal modeling is concerned with typically parametric time-series models that describe the seasonal behavior of the observed variable as well as the remaining characteristics. In the spectral density interpretation, a seasonal model captures the spectral mass at the seasonal frequencies as well as the remaining characteristics of the spectral density, for example the low frequencies that represent the long run.

Seasonal adjustment builds on the concept of a decomposition of the data-generating process into a seasonal and a non-seasonal component. This decomposition can be additive ($X = X^s + X^{ns}$) or multiplicative ($X = X^s \cdot X^{ns}$). The aim of adjustment is to retrieve the non-seasonal part X^{ns} from the observed X.

Seasonal Adjustment

Seasonality is not confined to economics data. Examples for seasonal variables range from river-flow data to incidences of flu epidemics. The practice of seasonal adjustment, however, is mainly restricted to economic aggregates.

In economics, seasonal adjustment is so popular that many variables – for example, some variables of national accounts – are only available in their adjusted form, that is as an estimate of X^{ns}. It has often been pointed out that this preference tacitly assumes that X^s is generated by forces outside the economic world, such that the seasonal component of a variable does not contain useful information on the non-seasonal component of the same and of other variables. A famous citation by Svend Hylleberg (Hylleberg 1986) sees seasonal cycles as affected by cultural traditions, technological developments, and the preferences of economic agents, which can be viewed as a critique of this approach.

Currently, seasonal adjustment of economic data is mainly enacted by standardized methods, typically *X-12* in the U.S. and *TRAMO-SEATS* in Europe. The conceptual basis of X-12 is a sequence of two-sided linear filters, outlier adjustments, and the application of linear time-series models to isolate the components (see Findley et al. 1998). TRAMO-SEATS aims at isolating the components using the concepts of unobserved-components representations and of signal extraction. The assessment of the strengths and weaknesses of these procedures is difficult, as the true components are never observed.

Seasonal Modeling

The current literature on seasonal modeling builds on the SARIMA (seasonal autoregressive integrated moving-average) models by Box and Jenkins (1970), who

recommend usage of the *seasonal difference* $X_t - X_{t-S}$, followed by traditional linear modeling of the filtered series. The application of this filter assumes the existence of the factor $1-B^S$ in the generalized ARMA representation of the original series, where B denotes the lag operator. This factor has zeros at S equidistant points around the unit circle, hence the name *seasonal unit roots*. Apart from +1 and possibly −1, these unit roots come in complex pairs, such that the S roots correspond to $[S/2]+1$ frequencies or unit-root events, if $[.]$ denotes the largest integer.

The 1980s saw an increasing interest in replacing the Box-Jenkins visual analysis on differencing by statistical hypothesis tests. An offspring of the unit-root test by Dickey and Fuller is the test for seasonal unit roots by Hylleberg et al. (1990), the HEGY test. A regression is run for seasonal differences of the variable on S specific transforms. F– and t–statistics allow investigating the unit-root events separately. Under the null of seasonal unit roots will the HEGY statistics follow non-standard distributions that can be represented as Brownian motion integrals or as mixtures of normal distributions.

For example, consider quarterly data ($S = 4$). In the HEGY regression, $X_t - X_{t-4}$ is regressed on four lagged 'spectral' transforms, i.e., on $X_{t-1} + X_{t-2} + X_{t-3} + X_{t-4}$, on $-X_{t-1} + X_{t-2} - X_{t-3} + X_{t-4}$, on $X_{t-1} - X_{t-3}$ and on $X_{t-2} - X_{t-4}$. The t–statistic on the first regressor tests for the unit root at +1, the t on the second regressor for the root at −1, and an F–statistic on the latter two terms tests for the complex root pair at $\pm i$.

Testing for seasonal unit roots can be interpreted as testing whether seasonal cycles experience persistent changes over time or whether seasonal differencing is really necessary to yield a stationary variable. A process with seasonal unit roots is often called *seasonally integrated*. A variable transformed into white noise by seasonal differencing is a special seasonally integrated process and is called a *seasonal* ►*random walk*.

The HEGY test was generalized to multivariate models, to cointegration testing, and recently to panel analysis. Other tests for seasonal unit roots have been developed, some of them with unit roots as the alternative (for example, Canova and Hansen 1995). A detailed description of many of these tests and also of other issues in seasonality can be found in Ghysels and Osborn (2001).

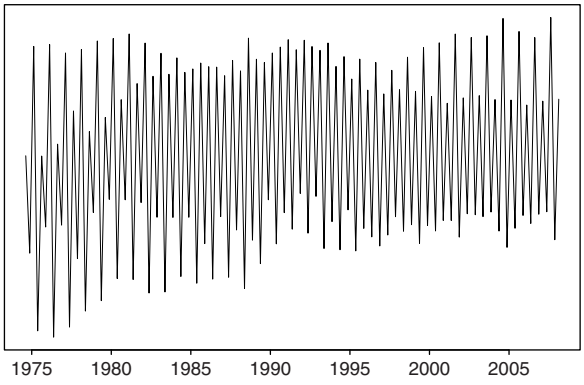
While the seasonal unit-root analysis is confined to extensions of the Box-Jenkins SARIMA class, more sophisticated seasonal models have been suggested, for example models with evolving seasonality, seasonal long memory, and seasonality in higher moments. The most intensely investigated class among them is the *periodic model* (see Franses 1996).

An Example

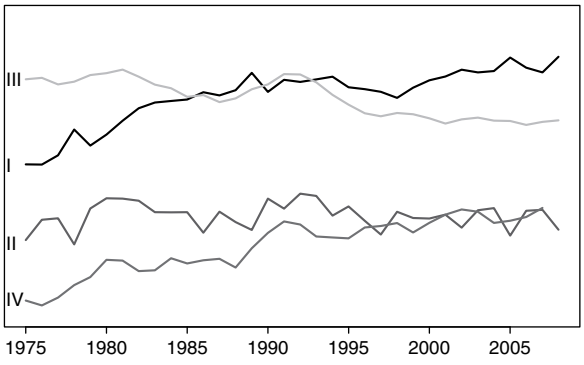
The time series variable is the quarterly number of overnight stays in the Austrian region of Tyrol for the years 1975 to 2008, which is constructed from the Austrian WIFO data base. The time-series plot in Fig. 1 shows the seasonal structure clearly.

It is a common and recommended practice to plot such series by quarters. The changes of ranks of quarters reflect the changes in the seasonal cycle. Figure 2 shows the increasing importance of winter tourism (skiing) over the observation period.

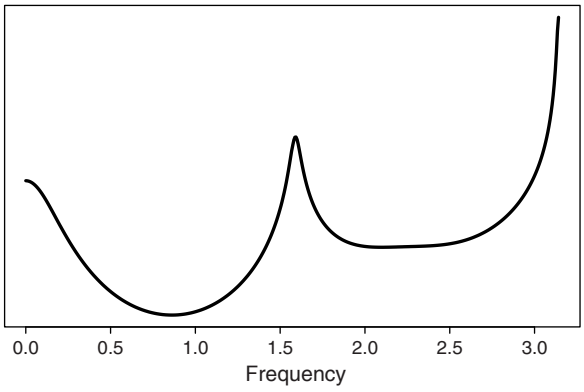
In an estimate of the spectral density (see Fig. 3), the seasonal peaks at π and $\pi/2$ are recognizable, as is another non-seasonal peak at the zero frequency (the



Seasonality. Fig. 1 Overnight stays in Tyrol, quarterly observations 1975–2008



Seasonality. Fig. 2 Overnight stays in Tyrol, plotted by quarters. Curves represent quarters I (solid), II (dashes), III (dots), and IV (dash-dotted)



Seasonality. Fig. 3 Spectral density estimate for the series on Tyrolean overnight stays

trend). Similar information is provided by the correlogram. Statistical tests confirm that this variable appears to have 'seasonal unit roots'. For example, the HEGY regression introduced above, with quarterly dummies, a trend, and a lagged $X_{t-1} - X_{t-5}$ as additional regressors, delivers t–statistics of −2.34 and −3.04, and an F–statistic of 2.56. All of these values are insignificant at 5%. The seasonal differencing operator is required to yield a stationary variable.

About the Author

Robert M. Kunst is a Professor at the University of Vienna, Austria, and a consultant and lecturer at the Institute for Advanced Studies Vienna. He is the coordinating editor of Empirical Economics. He is a Fellow of the Royal Statistical Society. He has authored or co-authored various articles on the topic of seasonality.

Cross References

►Bayesian Approach of the Unit Root Test
►Box–Jenkins Time Series Models
►Exponential and Holt-Winters Smoothing
►Moving Averages
►Seasonal Integration and Cointegration in Economic Time Series
►Time Series

References and Further Reading

Box GEP, Jenkins G (1970) Time series analysis: forecasting and control. Holden-Day, San Francisco, CA

Canova F, Hansen BE (1995) Are seasonal patterns constant over time? A test for seasonal stability. J Bus Econ Stat 13:237–252

Findley DF, Monsell BC, Bell WR, Otto MC, Chen B-C (1998) New capabilities and methods of the X-12-ARIMA seasonal adjustment program. J Bus Econ Stat 16:127–177

Franses PH (1996) Periodicity and stochastic trends in economic time series. Oxford University Press, Oxford

Ghysels E, Osborn DR (2001) The econometric analysis of seasonal time series. New York, Cambridge University Press

Hylleberg S (1986) Seasonality in regression. Academic Press, New York

Hylleberg S, Engle RF, Granger CWJ, Yoo BS (1990) Seasonal integration and cointegration. J Econom 44:215–238

Selection of Appropriate Statistical Methods in Developing Countries

Raymond Zepp
Dewey International University, Battambang, Cambodia

Statistical procedures are dictated by the nature of the research design. To the extent that comparisons of group means, searching for trends, or measuring central tendency and dispersion are universal objectives in all societies, it might be argued that the choice of statistical methods should be independent of the country or culture in question.

On the other hand, research in developing countries presents several challenges that are not as prevalent in developed countries, and therefore, the appropriateness of the statistical treatment may vary according to the type of data available.

First, data collected in developing countries can suffer from deficiencies of reliability. Industries, for example, may submit their production figures to the national statistics office in a variety of units of measurement (kilograms, tons, pounds), and these discrepancies are not always noticed by untrained workers in the statistics office.

As a result, statistics should be kept simple and transparent, so that problems of reliability can surface and be spotted easily. Research reports should include ►sensitivity analysis, that is, an analysis of how much variation in outputs could be caused by small variations in inputs.

Second, experimenters may find it more difficult in developing countries to control all variables. For example, social research may find it difficult to control the socioeconomic status of the subjects of a study. In this case, it may be more difficult to identify the real variable that gives rise to group differences. Thus, factoring out extraneous variables, for example by the ►analysis of covariance, may be a primary focus of research designs in developing countries.

Third, probability distributions may stray from the normal bell-shaped curve. Many developing countries have not only widely disparate populations, but may have two or three subpopulations such as tribal cultures or rich-poor splits that can yield bimodal distributions, or even distributions with most of the data occurring at the extremes of the curve rather than in the middle.

For this reason, there may be a tendency to use nonparametric statistical models in the analysis of data. Or, if parametric methods are to be used, careful study of the robustness of the procedure should be taken into account. If slight discrepancies from normality can result in large deviations in results, then the use of the parametric statistics should be called into question.

Fourth, technical and educational facilities in developing countries may limit the capacity to use more sophisticated statistical methods. For one thing, computer capability may be limited in either hardware or software, or else local statisticians may not be fully conversant with statistical software packages. In either case, it is probably more appropriate to adopt statistical methods that are as simple as possible.

A note needs to be made concerning statistical education in developing countries. Because schools and even universities lack the necessary computers, statistics as a subject is often taught by the old-fashioned method of calculations by hand-held calculators or even by pencil-and-paper. In such an educational system, the emphasis is often on the calculation algorithms of, say, means and standard deviations, rather on the interpretation of results. In developed countries where the entire class has unlimited access to computers with statistical software, the calculations can be done very easily, so that the emphasis can be placed on interpreting the results, or on assessing the appropriateness of the statistical method in question. In developing countries, however, students often "lose sight of the forest for the trees," that is, their academic assessment is entirely dependent on their ability to calculate algorithms that they do not focus on design of experiments and interpretation of results.

A second point about education in developing countries is the lack of teachers trained in locally appropriate methods. A university teacher quite likely has been trained in the developed world, and therefore wishes to teach students the most sophisticated and up-to-date methods, even though those methods may not be the most appropriate in the local context.

Related to the above point is the fact that the publication of research results is often biased by the complexity of the statistical methods used. A journal editor may reject a research study simply because the statistics used do not appear sophisticated enough to merit publication. Thus, a researcher may reject a simple but appropriate method in favor of a more complicated one in order to impress the readers.

One may summarize the above points in four recommendations:

1. When in doubt, opt for the simpler statistical procedure.
2. Be prepared to use nonparametric statistics.
3. Sensitivity Analysis should be carried out to compensate for possibilities of unreliable data.
4. Students should be trained in the appropriateness of statistical design and interpretation of results, not just in the calculation of statistical algorithms.

About the Author

Raymond Zepp holds a Bachelor's Degree in Mathematics from Oberlin College, a Master's Degree in Mathematics from the University of Cincinnati, and a Ph.D. in Mathematics Education from the Ohio State University. He is Vice President of the newly-opened Dewey International University in Battambang, Cambodia. As founder of DIU (www.diucambodia.org), he has incorporated his vision of "Learning by Doing" into a strong emphasis on community service learning and research. Dr. Zepp has taught statistics in developing universities, governments, and non-governmental organizations around the world, for example, in Nigeria, Lesotho, Macau, Papua New Guinea, Micronesia, Mozambique, Uganda, Qatar, and Cyprus, and as Fulbright Professor in the Ivory Coast. He has set up research institutes at the University of Cambodia and at Qatar University, and has designed new universities in Nigeria (Maiduguri) and Papua New Guinea (Goroka), and of course Cambodia (Dewey International). He has done statistical consulting for USAID, UNDP, Asia Development Bank, the World Bank, and others. Dr. Zepp has authored or co authored over 40 books (e.g., *Business Research and Statistics*, Hong Kong: Asia Pacific International Press, 1988) and over 100 journal articles, conference papers, etc. He currently resides in Battambang, Cambodia.

Cross References

►African Population Censuses
►Decision Trees for the Teaching of Statistical Estimation
►Learning Statistics in a Foreign Language
►Nonparametric Statistical Inference
►Promoting, Fostering and Development of Statistics in Developing Countries
►Role of Statistics: Developing Country Perspective
►Sensitivity Analysis

Semiparametric Regression Models

Yingcun Xia
Professor
National University of Singapore, Singapore, Singapore

In statistics, semiparametric regression includes regression models that combine parametric and nonparametric models. They are often used in situations where the fully nonparametric model may not perform well or when the researcher wants to use a parametric model but the functional form with respect to a subset of the regressors or the density of the errors is not known. Suppose Y is a response and $X = (\mathbf{x}_1, ..., \mathbf{x}_p)$ are covariates. A basic goal is to estimate $m(x) = E(Y|X = x)$ or the model $Y = m(X) + \varepsilon$ with $E(\varepsilon|X) = 0$ almost surely. Without any information about the structure of the function, it is difficult to estimate $m(x)$ well when $p > 1$, and as a consequence many parametric and semiparametric models have been proposed that impose structural constraints or special functional forms upon $m(x)$. Popular semiparametric models include *partially linear models*, see for example Speckman (1988), in which

$$Y = \beta_1 \mathbf{x}_1 + ... + \beta_{p-1}\mathbf{x}_{p-1} + g_p(\mathbf{x}_p) + \varepsilon,$$

additive models, see for example Hastie and Tibshirani (1990), in which

$$Y = g_1(\mathbf{x}_1) + g_2(\mathbf{x}_2) + ... + g_p(\mathbf{x}_p) + \varepsilon,$$

single-index models, see for example Ichimura (1993), in which

$$Y = g(\beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p) + \varepsilon,$$

varying coefficient models, see for example Chen and Tsay (1993) and Hastie and Tibshirani (1993), in which

$$Y = g_1(\mathbf{x}_1) + g_2(\mathbf{x}_1)\mathbf{x}_2 + ... + g_p(\mathbf{x}_1)\mathbf{x}_p + \varepsilon.$$

and *extended partially linear single-index model*, see Xia et al. (1999), in which

$$Y = \beta_1 \mathbf{x}_1 + ... + \beta_p \mathbf{x}_p + g(\theta_1 \mathbf{x}_1 + ... + \theta_p \mathbf{x}_p) + \varepsilon.$$

In all the above models, $g_1, ..., g_p$ and g are unknown functions and $\beta_1, ..., \beta_p, \theta_1, ..., \theta_p$ are parameters need to be estimated. A general form of the semiparametric model including all the models above is

$$\mu\{E(Y|\mathbf{x}_1, ..., \mathbf{x}_p)\} = G(g, \beta, X),$$

where $g = (g_1, ..., g_q)^T$ are unknown smooth functions, G is known up to a parameter vector β, function μ is known and usually monotonic.

Both splines smoothing and Kernel smoothing can be used to estimate these models. The general model can be estimated by the method proposed by Xia et al. (2002). Theoretically, all these models can avoid the "curse of dimensionality" in the estimation. The estimators of the unknown functions $g_1, ..., g_p$ and g can achieve the optimal consistency rate of univariate function, and the parameters such as $\beta_1, ..., \beta_p$ and θ are root-n consistent.

These models have been found very useful in application; see for example Hastie and Tibshirani (1990), Fan and Gijbels (1996) and Ruppert et al. (2003).

About the Author

Yingcun Xia is Professor of statistics at the National University of Singapore. He was elected member of the International Statistics Institute (2005–). He was Associated Editor for the *Annals of Statistics* (2007–2009). His research interest includes semiparametric modeling, nonlinear time series analysis and statistical modeling of infectious diseases. His work on nonlinear dimension reduction (called MAVE) and on the modeling of transmission of infectious diseases based on gravity mechanism has received wide recognition.

Cross References

►Absolute Penalty Estimation
►Bayesian Semiparametric Regression
►Nonparametric Regression Using Kernel and Spline Methods
►Smoothing Splines

References and Further Reading

Chen R, Tsay R (1993) Functional coefficient autoregressive models: estimation and tests of hypotheses. J Am Stat Assoc 88:298–308
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman and Hall, London
Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman and Hall/CRC, Boca Rotan, FL
Ichimura H (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. J Econometrics 58:71–120
Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression Cambridge University Press, UK
Speckman P (1988) Kernel smoothing in partial linear models. J Roy Stat Soc Ser B 50:413–436
Xia Y, Tong H, Li WK (1999) On extended partially linear single-index models. Biometrika 86:831–842
Xia Y, Tong H, Li WK, Zhu L (2002) An adaptive estimation of dimension reduction space (with discussion). J Roy Stat Soc Ser B 64:363–410

Semi-Variance in Finance

Vijay K. Rohatgi
Professor Emeritus
Bowling Green State University, Bowling Green, OH, USA

For any random variable X with finite variance, and any constant t

$$E\{(X-t)\}^2 = E\{(X-t)^-\}^2 + E\{(X-t)^+\}^2.$$

If $t = \mu = EX$, then $E\{(X-t)\}^2 = \sigma^2$, the variance of X. The quantity $E\{(X-\mu)^-\}^2$ is called the (lower) semi-variance of X whereas $E\{(X-\mu)^+\}^2$ is called the upper semi-variance of X. In financial applications where X represents return on an investment, σ is widely used as a measure of risk of an investment (portfolio). In that context σ is called volatility since it measures volatility of returns. Risk-averse investors like consistency of returns and hence lower volatility. In order to compare two or more investments one compares their returns per unit of risk, that is, μ/σ = 1/coefficient of variation. A modified version of this measure is due to Sharpe (1994) who uses the ratio excess returns (over risk free returns) divided by volatility. Another widely used measure of investors' risk is beta, the coefficient of linear regression of returns over some benchmark returns such as Standard and Poor 500 index. Thus, a value of beta over 1 means that the investment under consideration is more volatile (risky) than the benchmark.

For risk-averse investors neither of these two measures fits their need. They are more interested in the downside risk, the risk of losing money or falling below the target return. For instance, variance assigns equal weight to both deviations, those above the mean and those below the mean. In that sense it is more suitable for symmetric return distributions in which case $\sigma^2 = 2E\{(X-\mu)^-\}^2$. In practice the return distributions are often skewed to the right. No investor is averse to returns in excess of the target. He or she prefers positive skewness because the chance of large deviations from the target rate is much less.

Markowitz (1959) introduced

$$\sigma_D^2(t) = E\{(X-t)^-\}^2$$

as a measure of downside risk. Here t may be called the target rate of return which could be the riskless rate such as the three month T-bill rate or the Libor rate. Recall that $E\{(X-t)\}^2$ is minimized for $t = \mu$. On the other hand $\sigma_D^2(t)$ is an increasing function of t and a Chebyshev type inequality holds:

$$P(X < \mu - k\sigma_D(t)) \le 1/k^2 \quad \text{for } k \ge 1.$$

As an estimate of $\sigma_D^2(t)$ one generally uses the substitution principle estimator

$$(1/n)\sum_{i=1}^{n}\{(x_i-t)^-\}^2$$

and when $t = \mu$ we use the estimator

$$(1/n)\sum_{1}^{n}\{(x_i-\bar{x})^-\}^2.$$

Markowitz (1952) was the first to propose a method of construction of portfolios based on mean returns, and their variances and covariances (*see* ►Portfolio theory). In 1959 he proposed semivariance as a measure of downside risk and advocated its use in portfolio selection. Due to computational complexity of semivariance and semicovariance, however, he used the variance measure of risk instead. After the advent of desktop computers and their computational power in 1980s the focus shifted to portfolio selection based on semivariance as a measure of downside risk. See for example Markowitz et al. (1993).

Both $\sigma_D(t)$ and $\sigma_U(t)$ $\left(\sigma_U^2(t) = E\{(X-t)^+\}^2\right)$ have been used in Quality Control (*see* ►Statistical Quality Control) in constructing process capability indices. See for example, Kotz and Cynthia (1998). Other uses are in spatial statistics and in construction of confidence intervals in simulation output analysis Coobineh and Branting (1991). The semi-standard deviation $\sigma_D(\mu)$ can also be used in setting up dynamic stop loss points in security trading.

About the Author

Vijay K. Rohatgi is a Professor Emeritus, Department of Mathematics and Statistics, Bowling Green State Univerity (BGSU), Ohio. He was Chair of the Department (1983–1985) and played a key role in the creation of the doctoral program of the Department. He is internationally well-known as an author/co-author of five successful books on statistics, probability and the related, including *An Introduction to Probability and Statistics* (with A.K.Md. Ehsanes Saleh, Wiley, 2nd edition, 2001) and *Statistical Inference* (Dover Publications, 2003).

Cross References

►Banking, Statistics in
►Coefficient of Variation

►Portfolio Theory
►Variance

References and Further Reading

Coobineh F, Branting D (1991) A split distribution method for constructing confidence intervals for simulation output analysis. Int J Sys Sci 22:367–374
Kotz S, Cynthia L (1998) Process capability indices in theory and practice. Arnold, New York
Markowitz HM (1952) Portfolio selection. J Finance 7:77–91
Markowitz HM (1959) Portfolio selection, efficient diversification of investments. Cowles Foundation Monograph 16. Yale University Press, New Haven
Markowitz H, Todd P, Xu G (1993) Computation of mean-semivariance efficient set by the critical line algorithm. Ann Oper Res 45:307–317
Sharpe WF (1994) The sharpe ratio. J Portfolio Manag 21:49–58

Sensitivity Analysis

Andrea Saltelli[1], Paola Annoni[2]
[1]Head of the Unit of Econometrics and Applied Statistics Joint Research Centre of the European Commission, Institute for the Protection and the Security of the Citizen, Ispra, Italy
[2]Joint Research Centre of the European Commission, Institute for the Protection and the Security of the Citizen, Ispra, Italy

Existing guidelines for impact assessment recommend that mathematical modeling of real or man-made system be accompanied by a 'sensitivity analysis' - SA (EC 2009; EPA 2009; OMB 2006). The same recommendation can be found in textbooks for practitioners (e.g., Kennedy 2007, Saltelli et al. 2008). Mathematical models can be seen as machines capable of mapping from a set of assumptions (data, parameters, scenarios) into an inference (model output).

In this respect modelers should tackle:

- **Uncertainty.** Characterize the empirical probability density function and the confidence bounds for a model output. This can be viewed as the numerical equivalent of the measurement error for physical experiments. The question answered is "How uncertain is this inference?"
- **Sensitivity.** Identify factors or groups of factors mostly responsible for the uncertainty in the prediction. The question answered is "Where is this uncertainty coming from?"

The two terms are often used differently, with sensitivity analysis used for both challenges (e.g., Leamer 1990). We focus on sensitivity analysis proper, i.e., the effect of individual factors or group of factors in driving the output and its uncertainty.

Basic Concepts

The ingredients of a sensitivity analysis are the model's uncertain input factors and model's outputs. Here and in the following we shall interpret as factor all that can be plausibly changed at the level of model formulation or model input in the quest to map the space of the model predictions. Thus a factor could be an input datum acquired with a known uncertainty, as well as a parameter estimated with known uncertainty in a previous stage of modeling, as well a trigger acting on the model's structure (e.g., a mesh size choice), or a trigger selecting the choice of a model versus another, or the selection of a scenario. Modelers usually have considerable latitude of choice as to how to combine factors in a sensitivity analysis, e.g., what to vary, what to keep fixed. Also a modeler's choice is, to some extent, whether to treat factors as dependent upon one another or as independent. The design and the interpretation of this ensemble of the model simulations constitute a sensitivity analysis.

Use of Sensitivity Analysis

Sensitivity analysis is a tool to test the quality of a model or better the quality of an inference based on a model. This is investigated by looking at the robustness of an inference. There is a trade off here between how scrupulous an analyst is in exploring the input assumptions and how wide the resulting inference will be. Edward E. Leamer (1990) calls this an organized sensitivity analysis:

> *I have proposed a form of organized sensitivity analysis that I call 'global sensitivity analysis' in which a neighborhood of alternative assumptions is selected and the corresponding interval of inferences is identified. Conclusions are judged to be sturdy only if the neighborhood of assumptions is wide enough to be credible and the corresponding interval of inferences is narrow enough to be useful.*

In fact it is easy to invalidate a model demonstrating that it is fragile with respect to the uncertainty in the assumptions. Likewise one can criticize a sensitivity analysis by showing that its assumptions have not been taken 'wide enough.'

Examples of application of SA are: robustness assessment in the context of impact assessment; model simplification in the context of complex and computer demanding models; quality assurance for detecting coding errors or

misspecifications. Sensitivity analysis can also highlight the region in the space of input factors for which the model output assumes extreme values, as can be relevant in ▶risk analysis. Likewise it can identify model instability regions within the space of the factors for use in a subsequent calibration study.

Local Vs Global Methods

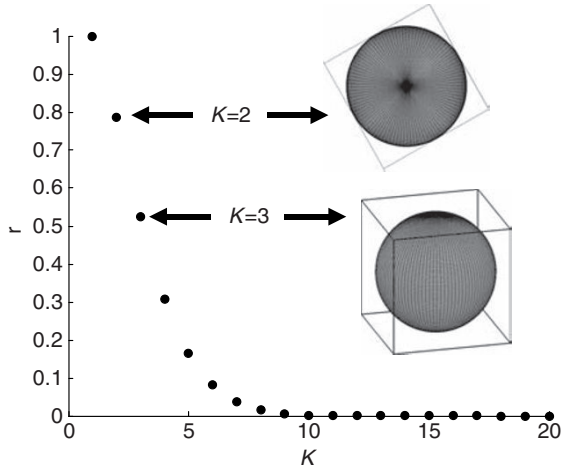
In the model $Y = f(X_1, X_2, \ldots, X_k)$ Y is the output and X_is are the input factors. The model is linear if each factor X_i enters linearly in f. The model is additive if the function f may be decomposed into a sum of k functions $f_i \equiv f_i(X_i)$, each f_i depending only on its own factor X_i.

There are 'local' and 'global' methods for SA. If the model is *linear*, a *local approach* based on first derivatives of the output with respect to the input factors will provide all the information that is needed for SA. If the model is *non linear but additive*, i.e., there are no interactions among factors, then *derivatives of higher and cross order* will be needed. When a-priori information on the nature of the model is not available (*model-free setting*) or the model is acknowledged to be non additive, then *global methods* are needed whereby all the space of the uncertain input factors is explored. Note that often modelers cannot assume linearity and additivity as their models come in the form of computer programs, possibly including several computational steps. In this situation it is better to use 'global' methods (EPA 2009; Saltelli et al. 2008).

A Very Popular Practice: OAT-SA

Most sensitivity analysis met in the literature are realized by varying one factor at a time – OAT approaches. Modelers have many good reasons to adopt OAT, including the use of a common '*baseline*' value from which all factors are moved. Derivative based approaches - when the derivatives stop at the first order - are a particular case of OAT. Typical arguments in favor of OAT are: (1) The baseline vector is a safe starting point where the model properties are well known; (2) Whatever effect is detected on the output, this is solely due to that factor which was moved and to none other; (3) The chances of the model to crash or to give unacceptable results are minimized as these generally increase with the distance from the baseline.

Despite all these points in favor to an OAT sensitivity analysis we would like to discourage as much as possible this practice (Saltelli and Annoni 2010). OAT is inefficient in exploring the input space as the coverage of the design space is extremely poor already with few input factors. The issue of uniformly covering the hyperspace in high dimensions is a well known and widely discussed matter under the name *curse of dimensionality* (Hastie et al. 2001). There



Sensitivity Analysis. Fig. 1 Curse of dimensionality-horizontal axis = number of dimensions; vertical axis = volume of the inscribed unitary sphere

are various ways to visualize this 'curse'. Figure 1 may be effective. It shows that, as the number of dimensions k increases, the volume of the hyper-sphere inscribed in the unitary hyper-cube goes rapidly to zero (it is less than 1% already for $k = 10$).

The OAT approach – moving always one step away from the same baseline – can be shown to generate points inside the hyper-sphere. Of course when one throws a handful of points in a multidimensional space these points will be sparse, and in no way the space will be fully explored. Still, even if one has only a handful of points at disposal, there is no reason why one should concentrate all these points in the hyper-sphere, i.e., closer to the origin on average than randomly generated points in the cube.

An additional shortcoming of OAT is that it cannot detect factor interactions. It may be the case that a factor is detected as no influential while it is actually relevant but only through its interaction with the other factors. In a model free setting, OAT is by no means the winning choice.

Design and Estimators

Unlike OAT, a good experimental design will tend to change more factors simultaneously. This design can be realized using the same techniques used for experimental design (e.g., a saturated two-level design or an unsaturated design with more levels). A practical alternative for numerical experiments is a Monte Carlo method. Beside design, sensitivity analysis needs sensitivity estimators which will translate the function values computed at the design points into sensitivity coefficients for the various factors.

Model's predictions have to be evaluated at different points within the parameter space, whose dimensionality is equal to the number k of input factors. To explore the k-dimensional factor space (the hyperspace) the first step in usually to reduce the problem to traveling across the k-dimensional unit cube by using the inverse cumulative distribution function of input factors. The input space can be explored using ad hoc trajectories (such as in the elementary effects method below), random numbers or quasi-random numbers. Quasi-random numbers are specifically designed to generate samples from the space of input factors as uniformly as possible. For a review on quasi random sequences and their properties see Bratley and Fox (1988).

After sampling the space of input factors, various methods may be applied to compute different sensitivity measures. Selected practices are given next.

Morris' Elementary Effects

The Elementary Effect method (Morris 1991) provides a ranking of input factors according to a sensitivity measure simply based on averages of derivatives over the space of factors. In the Morris setting each input factor is discretized into p levels and the exploration of the input space is carried out along r trajectories of $(k+1)$ points, where each point differs from the previous one in only one component. Each trajectory provides rough sensitivity measures for each factor called elementary effect *EE*. The elementary effect of trajectory j for factor i is:

$$EE_i^{(j)} = \frac{Y(X_1,\ldots,X_{i-1},X_i+\Delta,X_{i+1},\ldots,X_k) - Y(X_1,\ldots,X_k)}{\Delta} \tag{1}$$

where convenient choices for p and Δ are p even and Δ equal to $p/[2(p-1)]$. The point $(X_1,\ldots,X_k)$ is any point in the input space such that the incremental point $(X_1,\ldots,X_{i-1},X_i+\Delta,X_{i+1},\ldots,X_k)$ still belongs to the input space (for each $i=1,\ldots,k$). Elementary effect $EE_i^{(j)}$ provides a sensitivity index which highly depends on the particular trajectory, being in this sense *local*. To compute a more *global* sensitivity measure, many trajectories are chosen and the average value of $EE_i^{(j)}$ over j is computed. Following a recent revision of original Morris' measure, factors may be ranked according to μ^* (Campolongo et al. 2007):

$$\mu_i^* = \frac{1}{r}\sum_{j=1}^{r}\left|EE_i^{(j)}\right| \tag{2}$$

The elementary effects sensitivity measure is an efficient alternative to OAT. It is used for factor screening, especially with large and complex models. When modellers are constrained by computational costs, a recommended practice is to perform a preliminary analysis by means of Morris' trajectories to detect possible non influential factors. More computationally intensive methods may be then applied to a smaller set of input factors.

Monte Carlo Filtering

An alternative setting for sensitivity analysis is the 'factor mapping' which relates to situations when there is a special concern towards a particular portion of the distribution of the output Y, e.g., one is concerned with Y above or below a given threshold – e.g., an investment loss or a toxicity level not to be exceeded. This is the typical setting of Monte Carlo Filtering MCF (see Saltelli et al. 2004 for a review). The realizations of Y are classified into 'good' – behavioral – and 'bad - non-behavioral depending on the value of Y with respect to the threshold. A MCF analysis is divided into the following steps:

1. Compute different realizations of Y corresponding to different sampled points in the space of input factor by means of a Monte Carlo experiment;
2. Classify each realization as either behavioral (B) or non behavioral $(\overline{B})$;
3. For each X_i define two subsets, one including all the values of X_i which give behavioral Y, denoted $(X_i|B)$, the other including all the remaining values $(X_i|\overline{B})$;
4. Compute the statistical difference between the two empirical distribution functions of $(X_i|B)$ and $(X_i|\overline{B})$. A factor is considered influential if the two distribution functions are statistically different. Classical statistical tests, such as Smirnov two-sample test may be used to the purpose.

Variance-Based Sensitivity Measures

With variance-based sensitivity analysis (VB-SA) input factors can be ranked according to their contribution to the output variance. VB-SA also tackles interaction effects instructing the analyst about cooperative behavior of factors. Interactions can lead to extremal values of model output and are thus relevant to the analysis. In VB-SA sensitivity analysis the two most relevant measures are 'first order' and 'total order' indices.

The best systematization of the theory of variance-based methods is due to Sobol' (Sobol 1990), while total sensitivity indices were introduced by Homma and Saltelli (1996). For reviews see also Saltelli et al. (2005) or Helton et al. (2006). Variance-based SA uses measures as

$$S_i = \frac{V_{X_i}\left(E_{\mathbf{X}_{\sim i}}(Y|X_i)\right)}{V(Y)} \tag{3}$$

and

$$S_{T_i} = \frac{E_{\mathbf{X}_{\sim i}}\left(V_{X_i}(Y|\mathbf{X}_{\sim i})\right)}{V(Y)} = 1 - \frac{V_{\mathbf{X}_{\sim i}}\left(E_{X_i}(Y|\mathbf{X}_{\sim i})\right)}{V(Y)} \quad (4)$$

where $\mathbf{X}_{\sim i} = \{X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k\}$.

$E_{\mathbf{X}_{\sim i}}(Y|X_i)$ is the value of Y obtained by averaging over all factors but X_i, and is thus a function of X_i alone. $V_{X_i}\left(E_{\mathbf{X}_{\sim i}}(Y|X_i)\right)$ is the variance of this function over X_i itself. Intuitively a high value of this statistics implies an influent factor.

The quantity S_i corresponds to the fraction of $V(Y)$ that can be attributed to X_i alone. It can be viewed as a measure of how well $E_{\mathbf{X}_{\sim i}}(Y|X_i)$ fits Y: if the fitting is optimal then $S_i \cong 1$ and factor X_i is highly relevant. The quantity S_{T_i} corresponds to the fraction of $V(Y)$ that can be attributed to X_i and all its interactions with other factors. For additive models the two measures S_i and S_{T_i} are equal to one another for each factor X_i. For an interacting factor the difference $S_{T_i} - S_i$ is a measure of the strength of the interactions.

The estimation of S_i and S_{T_i} requires the computation of k-dimensional integrals. They are generally approximated assuming independency among input factors and using Monte-Carlo or quasi-Monte-Carlo sampling from the joint distribution of the space of input factors. Alternative procedures for the computation of S_i and S_{T_i} are available which use direct calculations. They all derive from metamodels, which provide cheap emulators of complex and large computational models (see for example Oakley and O'Hagan 2004; Storlie et al. 2009).

About the Author

Andrea Saltelli, has worked on physical chemistry, environmental sciences and applied statistics. His main disciplinary foci are sensitivity analysis and composite indicators. He is the leading author of three successful volumes published by Wiley on sensitivity analysis: *Sensitivity Analysis: Gauging the Worth of Scientific Models* (2000), *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models* (2004), and *Global Sensitivity Analysis: The Primer* (2008), and of several papers on the same subject. Paola Annoni has produced original work in the field of sensitivity analysis and partial ordering. Both work at the Joint Research Centre of the European Commission in Ispra, Italy.

Cross References

►Bayesian Statistics
►Bias Analysis
►Composite Indicators
►Design of Experiments: A Pattern of Progress
►Interaction
►Misuse of Statistics
►Model Selection
►Monte Carlo Methods in Statistics
►Selection of Appropriate Statistical Methods in Developing Countries

References and Further Reading

Bratley P, Fox BL (1988) Algorithm 659 implementing Sobol's quasi-random sequence generator. ACM Trans Math Soft 14(1):88–100

Campolongo F, Cariboni J, Saltelli A (2007) An effective screening design for sensitivity analysis of large models. Environ Model Soft 22:1509–1518

EC 2009 Impact assessment guidelines. SEC, p 24. http://ec.europa.eu/governance/impact/docs/key_docs/iag_2009_en.pdf. Accessed 15 Jan 2009

EPA 2009 Guidance on the development, evaluation, and application of environmental models. Technical Report, Office of the science advisor, Council for Regulatory Environmental Modeling. EPA /100/K-09/003, p 26. http://www.epa.gov/crem/library/cred_guidance_0309.pdf

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, New York

Helton JC, Johnson JD, Salaberry CJ, Storlie CB (2006) Survey of sampling based methods for uncertainty and sensitivity analysis. Reliab Eng Syst Saf 91:1175–1209

Homma T, Saltelli A (1996) Importance measures in global sensitivity analysis of model output. Reliab Eng Syst Saf 52(1):1–17

Kennedy P (2007) A guide to econometrics, 5th edn. Blackwell Publishing, Oxford

Leamer E (1990) Let's take the con out of econometrics, and sensitivity analysis would help. In: Granger C (ed) Modelling economic series. Clarendon Press, Oxford

Morris MD (1991) Fractional sampling plan for preliminary computational experiments. Technometrics 33:161–174

Oakley JE, O'Hagan A (2004) Probabilistic sensitivity analysis of complex models: a Bayesian approach. J Roy Stat Soc B 66: 751–769

OMB – Office of Management and Budget (2006) Proposed risk assessment bulletin (http://www.whitehouse.gov/omb/inforeg/proposed_risk_assessment_bulletin_010906.pdf)

Saltelli A, Annoni P (2010) How to avoid a perfunctory sensitivity analysis, Environ Model Softw, doi:10.1016/j.envsoft.2010.04.012

Saltelli A, Tarantola S, Campolongo F, Ratto M (2004) Sensitivity analysis in practice. A guide to assessing scientific models. Wiley, Chichester

Saltelli A, Ratto M, Tarantola S, Campolongo F (2005) Sensitivity analysis for chemical models. Chem Rev 105(7):2811–2828

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2008) Global sensitivity analysis. The primer. Wiley, Chichester

Sobol' IM (1990) Sensitivity estimates for nonlinear mathematical models. Matem Mod 2:112–118. (in Russian). (trans: Sobol' IM (1993) Sensitivity analysis for non-linear mathematical models. Math Model Comp Exper 1:407–414)

Storlie CB, Swiler LP, Helton JC, Sallaberry CJ (2009) Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. Reliab Eng Syst Saf 94:1735–1763

Sensometrics

Per Bruun Brockhoff
Professor, Head of Statistics Section
Technical University of Denmark, Lyngby, Denmark

Introduction

The use of humans as measurement instruments is playing an increasing role in product development and user-driven innovation in many industries. This ranges from the use of experts and trained human test panels to market studies where the consumer population is tested for preference and behavior patterns. This calls for improved understanding on one side of the human measurement instrument itself and on the other side the modeling and empirical treatment of data. The scientific grounds for obtaining improvements within a given industry span from experimental psychology to mathematical modeling, statistics, chemometrics, and machine learning together with specific product knowledge be it food, TVs, hearing aids, mobile phones, or whatever.

In particular in the food industry, sensory and consumer data is frequently produced and applied as the basis for decision making. And in the field of food research, sensory and consumer data is produced and used similar to the industrial use, and academic environments specifically for sensory and consumer sciences exist worldwide. The development and application of statistics and data analysis within this area is called sensometrics.

Sensory Science and Sensometrics

As the name indicates, sensometrics really grew out of and is still closely linked to sensory science, where the use of trained sensory panels plays a central role. Sensory science is the cross-disciplinary scientific field dealing with human perception of stimuli and the way they act upon sensory input. Sensory food research focuses on better understanding of how the senses react during food intake, but also how our senses can be used in quality control and innovative product development. Historically it can be viewed as a merger of simple industrial product testing with psychophysics as originated by G.T. Fechner and S.S. Stevens in the nineteenth century. Probably the first exposition of the modern sensory science is given by Amerine et al. (1965). Rose Marie Pangborn (1932–1990) was considered one of the pioneers of sensory analysis of food and the main global scientific conference in sensory science is named after her. The first Pangborn Symposium was held in Helsinki, Finland, in 1992 and these conferences are approaching in the order of 1,000 participants – the ninth was planned for in Bangkok, Thailand, in 2011. Jointly with this, international sensometrics conferences have been held also since 1992, where the first took place in Leiden, Holland (as a small workshop), and the tenth took place in Rotterdam, Holland, in 2010. The sensometrics conferences have a participation level of around 120–150. Both conferences are working together with the Elsevier Journal *Food Quality and Preference*, which is also the official membership journal for the Sensometrics Society (www.sensometric.org).

Sensometrics: Statistics, Psychometrics, or Chemometrics?

The "sensometrician" is faced with a vast collection of data types from a large number of experimental settings ranging from a simple one-sample binomial outcome to complex dynamical and/or multivariate data sets; see, e.g., Bredie et al. (2010) for a recent review of quantitative sensory methodology. So what is really (good) sensometrics? The answer will depend on the background of the sensometrician, who for the majority, if not a food scientist, is coming from one of the following fields: generic statistics, psychophysics/experimental psychology, or chemometrics.

The generic statistician arch type would commonly carry out the data analysis as a purely "empirical" exercise in the sense that methods are not based on any models for the fundamental psychological characteristics underlying the sensory phenomena that the measurements express. The advantage of a strong link to the generic scientific fields of mathematical and applied statistics is the ability to employ the most modern statistical techniques when relevant for sensory data and to be on top of sampling uncertainty and formal statistical inferential reasoning. And this is certainly needed for the sensory field as for any other field producing experimental data. The weakness is that the lack of proper psychophysical models may lead to inadequate interpretations of the analysis results. In, e.g., MacKay (2005) the first sentence of the abstract is expressing this concern rather severely: "Sensory and hedonic variability are fundamental psychological characteristics that must be explicitly modeled if one is to develop meaningful statistical models of sensory phenomena." A fundamental challenge of this ambitious approach is that the required psychophysical (probabilistic) models of behavior are on one hand only vaguely verifiable, since they are based on models of a (partly) unobserved system, the human brain and perceptual system, and on the other hand may lead to rather complicated statistical models. MacKay (2005) is published in a special sensory data issue of *The Journal of Chemometrics*; see Brockhoff et al. (2005). Chemometricians are the third and final arch type

of a sensometrician. In chemometrics the focus is more on multivariate data analysis (see ►Multivariate Data Analysis: An Overview) and for some the explorative principle is at the very heart of the field; see, e.g., Munck (2007) and Martens and Martens (2001). The advantage of the chemometrics approach is that usually all multivariate features of the data are studied without forcing certain potentially inadequate model structures on the data. The weakness is exactly also this lack of modeling rendering potentially certain well-understood psychophysical phenomena for the explorative modeling to find out by itself. Also, linked with the explorative approach, the formal statistical inferential reasoning is sometimes considered less important by the chemometrician.

Now, none of these arch types are (at their best) unintelligent and they would, all three of them, understand (some of) the limitations of their pure versions of analysis approach. And they all have ways of dealing with (some of) these concerns for practical data analysis, such that often, at the end of the day, the end results may not differ that much. There is though, in the point of view of this author, a lack of comprehensive comparisons between these different approaches where they all are used at their best.

Example 1: Sensory Profile Data

As an example, consider the so-called descriptive sensory analysis, also called sensory profiling. In sensory profiling the panelists develop a test vocabulary (defining attributes) for the product category and rate the intensity of these attributes for a set of different samples within the category. Thus, a sensory profile of each product is provided for each of the panelists, and most often this is replicated; see Lawless and Heymann (1999). Hence, data is inherently multivariate as many characteristics of the products are measured.

The statistics arch type would focus on the ANOVA structure of the setting and perform univariate and multivariate analyses of variance (ANOVA) and would make sure that the proper version of a mixed model ANOVA is used; see, e.g., Lea et al. (1997) and Næs et al. (2010). For studying the multivariate product structure the Canonical Variates Analysis (CVA) within the Multivariate ANOVA (MANOVA) framework would be the natural choice (see, e.g., Schlich (1998)) since it would be an analysis that incorporates the within-product (co)variability.

The chemometrics arch type would begin with principal components analysis (PCA) on averaged and/or unfolded data. For more elaborate analysis maybe three-way methods (see Brockhoff et al. (1996), Bro et al. (2002)) or other more ANOVA-like extensions would be used (see, e.g., Luciano and Næs (2008)). Analysis accounting for within-product (co)variability could be provided by extensions as presented in Bro et al. (2002) or in Martens et al. (2003).

In MacKay (2005) the approach for this type of data is that of probabilistic multidimensional scaling (PROSCAL). In short, a formal statistical model for product differences is expressed as variability on the (low-dimensional) underlying latent sensory scale. It is usually presented as superior to the use of, e.g., standard PCA, focusing on the point that it naturally includes models for different within-product variability, which in the standard PCA could be confounded with the "signal" – the inter-product distances.

Example 2: Sensory Difference and Similarity Test Data

The so-called difference and/or similarity tests are a commonly used sensory technique resulting in binary and/or categorical frequency data – the so-called triangle test is a classical example. In the triangle test an individual is presented with three samples, two of which are the same, and then asked to select the odd sample. The result is binary: correct or incorrect. Such sensory tests were already in the 1950s treated by the statistical community; see, e.g., Hopkins (1950) and Bradley (1958). These types of tests and results have also been treated extensively from a more psychophysical approach, often here denoted a Thurstonian approach. The focus in the Thurstonian approach is on quantifying/estimating the underlying sensory difference d between the two products that are compared in the difference test. This is done by setting up mathematical/psychophysical models for the cognitive decision processes that are used by assessors in each sensory test protocol see; e.g., Ennis (1993). For the triangle test, the usual model for how the cognitive decision process is taking place is that the most deviating product would be the answer – sometimes called that the assessors are using a so-called tau-strategy. Using basic probability calculus on three realizations from two different normal distributions, differing by exactly the true underlying sensory difference d, one can deduce the probability of getting the answer correct for such a strategy. This function is called the psychometric function and relates the observed number of correct answers to the underlying sensory difference d. Different test protocols will then lead to different psychometric functions. InBock and Jones (1968) probably the first systematic exposition of the psychological scaling theory and methods by Thurstone was given. This included a sound psychological basis as well as a statistical one with the use and theory of maximum likelihood methods. Within the field known as signal detection theory (see, e.g., Green and

Swets (1966) or Macmillan and Creelman (2005)), methods of this kind were further developed, originally with special emphasis on detecting weak visual or auditory signals. Further developments of such methods and their use within food testing and sensory science have developed over the last couple of decades with the numerous contributions of D. Ennis as a corner stone; see, e.g., Ennis (2003). In Brockhoff and Christensen (2010) it was emphasized and exploited that the Thurstonian-based statistical analysis of data from the basic sensory discrimination test protocols can be identified as ►generalized linear models using the inverse psychometric functions as link functions. With this in place, it is possible to extend and combine designed experimentation with discrimination/similarity testing and combine standard statistical modeling/analysis with Thurstonian modeling.

Summary

One recurrent issue in sensometrics is the monitoring and/or accounting for individual differences in sensory panel data, also called dealing with panel performance. A model-based approach within the univariate ANOVA framework was introduced in Brockhoff and Skovgaard (1994), leading to multiplicative models for interaction effect expressing the individual varying scale usage. In Smith et al. (2003) and in Brockhoff and Sommer (2008) random effect versions of such analyses were put forward leading to either a multiplicative (nonlinear) mixed model or a linear random coefficient model. Another recurring issue is the relation of multivariate data sets, e.g., trying to predict sensory response by instrumental/spectroscopic and/or chemical measurements. Similarly there is a wish to be able to predict how the market (consumers) will react to sensory changes in food products – then called Preference Mapping (McEwen 1996). This links the area closely to the chemometrics field and also naturally to the (machine) learning area, which in part is explored in Meullenet et al. (2007). Another commonly used sensory and consumer survey methodology is to use rankings or scoring on an ordinal scale. In Rayner et al. (2005) standard and extended rank-based non-parametrics is presented specifically for sensory and consumer data.

As indicated, there are yet many other examples of sensory and consumer data together with other purposes of analysis challenging the sensometrician whoever he or she is. Recently some open-source dedicated sensometrics software have appeared: the R-based SensoMiner (Lê and Husson 2008), the stand-alone tool PanelCheck (Tomic et al. 2007), and the R-package sensR (Christensen and Brockhoff 2009).

About the Author

Per Bruun Brockhoff is Professor in statistics at the Informatics Department at the Technical University of Denmark (since 2004), and Head of the Statistics Section (since 2008). He was the Chairman of DSTS, the Danish Society for Theoretical Statistics (2003–2007), and Chairman of the International Sensometrics Society (2006–2010). Professor Brockhoff co-authored around 60 peer reviewed scientific papers and 2 books. The books are both on Sensometrics: Statistics for Sensory and Consumer Science (with T. Næs and O. Tomic, John Wiley & Sons, 2010), and Nonparametrics for Sensory Science: A More Informative Approach (with J.C.W. Rayner, D.J. Best and G.D. Rayner, Blackwell Publishing, USA, 2005). He is an Elected member of ISI (2005) and currently member of the editorial boards of the two international journals: Food Quality and Preference and Journal of Chemometrics.

Cross References

►Analysis of Variance
►Chemometrics
►Multidimensional Scaling
►Nonlinear Mixed Effects Models
►Random Coefficient Models
►Random Coefficient Models

References and Further Reading

Amerine MA, Pangborn RM, Roessler EB (1965) Principles of sensory evaluation of food. Academic, New York

Bock DR, Jones LV (1968) The measurement and prediction of judgment and choice. Holden-Day, San Francisco

Bradley RA (1958) Triangle, duo-trio, and difference-from-control tests in taste testing. Biometrics 14:566

Bredie WLP, Dehlholm C, Byrne DV, Martens M (2010) Descriptive sensory analysis of food: a review. Submitted to Food Qual Prefer

Bro R, Sidiropoulos ND, Smilde AK (2002) Maximum likelihood fitting using ordinary least squares algorithms. J Chemometr 16(8–10):387–400

Bro R, Qannari EM, Kiers HA, Næs TA, Frøst MB (2008) Multiway models for sensory profiling data. J Chemometr 22: 36–45

Brockhoff PM, Skovgaard IM (1994) Modelling individual differences between assessors in sensory evaluations. Food Qual Prefer 5:215–224

Brockhoff PB, Sommer NA (2008) Accounting for scaling differences in sensory profile data. Proceedings of Tenth European Symposium on Statistical Methods for the Food Industry. pp 283–290, Louvain-La-Neuve, Belgium

Brockhoff P, Hirst D, Næs T (1996) Analysing individual profiles by three-way factor analysis. In: Næs T, Risvik E (eds) Multivariate analysis of data in sensory science, vol 16, Data handling in science and technology. Elsevier Science, B.V., pp 71–102

Brockhoff PB, Næs T, Qannari M (2005) Editorship. J Chemometr 19(3):121
Brockhoff PB, Christensen RHB (2010) Thurstonian models for sensory discrimination tests as generalized linear models. Food Qual Pref 21:330–338
Christensen RHB, Brockhoff PB (2009) sensR: An R-package for thurstonian modelling of discrete sensory data. R-package version 1.1.0. (www.cran.r-project.org/package=sensR/)
Ennis DM (1993) The power of sensory discrimination methods. J Sens Stud 8:353–370
Ennis DM (2003) Foundations of sensory science. In: Moskowitz HR, Munoz AM, Gacula MC (eds) Viewpoints and Controversies in Sensory Science and Consumer Product Testing. Food and Nutrition, Trumbull, CT
Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York
Hopkins JW (1950) A Procedure for quantifying subjective appraisals of odor, flavour and texture of foodstuffs. Biometrics 6(1):1–16
Lawless HT, Heymann H (1999) Sensory evaluation of food. Principles and Practices. Chapman and Hall, New York
Lê S, Husson F (2008) SensoMineR: a package for sensory data analysis. J Sens Stud 23(1):14–25
Lea P, Næs T, Rødbotten M (1997) Analysis of variance of sensory data. Wiley, New York
Luciano G, Næs T (2008) Interpreting sensory data by combining principal component analysis and analysis of variance. Food Qual Pref 20:167–175
MacKay DB (2005) Probabilistic scaling analyses of sensory profile, instrumental and hedonic data. J Chemometr 19(3):180–190
Macmillan NA, Creelman CD (2005) Detection theory, a user's guide, 2nd edn. Mahwah, N.J.: Lawrence Erlbaum Associates
Martens H, Martens M (2001) Multivariate analysis of quality: an introduction. Wiley, Chicester, UK
Martens H, Hoy M, Wise B, Bro R, Brockhoff PB (2003) Prewhitening of data by covariance-weighted pre-processing. J Chemometr 17(3):153–165
McEwen JA (1996) Preference mapping for product optimization. In: Næs T, Risvik E (eds) Multivariate analysis of data in sensory science, vol 16, Data handeling in science and technology. Elsevier Science, B.V., pp 71–102
Meullenet J-F, Xiong R, Findlay CJ (2007) Multivariate and probabilistic analysis of sensory science problems. Blackwell, Ames, USA
Munck L (2007) A new holistic exploratory approach to Systems biology by near infrared spectroscopy evaluated by chemometrics and data inspection. J Chemometr 21:406–426
Næs T, Tomic O, Brockhoff PB (2010) Statistics for sensory and consumer science. Wiley, New York
Rayner JCW, Best DJ, Brockhoff PB, Rayner GD (2005) Nonparametrics for Sensory science: a more informative approach. Blackwell, USA
Schlich P (1998) What are the sensory differences among coffees? Multi-panel analysis of variance and flash analysis. Food Qual Prefer 9:103
Smith A, Cullis B, Brockhoff P, Thompson R (2003) Multiplicative mixed models for the analysis of sensory evaluation data. Food Qual Prefer 14(5–6):387–395
Tomic O, Nilsen AN, Martens M, Næs T (2007) Visualization of sensory profiling data for performance monitoring. LWT – Food Sci Technol 40:262–269

Sequential Probability Ratio Test

Walter W. Piegorsch[1], William J. Padgett[2]
[1]Professor, Chair
University of Arizona, Tucson, AZ, USA
[2]Distinguished Professor Emeritus of Statistics
University of South Carolina, Columbia, SC, USA

Introduction: Sequential Testing and Sequential Probability Ratios

An important topic in statistical theory and practice concerns the analysis of data that are sampled sequentially. The development of powerful mathematical and statistical tools for the analysis of sequential data is a critical area in statistical research. Our emphasis in this short, introductory exposition is on sequential testing, and in particular on the best-known version for such testing, the *sequential probability ratio test*.

Suppose we are given two hypotheses about the underlying distribution of a random variable X: $H_0 : X \sim f_0(x)$ *vs* $H_a : X \sim f_1(x)$, for two probability density functions (pdfs) or probability mass functions (pmfs) $f_i(x), i = 0, 1$. To perform a sequential test of H_0 *vs.* H_a, we sample individual observations one at a time, and assess in a series of separate steps whether or not the accumulated information favors departure from H_0:

Step 0: Begin by setting two constants, A and B, such that $0 < A < 1 < B$.

Step 1: Observe X_1. Compute the probability ratio $\Lambda_1 = f_1(x_1)/f_0(x_1)$. Since very large values of this ratio support H_a, reject H_0 if $\Lambda_1 \geq B$. Alternatively, since very small values of this ratio support H_0, accept H_0 if $\Lambda_1 \leq A$. The sequential approach also allows for an indeterminate outcome, so if $A < \Lambda_1 < B$, continue sampling and go to Step 2.

Step 2: Observe X_2. Compute the probability ratio $\Lambda_2 = f_1(x_1, x_2)/f_0(x_1, x_2)$. As in Step 1, if $\Lambda_2 \geq B$, reject H_0, while if $\Lambda_2 \leq A$, accept H_0. If $A < \Lambda_2 < B$, continue sampling and observe X_3.

⋮

Step n: Observe X_n. Compute the probability ratio $\Lambda_n = f_1(x_1, x_2, \ldots, x_n)/f_0(x_1, x_2, \ldots, x_n)$. As in Step 1, if $\Lambda_n \geq B$, reject H_0, while if $\Lambda_n \leq A$, accept H_0. If $A < \Lambda_n < B$, continue sampling and observe X_{n+1}. (etc.)

This is known as a *Sequential Probability Ratio Test* (*SPRT*), due to Wald (1945a; 1945b).

Notice that in the typical setting where the individual observations are sampled independently from $f_0(x)$ or $f_1(x)$, the probability ratios take the form

$\Lambda_n = \prod_{i=1}^{n}\{f_1(x_i)/f_0(x_i)\}$. Then, the continuance condition $A < \Lambda_n < B$ is equivalent to $\log\{A\} < \log\left\{\prod_{i=1}^{n}[f_1(x_i)/f_0(x_i)]\right\} < \log\{B\}$. For $D_i = \log\{f_1(x_i)\} - \log\{f_0(x_i)\}$ at any $i = 1, 2, \ldots$, this simplifies to

$$\log\{A\} < \sum_{i=1}^{n} D_i < \log\{B\}. \tag{1}$$

An idealized schematic of this procedure can be given, analogous to Fig. 6–13 of Lindgren (1976), for example. For specific choices of f_0 and f_1, one can often simplify (1) even further. Example 1 illustrates the approach.

Example 1 *The Exponential Family*

Suppose we test the simple hypotheses $H_0 : \theta = \theta_0$ *vs* $H_a : \theta = \theta_1$. Let the X_is be independent and identically distributed (i.i.d.) with underlying pdf or pmf taken from the *exponential family* of probability functions (Pierce 1998): $f(x) = h(x)c(\theta)e^{\omega(\theta)t(x)}$. Then, the continuance condition simplifies to $\log\{A\} < n\log\{c(\theta_1)/c(\theta_0)\} + [\omega(\theta_1) - \omega(\theta_0)]\sum_{i=1}^{n} t(X_i) < \log\{B\}$, which if $\omega(\theta_1) - \omega(\theta_0) > 0$ becomes

$$a_n < \sum_{i=1}^{n} t(X_i) < b_n, \tag{2}$$

where

$$a_n = \frac{\log\{A\} - n\log\left[\frac{c(\theta_1)}{c(\theta_0)}\right]}{\omega(\theta_1) - \omega(\theta_0)} \quad \text{and}$$

$$b_n = \frac{\log\{B\} - n\log\left[\frac{c(\theta_1)}{c(\theta_0)}\right]}{\omega(\theta_1) - \omega(\theta_0)}.$$

[If $\omega(\theta_1) - \omega(\theta_0) < 0$, then the inequalities in (2) are reversed.] Notice that the central quantity in (2) is the sufficient statistic $T_n = \sum_{i=1}^{n} t(X_i)$.

For instance, suppose we sample randomly from the single-parameter exponential distribution with mean $\theta, X_i \sim$i.i.d. Exp(θ), and wish to test $H_0 : \theta = \theta_0$ *vs* $H_a : \theta = \theta_1$, where $\theta_1 > \theta_0$. The pdf has the form $f(x|\theta) = \theta^{-1}\exp\{-x/\theta\}I_{(0,\infty)}(x)$, which is a member of the exponential family with $c(\theta) = \theta^{-1}$, $\omega(\theta) = -\theta^{-1}$, and $t(x) = x$. Thus $\log\{\Lambda_n\} = n\log\{\theta_0/\theta_1\} + \left[\theta_0^{-1} - \theta_1^{-1}\right]\sum_{i=1}^{n} X_i$. The continuance region's form can be simplified here by noting that since $\theta_1 > \theta_0$, we have $\omega(\theta_1) - \omega(\theta_0) = \theta_0^{-1} - \theta_1^{-1} > 0$, so (2) applies: continue sampling when $a_n < \sum_{i=1}^{n} X_i < b_n$, for

$$a_n = \frac{\log\{A\} - n\log\left[\frac{\theta_0}{\theta_1}\right]}{\theta_0^{-1} - \theta_1^{-1}} \quad \text{and}$$

$$b_n = \frac{\log\{B\} - n\log\left[\frac{\theta_0}{\theta_1}\right]}{\theta_0^{-1} - \theta_1^{-1}}.$$

Otherwise, reject H_0 when $\sum_{i=1}^{n} X_i \geq b_n$, or accept H_0 when $\sum_{i=1}^{n} X_i \leq a_n$.

Choosing the Sequential Limits *A* and *B*

For most hypothesis tests, concern centers on the testing error rates, i.e., the Type I error rate, $\alpha = P[\text{reject } H_0 | H_0 \text{ true}]$, and the Type II error rate, $\beta = P[\text{accept } H_0 | H_0 \text{ false}]$. For the *SPRT* these quantities will both be functions of A and B, thus one could in principle invert the relationships and select A and B as functions of α and β. Unfortunately, *SPRT* error rates in these forms are difficult to evaluate. It is possible to approximate them, however, as the following theorem shows.

Theorem 1 The *SPRT* as defined above relates its continuance limits and Type I and II error rates via

$$B \leq (1-\beta)/\alpha \qquad \text{and} \qquad A \geq \beta/(1-\alpha). \tag{3}$$

See, e.g., Wald (1947, §3.2) for a proof. The Theorem may be used to define A and B as functions of α and β by choosing A and B to satisfy the equalities in (3): given nominal error rates α^* and β^*, use (3) to set

$$B = (1-\beta^*)/\alpha^* \qquad \text{and} \qquad A = \beta^*/(1-\alpha^*). \tag{4}$$

Of course, these choices of A and B do not ensure that the actual underlying Type I and Type II error rates, α and β, respectively, will attain the nominally-chosen rates α^* and β^*. However, one can produce a series of upper bounds using (3) and (4) to obtain $\alpha + \beta \leq \alpha^* + \beta^*$, $\alpha \leq \alpha^*/(1-\beta^*)$ and $\beta \leq \beta^*/(1-\alpha^*)$. Wald (1947, §3.3) notes that for most typical values of α^* and β^* these bounds are often rather tight and may even be negligible in practice.

Example 2 Suppose we set the nominal error rates to $\alpha^* = 0.01$ and $\beta^* = 0.05$. Then we find $\alpha + \beta \leq 0.06$, while the individual error rates are bounded as $\alpha \leq (0.01)/(0.95) = 0.0105$ and $\beta \leq (0.05)/(0.99) = 0.0505$.

Finite Termination and Average Sample Number (ASN)

Notice that the (final) sample size N of any sequential test procedure is not a fixed quantity, but is in fact a random

variable determined from the data. As such, an obvious concern with any form of sequential test is whether or not the method eventually terminates. Luckily, for i.i.d. sampling the *SPRT* possesses a finite termination characteristic in that $P[N < \infty] = 1$. This holds under either H_0 or H_a, and is based on a more general result given by Wald (1944); also see Lehmann (1959, §3.10). The larger literature on finite termination of sequential tests is quite diverse; some historically interesting expositions are available in, e.g., David and Kruskal (1956), Savage and Savage (1965), or Wijsman (1967).

When $P[N < \infty] = 1$, it is reasonable to ask what the *expected* sample size, $E[N]$, is for a given *SPRT*. This is known as the *average sample number* (*ASN*) or *expected sample number* (*ESN*). A basic result for the *ASN* is available via the following theorem (Wald 1945b):

Theorem 2 (Wald's Equation): Let $D_1, D_2, \ldots$ be a sequence of i.i.d. random variables with $E[\,|D_i|\,] < \infty$. Let $N > 0$ be an integer-valued random variable whose realized value, n, depends only on $D_1, \ldots, D_n$, with $E[N] < \infty$. Then $E[D_1 + D_2 + \ldots + D_N] = E[N] \cdot E[D_1]$.

A consequence of Wald's Equation is the immediate application to the *SPRT* and its *ASN*. Clearly $\log\{\Lambda_N\} = \log\{f_1(x_1)/f_0(x_1)\} + \cdots + \log\{f_1(x_N)/f_0(x_N)\} = \sum_{i=1}^{N} D_i$. So, applying Wald's equation yields $E[N] = E[\log\{\Lambda_N\}]/E[D]$, where $D = \log\{f_1(X)/f_0(X)\}$. This result lends itself to a series of approximations. For instance, if H_0 is rejected at some N, $\log\{\Lambda_N\} \approx \log\{B\}$. Or, if H_0 is accepted at some N, $\log\{\Lambda_N\} \approx \log\{A\}$. Thus, under H_0, $E[\log\{\Lambda_N\}|H_0] \approx \alpha \cdot \log\{B\} + (1-\alpha)\log\{A\}$, so $E[N|H_0] \approx [\alpha \cdot \log\{B\} + (1-\alpha)\log\{A\}]/E[D|H_0]$. Similarly, $E[N|H_a] \approx [(1-\beta)\log\{B\} + \beta \cdot \log\{A\}]/E[D|H_a]$. For any given parametric configuration, these relationships may be used to determine approximate values for *ASN*. Wald (1946) gives some further results on ways to manipulate the *ASN*.

An important reason for employing the *SPRT*, at least for the case of testing simple hypotheses, is that it achieves optimal *ASN*s: if the X_is are i.i.d., then for testing $H_0 : \theta = \theta_0$ *vs.* $H_a : \theta = \theta_1$ both $E[N|H_0]$ and $E[N|H_a]$ are minimized among all sequential tests whose error probabilities are at most equal to those of the *SPRT* (Wald and Wolfowitz 1948). For testing composite hypotheses, the theory of *SPRT*s is more complex, although a variety of interesting results are possible (Stuart et al. 1999, §24.23–24; Lai 2001, §2). In his original article, Wald (1945a) himself discussed the problem of sequential testing of composite hypotheses on a binomial parameter; also see Siegmund (1985, §II.3). For testing with normally distributed samples, various forms of sequential t-tests have been proposed; see Jennison and Turnbull (1991) and the references therein for a useful discussion on sequential t-tests (and sequential χ^2- and F-tests) that includes the important problem of *group sequential testing*.

Since Wald's formalization of the *SPRT*, a number of powerful, alternative formulations/constructions have led to wide application of the method. We provide here a short introduction to the basic mathematical underpinnings; however, comprehensive reviews on the larger area of sequential analysis date as far back as Johnson (1961), along with more modern expositions given by Lai (1998, 2001, 2004) and Ghosh (2004). For a perspective emphasizing ▶sequential sampling, see Mukhopadhyay (2002). Also see the book-length treatments by Siegmund (1985), Ghosh and Sen (1991), or Mukhopadhyay and de Silva (2008), along with Wald's (1947) classic text. For cutting-edge developments a dedicated scientific journal exists: *Sequential Analysis*, with more information available online at the website http://www.informaworld.com/smpp/title~db=all~content=t713597296.

Acknowledgments

Thanks are due the Editor and an anonymous referee for their helpful suggestions on an earlier draft of the manuscript. The first author's research was support in part by grant #RD-83241902 from the U.S. Environmental Protection Agency and by grant #R21-ES016791 from the U.S. National Institute of Environmental Health Sciences. The contents herein are solely the responsibility of the authors and do not necessarily reflect the official views of these agencies.

About the Authors

Walter W. Piegorsch is Chair of the Graduate Interdisciplinary Program (GIDP) in Statistics at the University of Arizona, Tucson, AZ. He is also a Professor of Mathematics, a Professor of Public Health, and Director of Statistical Research & Education at the University's BIO5 Institute for Collaborative Bioresearch. Professor Piegorsch has held a number of professional positions, including Chairman of the American Statistical Association Section on Statistics & the Environment (2004), Vice-Chair of the American Statistical Association Council of Sections Governing Board (1997–1999), and election to the Council of the International Biometric Society (2002–2005). He serves as Editor-in-Chief of *Environmetrics*, and also has served as Joint-Editor of the *Journal of the American Statistical Association* (Theory & Methods Section).

Dr. Piegorsch was named a Fellow of the American Statistical Association (1995), a Member (by Election, 1995) of the International Statistical Institute, and has received the Distinguished Achievement Medal of the American Statistical Association Section on Statistics and the Environment (1993), and was a Co-recipient of The Ergonomics Society/Elsevier Ltd. Applied Ergonomics Award (2007).

For biography of William J. Padgett *see* the entry ►Weibull distribution.

Cross References

►Exponential Family Models
►Optimal Stopping Rules
►Sequential Sampling

References and Further Reading

David HT, Kruskal WH (1956) The WAGR sequential *t*-test reaches a decision with probability one. Ann Math Stat 27: 797–805
Ghosh BK (2004) Sequential analysis. In: Kotz S, Read CB, Balakrishnan N, Vidakovic B (eds) Encyclopedia of statistical sciences vol 11, 2nd edn. Wiley, New York, pp 7605–7613
Ghosh BK, Sen PK (eds) (1991) Handbook of sequential analysis, M. Dekker, New York
Jennison C, Turnbull BW (1991) Exact calculations for sequential *t*, χ^2 and *F* tests. Biometrika 78:133–141
Johnson NL (1961) Sequential analysis: a survey. J Roy Stat Soc Ser A (General) 124:372–411
Lai TL (1998) Sequential analysis. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 5. Wiley, New York, pp 4074–4079
Lai TL (2001) Sequential analysis: some classical problems and new challenges (with discussion). Stat Sin 11:303–408
Lai TL (2004) Likelihood ratio identities and their applications to sequential analysis (with discussion). Sequential Anal 23: 467–556
Lehmann EL (1959) Testing statistical hypotheses, 1st edn. Wiley, New York
Lindgren BW (1976) Statistical theory, 3rd edn. Macmillan, New York
Mukhopadhyay N (2002) Sequential sampling. In: El-Shaarawi AH, Piegorsch WW (eds) Encyclopedia of environmetrics, vol 4. Wiley, Chichester, pp 1983–1988
Mukhopadhyay N, de Silva BM (2008) Sequential methods and their applications. Chapman & Hall/CRC, Boca Raton, FL
Pierce DA (1998) Exponential family. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 2. Wiley, New York, pp 1448–1455
Savage IR, Savage LJ (1965) Finite stopping time and finite expected stopping time. J Roy Stat Soc Ser B (Meth) 27:284–289
Siegmund D (1985) Sequential analysis: tests and confidence intervals. Springer-Verlag, New York
Stuart A, Ord JK, Arnold S (1999) Kendall's advanced theory of statistics: Volume 2A-Classical inference and the linear model, 6th edn. Arnold, London
Wald A (1944) On cumulative sums of random variables. Ann Math Stat 15:283–296
Wald A (1945a) Sequential tests of statistical hypotheses. Ann Math Stat 16:117–186
Wald A (1945b) Some generalizations of the theory of cumulative sums of random variables. Ann Math Stat 16:287–293
Wald A (1946) Some improvements in setting limits for the expected number of observations required by a sequential probability ratio test. Ann Math Stat 17:466–474
Wald A (1947) Sequential analysis. Wiley, New York
Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. Ann Math Stat 19:326–339
Wijsman RA (1967) General proof of termination with probability one of invariant sequential probability ratio tests based on multivariate normal observations. Ann Math Stat 38: 8–24

Sequential Ranks

Estate V. Khmaladze
Professor
Victoria University of Wellington, Wellington, New Zealand

To discuss sequential ranks it will be more helpful to present them in comparison with ordinary ranks.

Suppose $X_1, \ldots, X_n$ is a sequence of random variables. Denote by $\mathbb{I}_A$ the indicator function of an event A. For each X_i consider now what one can call its "ordinary" rank:

$$R_{in} = \sum_{j=1}^{n} \mathbb{I}_{\{X_j \le X_i\}}.$$

So, R_{in} counts the number of our random variables that take values not exceeding X_i. For example, if X_i happens to be the smallest, its rank will be 1, and if it happens to be the largest, its rank will be n. If the joint distribution of $X_1, \ldots, X_n$ is absolutely continuous, then with probability 1 all values of our random variables will be different. Therefore, for any integer $k = 1, \ldots, n$ there will be one and only one random variable with rank equal to k. For example, for $n = 5$, if our X_i-s happened to be

$$-1.31,\ 0.24,\ -3.52,\ 4.11\ and\ 2.25,$$

their ranks will be

$$2,\quad 3,\quad 1,\quad 5\quad and\quad 4.$$

Hence, the vector of "ordinary" ranks $\mathbb{R}_n = \{R_{1n}, \ldots, R_{nn}\}$ is a random permutation of the numbers $\{1, \ldots, n\}$. Thus, its distribution possesses a certain degeneracy. In particular, even if $X_1, \ldots, X_n$ are independent and identically distributed, the ordinary ranks are dependent random variables – for example, if $R_{in} = 3$ it precludes any other rank $R_{jn}, j \neq i$, from taking the value 3, so that the conditional probability $P(R_{jn} = 3 | R_{in} = 3) = 0$, while without this condition $P(R_{jn} = 3)$ does not need to be 0 at all.

Moreover, any symmetric statistic from the vector $\mathbb{R}_n$ is not random and, for given n, must be constant: if ψ is a symmetric function of its n arguments, then

$$\psi(R_{1n},\ldots,R_{nn}) = \psi(1,\ldots,n), \quad \text{e.g.,} \quad \sum_{i=1}^{n}\phi(R_{in}) = \sum_{i=1}^{n}\phi(i).$$

The definition of sequential ranks is slightly different, but the difference in their properties is quite remarkable. Namely, the sequential rank of X_i is defined as

$$S_i = \sum_{j=1}^{i}\mathbb{I}_{\{X_j\le X_i\}}.$$

Therefore, it is the rank of X_i among only "previous" observations, including X_i itself, but not "later" observations $X_{i+1},\ldots,X_n$. For the sample values given above, their sequential ranks are

$$1, \quad 2, \quad 1, \quad 4, \quad 4.$$

The relationship between the vectors of ordinary ranks and sequential ranks is one-to-one. Namely, given vector $\mathbb{R}_n = \{R_{1n},\ldots,R_{nn}\}$ of ordinary ranks, the sums

$$S_i = \sum_{j=1}^{i}\mathbb{I}_{\{R_{jn}\le R_{in}\}}$$

return sequential ranks of $X_1,\ldots,X_n$ and the other way around, given a vector of sequential ranks $\mathbb{S}_n$, if

$$S_{i,i+1} = S_i + \mathbb{I}_{\{S_i\ge S_{i+1}\}}, S_{i,i+2} = S_{i,i+1} + \mathbb{I}_{\{S_{i,i+1}\ge S_{i+2}\}},\ldots,$$

then finally

$$S_{i,n} = R_{in}.$$

Because of this one-to-oneness, the vector $\mathbb{S}_n$ also must have some sort of degeneracy. It does, but in a very mild form: S_1 is always 1.

Assume that $X_1,\ldots,X_n$ are independent and identically distributed random variables with continuous distribution function F. Then $U_1 = F(X_1),\ldots,U_n = F(X_n)$ are independent uniformly distributed on $[0,1]$ random variables. The values of R_{in} and S_i will not change, if we replace X_i-s by U_i-s. Therefore, the distribution of both ranks must be independent of F – they both are "distribution free." We list some properties of $\mathbb{S}_n$ in this situation – they can be found, e.g., in Barndolf-Nielsen (1963), Renyi (1962, 1976), Sen (1981).

The distribution of each S_i is $P(S_i = k) = 1/i, k = 1,\ldots,i$, and, therefore, the distribution function of $S_i/(i+1)$ quickly converges to the uniform distribution function:

$$P\left(\frac{S_i}{i+1} = \frac{k}{i+1}\right) = \frac{1}{i}, \text{ and } \left|P\left(\frac{S_i}{i+1}\le x\right) - x\right| \le \frac{1}{i+1}.$$

Recall that, similarly, for ordinary ranks $P(R_{in} = k) = 1/n, k = 1,\ldots,n$, see, e.g., Hajek and Shidak (1975). However, unlike ordinary ranks, sequential ranks $S_1,\ldots,S_n$ are independent random variables. Hence symmetric statistics from sequential ranks are non-degenerate random variables. For example,

$$\sum_{i=1}^{n}\phi(S_i)$$

is a sum of independent random variables. Also unlike ordinary ranks, with arrival of a new observation X_{n+1} sequential ranks $S_1,\ldots,S_n$ stay unchanged and only one new rank S_{n+1} is to be calculated.

Therefore, asymptotic theory of sequential ranks is relatively simple and computationally they are very convenient.

The ordinary ranks are used in testing problems, usually, through the application of two types of statistics–widely used linear rank statistics and goodness of fit statistics, based on the empirical field

$$z_R(t,u) = \sum_{i=1}^{nt}\left[\mathbb{I}_{\{R_{in}\le u(n+1)\}} - \frac{[nu]}{n+1}\right], \quad (t,u)\in[0,1]^2.$$

Linear rank statistics can also be thought of as based on the field $z_R(t,u)$, and, more exactly, are linear functionals from it:

$$\psi(\mathbb{R}_n) = \int \psi(t,u)z_R(dt,du) = \sum_{i=1}^{n}\left[\psi\left(\frac{i}{n},\frac{R_{in}}{n+1}\right) - E\psi\left(\frac{i}{n},\frac{R_{in}}{n+1}\right)\right]$$

(the term "linear" would not be very understandable otherwise). Without loss of generality one can assume that $\int_0^1\psi(t,u)dt = 0$.

One of the central results in the theory of rank tests, see Hajek and Shidak (1975), is the optimality statement about linear rank statistics. If under the null hypothesis the sample is i.i.d.(F) while under the alternative hypothesis the distribution A_i of each X_i is such that

$$\frac{dA_i(x)}{dF(x)} = 1 + \frac{1}{\sqrt{n}}a\left(\frac{i}{n},F(x)\right) + \textit{smaller terms}, \quad \text{as} n\to\infty, \tag{1}$$

where $\int_0^1 a(t,F(x))dt = 0$, then the linear rank statistic, with ψ equal to a from (1),

$$a(\mathbb{R}_n) = \sum_{i=1}^{n}a\left(\frac{i}{n},\frac{R_{in}}{n+1}\right),$$

is asymptotically optimal against this alternative. Indeed, the statistic

$$\sum_{i=1}^{n}a\left(\frac{i}{n},F(X_i)\right)$$

is the statistic of the asymptotically optimal test for our alternative, based on the observations $X_1, \ldots, X_n$ "themselves," and $R_{in}/(n+1)$ is a "natural" approximation for $F(X_i)$.

Returning to sequential ranks, one can again consider the empirical field

$$z_S(t,u) = \sum_{i=1}^{nt} \left[\mathbb{I}_{\{S_i \le u(i+1)\}} - \frac{[iu]}{i+1} \right], \quad (t,u) \in [0.1]^2,$$

and sequential linear rank statistics, based on it:

$$\phi(\mathbb{S}_n) = \int \phi(t,u) z_S(dt,du) = \sum_{i=1}^{n} \left[\phi\left(\frac{i}{n}, \frac{S_i}{i+1}\right) - E\phi\left(\frac{i}{n}, \frac{S_i}{i+1}\right) \right].$$

Although $S_i/(i+1)$ is no less "natural" an approximation for $F(X_i)$, the statistic

$$a(\mathbb{S}_n) = \sum_{i=1}^{n} a\left(\frac{i}{n}, \frac{S_i}{i+1}\right)$$

is not optimal for the alternative (1) any more. The papers (Khmaladze 1986) and (Pardzhanadze 1986) derived the form of this optimal statistic, and hence established the theory of sequential ranks to the same extent as the theory of "ordinary" rank statistics.

More Specifically, it was shown that the empirical fields z_R and z_S are asymptotically linear transformations of each other and, as a consequence, the two linear rank statistics $\psi(\mathbb{R}_n)$ and $\phi(\mathbb{S}_n)$ have the same limit distribution under the null hypothesis and under any alternative (1) as soon as functions ψ and ϕ are linked as below:

$$\psi(t,u) - \frac{1}{t}\int_0^t \psi(\tau,u)d\tau = \phi(t,u) \quad \text{or}$$
$$\phi(t,u) - \int_t^1 \frac{1}{\tau}\phi(\tau,u)d\tau = \psi(t,u).$$

In particular, both linear rank statistics

$$\sum_{i=1}^{n} a\left(\frac{i}{n}, \frac{R_{in}}{n+1}\right) \quad \text{and} \quad \sum_{i=1}^{n} \left[a\left(\frac{i}{n}, \frac{S_i}{i+1}\right) - \frac{n}{i}\int_0^{i/n} a\left(\tau, \frac{S_i}{i+1}\right) d\tau \right] \tag{2}$$

are asymptotically optimal test statistics against alternative (1).

Two examples of particular interest should clarify the situation further.

Example 1 *(Wilcoxon rank (or rank-sum) statistic).* In the two-sample problem, when we test if both samples came from the same distribution or not, the following Wilcoxon rank statistic

$$\sum_{i=1}^{m} \frac{R_{in}}{n+1}$$

is most widely used (see ►Wilcoxon–Mann–Whitney Test). Its sequential analogue is not mentioned often, but according to (2) there is such an analogue, which is

$$-\sum_{i=m+1}^{n} \frac{m}{i}\frac{S_i}{i+1}.$$

In general, the following two statistics are asymptotically equivalent:

$$\sum_{i=1}^{m} a\left(\frac{R_{in}}{n+1}\right) \quad \text{and} \quad -\sum_{i=m+1}^{n} \frac{m}{i} a\left(\frac{S_i}{i+1}\right).$$

Note again, that if the size m of the first sample is fixed, but we keep adding new observations to the second sample, so that $n-m$ keeps increasing, we would only need to add new summands to the sequential rank statistics, on the right, without changing the previous summands.

Example 2 *(Kendall's τ and Spearman's ρ rank correlation coefficients).* The latter correlation coefficient has the form

$$\rho_n = \sum_{i=1}^{n} \frac{i}{n}\left(\frac{R_{in}}{n+1} - \frac{1}{2}\right)$$

while the former is

$$\tau_n = \sum_{i=1}^{n} \frac{i}{n}\left(\frac{S_i}{i+1} - \frac{1}{2}\right).$$

These two coefficients are usually perceived as different statistics. However, from (2) it follows that they also are asymptotically equivalent.

Among other papers that helped to form and advance the theory of sequential ranks we refer to Müller-Funk (1983), Renyi (1962, 1976), and Reynolds (1975). Among more recent papers and applications to change-point problem we would point to Bhattacharya and Zhou (1994), Gordon and Pollak (1994), and Malov (1993).

About the Author

For biography *see* the entry ►Testing Exponentiality of Distribution.

Cross References

►Kendall's Tau
►Measures of Dependence
►Record Statistics
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Barndolf-Nielsen O (1963) On limit behaviour of extreme order statistics. Ann Math Stat 34:992–1002

Bhattacharya PK, Zhou H (1994) A rank cusum procedure for detecting small changes in a symmetric distribution, Change-Point problems. IMS Lecture notes, vol 23

Gordon L, Pollak M (1994) An efficient sequential nonparametric scheme for detecting a change in distribution. Ann Stat 22: 763–804

Hajek J, Shidak Z (1975) Theory of rank tests Academic, CSZV, Prague

Khmaladze EV, Parjanadze AM (1986) Functional limit theorems for linear statistics of sequential ranks. Probab theor relat fields 73:1285–1295

Malov SV (1993) Sequential ranks and order statistics. Notes Sci Seminars POMI 204:115–125

Müller-Funk U (1983) Sequential signed rank statistics. Sequential Anal Design Meth Appl 2:123–148

Pardzhanadze AM, Khmaladze EV (1986) On the asymptotic theory of statistics based on sequential ranks. Theor Probab Appl 31:669–682

Renyi A (1962, 1976) On the extreme elements of observations Academiai Kiado. Selected papers of Alfred Renyi

Reynolds M (1975) A sequential rank test for symmetry. Ann Stat 3:382–400

Sen PK (1981) Sequential non-parametrics. Wiley, New York

Sequential Sampling

Nitis Mukhopadhyay
Professor
University of Connecticut-Storrs, Storrs, CT, USA

Introduction

Sequential sampling entails observing data in a sequence. How long should one keep observing data? That will largely depend on the preset levels of errors that one may be willing to live with and the optimization techniques that may be required. In the early 1940s, Abraham Wald developed the theory and practice of the famous *sequential probability ratio test* (SPRT) to decide between a simple null hypothesis and a simple alternative hypothesis (Wald 1947). Wald and Wolfowitz (1948) proved optimality of Wald's SPRT within a large class of tests, including Neyman and Pearson's (1933) UMP test, in the sense that the SPRT needs on an average fewer observations under either hypothesis. These were mentioned in another chapter.

For a comprehensive review, one should refer to the *Handbook of Sequential Analysis*, a landmark volume that was edited by Ghosh and Sen (1991). This nearly 20 years old handbook is still one of the most prized resource in this whole field.

Section ▶"Why Sequential Sampling?" explains with Examples 1 and 2 why one must use sequential sampling strategies to solve certain statistical problems. We especially highlight the Stein (1945, 1949) path-breaking two-stage and the Ray (1957) and Chow and Robbins (1965) purely sequential fixed-width confidence interval procedures in sections ▶"Stein's Two-stage Sampling" and "Purely Sequential Sampling" respectively.

Sections ▶"Two-stage Sampling" and "Purely Sequential Sampling" analogously highlight the Ghosh and Mukhopadhyay (1976) two-stage and the Robbins (1959) purely sequential bounded-risk point estimation procedures respectively. Both sections ▶"Two-stage and Sequential Fixed-width Confidence Interval" and "Two-stage and Sequential Bounded Risk Point Estimation" handle the problems of estimating an unknown mean of a normal distribution whose variance is also assumed unknown.

Section ▶"Which Areas Are Hot Beds for Sequential Sampling?" briefly mentions applications of sequential and multi-stage sampling strategies in concrete problems that are in the cutting edge of statistical research today.

Why Sequential Sampling?

There is a large body of statistical inference problems that cannot be solved by any fixed-sample-size procedure. We will highlight two specific examples. Suppose that $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$ where $-\infty < \mu < \infty, 0 < \sigma^2 < \infty$ are both unknown parameters, and $n(\geq 2)$ is fixed.

Example 1 We want to construct a confidence interval I for μ such that (i) the length of I is $2d(> 0)$ where d is preassigned, and (ii) the associated confidence coefficient, $P_{\mu,\sigma^2}\{\mu \in I\} \geq 1 - \alpha$ where $0 < \alpha < 1$ is also preassigned. Dantzig (1940) showed that this problem has no solution regardless of the form of the confidence interval I when n is fixed in advance.

Example 2 Suppose that $\overline{X}_n$, the sample mean, estimates μ and we want to claim its bounded-risk property, namely that $\sup_{\mu,\sigma^2} E[(\overline{X}_n - \mu)^2] \leq \omega$ where $\omega(> 0)$ is a preassigned risk-bound. This problem also has no solution regardless of the form of the estimator of μ.

Theorem 1 *Suppose that $X_1, \cdots, X_n$ are iid with a probability density function $\frac{1}{\sigma} f\left(\sigma^{-1}(x - \theta)\right)$ where $-\infty < \theta < \infty, 0 < \sigma < \infty$ are two unknown parameters. For estimating θ, let the loss function be given by $W(\theta, \delta(\boldsymbol{x})) = H(|\delta(\boldsymbol{x}) - \theta|)$ where $\boldsymbol{x} = (x_1, \cdots, x_n)$ is a realization of $\boldsymbol{X} = (X_1, \cdots, X_n)$. Assume that $H(|u|) \uparrow |u|$, and let $M =$*

$\sup_{-\infty<u<\infty} H(|u|)$, *which may be infinite. Then, for any fixed* $L < M$, *there does not exist an estimator* $\delta(\boldsymbol{X})$ *such that* $\sup_{\theta,\sigma} E_{\theta,\sigma}\{W(\theta, \delta(\boldsymbol{X}))\} \le L$.

This statement is similar to that of Theorem 3.7.1 in Ghosh et al. (1997) and Theorem 2.3.1 in Mukhopadhyay and de Silva (2009). It was originally proved in Lehmann (1951).

Theorem 1 proves immediately the non-existence of a fixed-sample-size methodology to solve the problems mentioned in Examples 1–2 exactly. There are these and numerous other inference problems where we have no fixed-sample-size procedure at all to talk about. In order to address this class of important inference problems, an appropriately designed sequential sampling procedure is a must.

Two-Stage and Sequential Fixed-Width Confidence Interval

In the context of Example 1, we first summarize Stein's (1945, 1949) two-stage procedure and then the purely sequential procedure due to Ray (1957) and Chow and Robbins (1965).

Stein's Two-Stage Sampling

Stein (1945, 1949) gave his path-breaking two-stage sampling design to solve *exactly* the problem mentioned in Example 1. One begins with pilot observations $X_1, \cdots, X_m$ with a pilot or initial sample size $m(\ge 2)$. Let $a_{m-1} \equiv a_{m-1,\alpha/2}$ be the upper $50\alpha\%$ point of the Student's t distribution with $m-1$ degrees of freedom. Now, based on $X_1, \cdots, X_m$, we obtain the sample variance, $S_m^2 = (m-1)^{-1}\Sigma_{i=1}^m (X_i - \overline{X}_m)^2$ which estimates unknown σ^2. Let us denote $\langle u \rangle$ = the largest integer $< u$, $u > 0$.

We define the final sample size as

$$N \equiv N(d) = \max\left\{m, \left\langle \frac{a_{m-1}^2 S_m^2}{d^2} \right\rangle + 1\right\}. \tag{1}$$

It is easy to see that N is finite with probability one. This two-stage procedure is implemented as follows:

If $N = m$, it indicates that we already have too many observations at the pilot stage. Hence, we do not need any more observations at the second stage.

But, if $N > m$, it indicates that we have started with too few observations at the pilot stage. Hence, we sample the difference at the second stage by gathering new observations $X_{m+1}, \cdots, X_N$ at the second stage.

Case 1. If $N = m$, the final dataset is $X_1, \cdots, X_m$

Case 2. If $N > m$, the final dataset is $X_1, \cdots, X_m, X_{m+1}, \cdots, X_N$

Combining the two possibilities, one can say that the final dataset is composed of N and $X_1, \cdots, X_N$. This gives rise to the sample mean $\overline{X}_N$ and the associated fixed-width interval $I_N = \left[\overline{X}_N \pm d\right]$.

It is clear that (i) the event $\{N = n\}$ depends only on the random variable S_m^2, and (ii) $\overline{X}_n, S_m^2$ are independent random variables, for all fixed $n(\ge m)$. So, any event defined only through $\overline{X}_n$ must be independent of the event $\{N = n\}$. Using these tools, Stein (1945, 1949) proved the following result that is considered a breakthrough. More details can be found in Mukhopadhyay and de Silva (2009, Sect. 6.2.1).

Theorem 2 $P_{\mu,\sigma^2}\left\{\mu \in \left[\overline{X}_N \pm d\right]\right\} \ge 1 - \alpha$ *for all fixed* $d > 0$, $0 < \alpha < 1$, μ, *and* σ^2.

It is clear that the final sample size N from (1) tried to mimic the optimal fixed sample size C, the smallest integer $\ge z_{\alpha/2}^2 \sigma^2 d^{-2}$, had σ^2 been known. This procedure, however, is known for its significant oversampling on an average.

Purely Sequential Sampling

In order to overcome significant oversampling, Ray (1957) and Chow and Robbins (1965) proposed a purely sequential procedure. One begins with pilot observations $X_1, \cdots, X_m$ with a pilot or initial sample size $m(\ge 2)$, and then proceed by taking one additional observation at-a-time until the sampling process terminates according to the following stopping rule: With $\overline{X}_n = n^{-1}\Sigma_{i=1}^n X_i$ and $S_n^2 = (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X}_n)^2$, let

$$N \equiv N(d) = \inf\left\{n \ge m : n \ge \frac{z_{\alpha/2}^2 S_n^2}{d^2}\right\}. \tag{2}$$

It is easy to see that N is finite with probability one. Based on the final dataset composed of N and $X_1, \cdots, X_N$, one finds $\overline{X}_N$ and proposes the associated fixed-width interval $I_N = \left[\overline{X}_N \pm d\right]$. Now, one can prove that asymptotically, $P_{\mu,\sigma^2}\{\mu \in \left[\overline{X}_N \pm d\right]\} \to 1 - \alpha$ for all fixed $0 < \alpha < 1$, μ, and σ^2 as $C \to \infty$ when $m \ge 2$.

One can also prove that $E_{\sigma^2}[N - C] = -1.1825$ if $m \ge 4$. This property is referred to as the *asymptotic second-order efficiency* according to Ghosh and Mukhopadhyay (1981). One has to employ mathematical tools from nonlinear renewal theory to prove such a property. The nonlinear renewal theory has been fully developed by Woodroofe (1977) and Lai and Siegmund (1977, 1979).

Two-Stage and Sequential Bounded Risk Point Estimation

In the context of Example 2, we first summarize a two-stage procedure from Ghosh and Mukhopadhyay (1976) followed by a purely sequential procedure along the line of Robbins (1959).

Two-Stage Sampling

Ghosh and Mukhopadhyay (1976) discussed a two-stage sampling design analogous to (1) to solve *exactly* the problem mentioned in Example 2. We again start with pilot observations $X_1, \cdots, X_m$ where $m(\geq 4)$ is the pilot size and obtain S_m^2. Define the final sample size as:

$$N \equiv N(\omega) = \max\left\{m, \left\langle \frac{b_m S_m^2}{\omega} \right\rangle + 1\right\} \qquad (3)$$

where $b_m = \frac{m-1}{m-3}$. It is easy to see that N is finite with probability one.

The two-stage sampling scheme is implemented as before.

Case 1. If $N = m$, the final dataset is $X_1, \cdots, X_m$

Case 2. If $N > m$, the final dataset is $X_1, \cdots, X_m, X_{m+1}, \cdots, X_N$

Combining the two situations, one can see that the final dataset is again composed of N and $X_1, \cdots, X_N$ which give rise to an estimator $\overline{X}_N$ for μ.

Now, we recall that $\overline{X}_n$ is independent of the event $\{N = n\}$ for all fixed $n(\geq m)$. Hence, we can express the risk associated with the estimator $\overline{X}_N$ as follows:

$$E_{\mu,\sigma^2}\left\{(\overline{X}_N - \mu)^2\right\} = \sigma^2 E_{\mu,\sigma^2}\left[N^{-1}\right],$$

which will not exceed the set risk-bound ω for all fixed μ and σ^2. More details can be found in Mukhopadhyay and de Silva (2009, Sect. 6.3).

It is clear that the final sample size N from (3) tried to mimic the optimal fixed sample size n^*, the smallest integer $\geq \sigma^2\omega^{-1}$, had σ^2 been known. This procedure is also well-known for its significant oversampling on an average.

For either problem, there are more efficient two-stage, three-stage, accelerated sequential, and other estimation methodologies available in the literature. One may begin by reviewing this field from Mukhopadhyay and Solanky (1994), Ghosh et al. (1997), Mukhopadhyay and de Silva (2009), among other sources.

Purely Sequential Sampling

In order to overcome significant oversampling, along the line of Robbins (1959), one can propose the following purely sequential procedure. One begins with pilot observations $X_1, \cdots, X_m$ with a pilot or initial sample size $m(\geq 2)$, and then proceed by taking one additional observation at-a-time until the sampling process terminates according to the following stopping rule: Let

$$N \equiv N(\omega) = \inf\left\{n \geq m : \; n \geq \frac{S_n^2}{\omega}\right\}. \qquad (4)$$

It is easy to see that N is finite with probability one. Based on the final dataset composed of N and $X_1, \cdots, X_N$, one finds $\overline{X}_N$ and proposes the associated estimator $\overline{X}_N$ for μ. Now, one can prove that asymptotically, $\omega^{-1}E_{\mu,\sigma^2}\left\{(\overline{X}_N - \mu)^2\right\} \to 1$ for all fixed μ, and σ^2 as $n^* \to \infty$ when $m \geq 2$.

One can again prove that $E_{\sigma^2}[N - C]$ is bounded by appealing to nonlinear renewal theory. This property is referred to as the *asymptotic second-order efficiency* according to Ghosh and Mukhopadhyay (1981).

Which Areas Are Hot Beds for Sequential Sampling?

First, we should add that all computer programs necessary to implement the sampling strategies mentioned in sections ▶"Two-stage and Sequential Fixed-width Confidence Interval" and "Two-stage and Sequential Bounded Risk Point Estimation" are available in conjunction with the recent book of Mukhopadhyay and de Silva (2009).

Sequential and multi-stage sampling techniques are implemented practically in all major areas of statistical science today. Some modern areas of numerous applications *include* change-point detection, clinical trials, computer network security, computer simulations, ▶data mining, disease mapping, educational psychology, financial mathematics, group sequential experiments, horticulture, infestation, kernel density estimation, longitudinal responses, multiple comparisons, nonparametric functional estimation, ordering of genes, ▶randomization tests, reliability analysis, scan statistics, selection and ranking, sonar, surveillance, survival analysis, tracking, and water quality.

In a majority of associated statistical problems, sequential and multi-stage sampling techniques are absolutely essential in the sense of our prior discussions in section ▶"Why Sequential Sampling?". In other problems, appropriate sequential and multi-stage sampling techniques are more efficient than their fixed-sample-size counterparts, if any.

For an appreciation of concrete real-life problems involving many aspects of sequential sampling, one may refer to *Applied Sequential Methodologies*, a volume edited by Mukhopadhyay et al. (2004).

About the Author

Dr. Nitis Mukhopadhyay is professor of statistics, Department of Statistics, University of Connecticut, USA. He is Editor-in-Chief of *Sequential Analysis* since 2004. He is Associate Editor for *Calcutta Statistical Association Bulletin* (since 1998), *Communications in Statistics* (since 2002) and *Statistical Methodology* (since 2004). He is Chair of the National Committee on Filming Distinguished Statisticians of the American Statistical Association since 2002. In 2002, he has been named IMS fellow for "outstanding contribution in sequential analysis and multistage sampling; pathbreaking research in selection and ranking; authoritative books; exemplary editorial service; innovative teaching and advising; and exceptional dedication to preserve and celebrate statistical history through films and scientific interviews." He is also an Elected Fellow of The American Statistical Association (2003), and Elected Ordinary Member of The International Statistical Institute (2007), and a life member of: the International Indian Statistical Association, the Calcutta Statistical Association and the Statistical Society of Sri Lanka. Professor Mukhopadhyay was elected a Director of the Calcutta Statistical Association for the period 2005–2008. He has authored/coauthored about 170 papers in international journals and 7 books including, *Sequential Methods and Their Applications* (Chapman & Hall/CRC, Boca Raton, 2009).

Cross References

►Acceptance Sampling
►Loss Function
►Optimal Stopping Rules
►Ranking and Selection Procedures and Related Inference Problems
►Sampling Algorithms
►Sequential Probability Ratio Test

References and Further Reading

Chow YS, Robbins H (1965) On the asymptotic theory of fixed width sequential confidence intervals for the mean. Ann Math Stat 36:457–462
Dantzig GB (1940) On the non-existence of tests of Student's hypothesis having power functions independent of σ. Ann Math Stat 11:186–192
Ghosh BK, Sen PK (eds) (1991) Handbook of sequential analysis. Marcel Dekker, New York
Ghosh M, Mukhopadhyay N (1976) On two fundamental problems of sequential estimation. Sankhya B 38:203–218
Ghosh M, Mukhopadhyay N (1981) Consistency and asymptotic efficiency of two-stage and sequential procedures. Sankhya A 43:220–227
Ghosh M, Mukhopadhyay N, Sen PK (1997) Sequential estimation. Wiley, New York
Lai TL, Siegmund D (1977) A nonlinear renewal theory with applications to sequential analysis I. Ann Stat 5:946–954
Lai TL, Siegmund D (1979) A nonlinear renewal theory with applications to sequential analysis II. Ann Stat 7:60–76
Lehmann EL (1951) Notes on the theory of estimation. University of California, Berkeley
Mukhopadhyay N, Datta S, Chattopadhyay S (2004) Applied sequential methodologies, edited volume. Marcel Dekker, New York
Mukhopadhyay N, de Silva BM (2009) Sequential methods and their applications. CRC, New York
Mukhopadhyay N, Solanky TKS (1994) Multistage selection and ranking procedures: second-order asymptotics. Marcel Dekker, New York
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc A 231:289–337
Ray WD (1957) Sequential confidence intervals for the mean of a normal population with unknown variance. J R Stat Soc B 19:133–143
Robbins H (1959) Sequential estimation of the mean of a normal Population. In: Grenander U (ed) Probability and statistics (Harald Cramér Volume). Almquist and Wiksell, Uppsala, pp 235–245
Stein C (1945) A two sample test for a linear hypothesis whose power is independent of the variance. Ann Math Stat 16:243–258
Stein C (1949) Some problems in sequential estimation. Econometrica 17:77–78
Wald A (1947) Sequential analysis. Wiley, New York
Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. Ann Math Stat 19:326–339
Woodroofe M (1977) Second order approximations for sequential point and interval estimation. Ann Stat 5:984–995

Sex Ratio at Birth

Johan Fellman
Professor Emeritus
Folkhälsan Institute of Genetics, Helsinki, Finland

Sex Ratio in National Birth Registers

The sex ratio at birth, also called the secondary sex ratio, and here denoted SR, is usually defined as the number of males per 100 females. Among newborns there is almost always a slight excess of boys. Consequently, the SR is greater than 100, mainly around 106.

John Graunt (1620–1674) was the first person to compile data showing an excess of male births to female births and to note spatial and temporal variation in the SR. John Arbuthnot (1667–1735) demonstrated that the excess of males was statistically significant and asserted that the SR is uniform over time and space (Campbell 2001). Referring to christenings in London in the 82 years up to 1710, Arbuthnot suggested that the regularity in the SR and the dominance of males over females

could not be attributed to chance and must be an indication of divine providence. Nicholas Bernoulli's (1695–1726) counter-argument was that Arbuthnot's model was too restrictive. Instead of a fair coin model, the model should be based on an asymmetric coin. Based on the generalized model, chance could give uniform dominance of males over females. Later, Daniel Bernoulli (1700–1782), Pierre Simon de Laplace (1749–1827) and Siméon-Denis Poisson (1781–1840) also contributed to this discussion (David 1962; Hacking 1975).

Some general features of the SR can be noted. Stillbirth rates are usually higher among males than females, and the SR among stillborn infants is markedly higher than normal values, but the excess of males has decreased during the last decades. Hence, the SR among liveborn infants is slightly lower than among all births, but this difference is today very minute. Further, the SR among multiple maternities is lower than among singletons. In addition to these general findings, the SR shows marked regional and temporal variations.

In a long series of papers, attempts have been made to identify factors influencing the SR, but statistical analyses have shown that comparisons demand large data sets. Variations in the SR that have been reliably identified in family data have in general been slight and without notable influence on national birth registers. Attempts to identify reliable associations between SRs and stillbirth rates have been made, but no consistent results have emerged. Hawley (1959) stated that where prenatal losses are low, as in the high standard of living in Western countries, the SRs at birth are usually around 105 to 106. By contrast, in areas with a lower standard of living, where the frequencies of prenatal losses are relatively high, SRs are around 102. Visaria (1967) stressed that available data on late fetal mortality lend at best only weak support for these findings and concluded that racial differences seem to exist in the SR. He also discussed the perplexing finding that the SR among Koreans is high, around 113.

A common pattern observed in different countries is that during the first half of the twentieth century the SR showed increasing trends, but during the second half the trend decreased. Different studies have found marked peaks in the proportion of males during the First and Second World War. It has been questioned whether temporal or spatial variations of the SR are evident, and whether they constitute a essential health event. A common opinion is that secular increases are caused by improved socioeconomic conditions. The recent downward trends in the SRs have been attributed to new reproductive hazards, specifically exposure to environmental oestrogens. However, the turning point of the SR preceded the period of global industrialization and particularly the introduction of pesticides or hormonal drugs, rendering a causal association unlikely.

Sex Ratio in Family Data

In general, factors that affect the SR within families remain poorly understood. In a long series of papers, using family data, attempts have been made to identify factors influencing the SR. Increasing evidence confirms that exposure to chemicals, including pollutants from incinerators, dioxin, pesticides, alcohol, lead and other such workplace hazards, has produced children with reduced male proportion, Variables reported to be associated with an increase in the SR are large family size, high ancestral longevity, paternal baldness, excessive coffee-drinking, intensive coital frequency and some male reproductive tract disorders.

Some striking examples can be found in the literature of unisexual pedigrees extending over several generations. Slater (1943) stated that aberrant SRs tend, to some extent, to run in families. The finding by Lindsey and Altham (1998) that the probability of couples being only capable of having children of one sex is very low contradicts Slater's statement. The variation in the SR that has been reliably identified in family studies has invariable been slight compared with what we have observed in families with X-linked recessive retinoschisis (cleavage of retinal layers). We noted a marked excess of males within such families, in contrast to normal SRs in families with the X-linked recessive disorders haemophilia and color blindness (Eriksson et al. 1967; Fellman et al. 2002). However, with the exception of the X-linked recessive retinoschisis, no unequivocal examples exist of genes in man that affect the SR, and X-linked retinoschisis is universally very rare. Summing up, influential factors, although they have an effect on family data, have not been identified in large national birth registers.

About the Author

For biography *see* the entry ►Lorenz Curve.

Cross References

►Demography
►Sign Test
►Significance Tests, History and Logic of
►Statistics, History of

References and Further Reading

Campbell RB (2001) John Graunt, John Arbuthnott, and the human sex ratio. Hum Biol 73:605–610
David FN (1962) Games, gods and gambling. Charles Griffin, London

Eriksson AW, Vainio-Mattila B, Krause U, Fellman J, Forsius H (1967) Secondary sex ratio in families with X-chromosomal disorders. Hereditas 57:373–381
Fellman J, Eriksson AW, Forsius H (2002) Sex ratio and proportion of affected sons in sibships with X-chromosomal recessive traits: maximum likelihood estimation in truncated multinomial distributions. Hum Hered 53:173–180
Hacking I (1975) The emergence of probability. Cambridge University Press, Cambridge
Hawley AH (1959) Population composition. In: Hauser PM, Duncan OD (ed) The study of population: an inventory and appraisal. University of Chicago, Chicago, pp 361–382
Lindsey JK, Altham PME (1998) Analysis of the human sex ratio by using overdispersion models. Appl stat 47: 149–157
Slater E (1943) A demographic study of a psychopathic population. Ann Eugenic 12:121–137
Visaria PM (1967) Sex ratio at birth in territories with a relatively complete registration. Eugenic Quart 14:132–142

Sign Test

Peter Sprent
Emeritus Professor
University of Dundee, Dundee, UK

The sign test is a nonparametric test for hypotheses about a population median given a sample of observations from that population, or for testing for equality of medians, or for a prespecified constant median difference, given paired sample (i.e., matched pairs) values from two populations. These tests are analogues of the one-sample and matched pairs t-test for means in a parametric test such as the t-test.

The sign test is one of the simplest and oldest nonparametric tests. The name reflects the fact that each more detailed observation is effectively replaced by one of the signs plus (+) or minus (−). This was basically the test used by Arbuthnot (1710) to refute claims that births are equally likely to be male or female. Records in London showed that for each of 81 consecutive years an excess of male over female births. Calling such a difference a plus, Arbuthnot argued that if births were equally likely to be of either gender, then the probability of such an outcome was, $(0.5)^{81}$, or effectively zero.

Given a sample of n observations from any population which may be discrete or continuous and not necessarily symmetric, the test is used to test a hypothesis $H_0 : M = M_0$ where M is the population median. If H_0 holds the number of values less than M_0 will have a binomial distribution with parameters n and $p = 0.5$. The symmetry of the ►binomial distribution when $p = 0.5$ means the number of sample values greater than M_0 (a plus) may be used as an alternative equivalent statistic in a one or two-tail test.

Although not a commonly arising case, the test is still valid if each observation in a sample is from a different population providing each such population has the same median. For example, the populations may differ in ►variance or in ►skewness.

Among tests for location the sign test thus requires fewer assumptions for validity than any other well established test. The main disadvantage of the test is that it often has lower efficiency and lower power than tests that require stronger assumptions when those assumptions are valid. However, when the stronger assumptions are not valid the sign test may have greater power and efficiency. If the sample is from a normal distribution with known variance the asymptotic relative efficiency (ARE) of the sign test relative to the normal theory test is $2/\pi$. However if the sample is from a double exponential distribution the ARE of the sign test is twice that attained using the t-test.

For continuous data except in special cases like samples from a double exponential distribution the sign test is usually less efficient than some parametric test or nonparametric test that makes more use of information about the data. For example, the t-test is preferable for samples from a normal, or near normal, distribution and the ►Wilcoxon-signed-rank test performs better if an assumption of symmetry can be made.

Even when a sign test is less efficient than some other test it may prove economically beneficial if exact data of the type needed for that other test is expensive to collect but it is easy to determine whether such data, if it were available, would indicate a value less than or greater than an hypothesised median value M_0. For example, if in a manufacturing process rods produced should have a median diameter of 40 mm it may be difficult to measure diameters precisely, but easy to determine whether the diameter of each rod is less than 40 mm by attempting to pass it through a circular aperture of diameter 40 mm. Those that pass though have a diameter less than 40 mm (recorded as a minus); those that fail to pass through have a greater diameter (recorded as a plus). If diameters can be assumed to be normally distributed and a sample size of 30 is required to give the required power with a normal theory test when exact measurements are available, the ARE for a sign test (which gives a fairly good idea of the efficiency for a sample of this size) suggests that if we only have information on whether each item has diameter less than (or greater

than) 40 mm, then a sample of size $30 \times \pi/2 \approx 47$ should have similar power. An assumption here is that efficiency for smaller samples is close to the ARE, a result verified in some empirical studies. Thus if the cost of obtaining each exact measurement were twice that of determining only whether or not a diameter exceeded 40 mm there would be a clear cost saving in measuring simply whether diameters were more or less than 40 mm for a sample of 47 compared to that for taking exact measurements for a sample of 30.

Sample values exactly equal to M_0 are usually ignored when using the test and the sample size used in assessing significance is reduced by 1 for each such value.

In the case of matched pair samples from distributions that may be assumed to differ if at all only in their medians, the test may be applied using the signs of the paired differences to test if the difference is consistent with a zero median and by a slight modification to test the hypothesis that the median difference has some specified value θ_0. The test is available in most standard statistical software packages or may be conducted using tables for the binomial distribution when $p = 0.5$ and the relevant n (sample size). For continuous data one may determine confidence intervals based on this test with the aid of such tables. Details are given in most textbooks covering basic nonparametric methods such as Gibbons and Chakraborti (2004) or Sprent and Smeeton (2007).

An interesting case that leads to a test equivalent to the sign test with heavy tying was proposed by McNemar (1947) and is usually referred to as McNemar's test. This test is relevant where observations are made to test if there are nonneutralizing changes in attitudes of individuals before or after exposure to a treatment or stimulus. For example, a group of 200 motorists may be asked whether or not they think the legal maximum permissible level of blood alcohol for drivers should be lowered. The numbers answering *yes* or *no* are recorded. The group are then shown a video illustrating the seriousness of accidents where drivers have exceeded the legal limit. Their answers to the same question about lowering the level are now recorded and tabulated as shown in this table:

		Before video	Lower limit
		Yes	*No*
After video Lower limit	*Yes*	160	24
	No	11	5

If we denote a change from *No* before the video to *Yes* after the video by a plus there are 24 plus, and a change from *Yes* before to *No* afterwards there are 11 minus. Thus, although the video seems to have influenced some changes of opinion in both directions more (24) who did not support a reduction before seeing the video appear to have been persuaded to support a reduction after seeing the video, whereas 11 have switched opinions in the opposite direction, opposing a ban after seeing the video although they supported one before seeing the video.

A sign test may be applied on the basis of 24 plus and 11 minus being observed in an effective sample of size 35. The diagonal values of 160 and 5 represent "ties" in the sense that they represent drivers who are not influenced by the video and so are ignored.

About the Author

Dr. Peter Sprent is Emeritus Professor of Statistics and a former Head of the Mathematics Department at the University of Dundee, Scotland. Previously he worked as a consultant statistician at a horticultural research station in England. He has 28 years teaching experience in Australia and the United Kingdom. He has written or coauthored 12 books on statistics and related topics, the best known of which is *Applied Nonparametric Statistical Methods* (with Nigel C. Smeeton, Chapman and Hall/CRC; 4th edition, 2007). He has been on the editorial boards, or been an associate editor, of several leading statistical journals and served on the Council and various committees of the Royal Statistical Society. He is a Fellow of the Royal Society of Edinburgh and is an Elected member of the International Statistical Institute.

Cross References

►Asymptotic Relative Efficiency in Testing
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Sex Ratio at Birth
►Wilcoxon-Signed-Rank Test

References and Further Reading

Arbuthnot J (1710) An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. Philos Trans R Soc 27:186–190

Gibbons JD, Chakraborti S (2004) Nonparametric statistical inference, 4th edn. Marcel Dekker, New York

McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrica 12:153–157

Sprent P, Smeeton NC (2007) Applied nonparametric statistical methods, 4th edn. Chapman & Hall/CRC Press, Boca Raton

Significance Testing: An Overview

Elena Kulinskaya[1], Stephan Morgenthaler[2], Robert G. Staudte[3]
[1]Professor, Aviva Chair in Statistics
University of East Anglia, Norwich, UK
[2]Professor, Chair of Applied Statistics
Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[3]Professor and Head of Department of Mathematics and Statistics
La Trobe University, Melbourne, VIC, Australia

Introduction

A *significance test* is a statistical procedure for testing a hypothesis based on experimental or observational data. Let, for example, $\overline{X}_1$ and $\overline{X}_2$ be the average scores obtained in two groups of randomly selected subjects and let μ_1 and μ_2 denote the corresponding population averages. The observed averages can be used to test the null hypothesis $\mu_1 = \mu_2$, which expresses the idea that both populations have equal average scores. A *significant result* occurs if $\overline{X}_1$ and $\overline{X}_2$ are very different from each other, because this contradicts or falsifies the null hypothesis. If the two group averages are similar to each other, the null hypothesis is not contradicted by the data. What exact values of the difference $\overline{X}_1 - \overline{X}_2$ of the group averages are judged as significant depends on various elements. The variation of the scores between the subjects, for example, must be taken into account. This variation creates uncertainty and is the reason why the testing of hypotheses is not a trivial matter. Because of the uncertainty in the outcome of the experiment, it is possible that a seemingly significant result is obtained, even though the null hypothesis is true. Conversely, the null hypothesis being false does not mean that the experiment will necessarily result in a significant result.

The significance of a test is usually measured in terms of a tail-error probability of the null distribution of a test statistic. In the above example, assume the groups are normally distributed with common known variance σ^2. The Z-test statistic is $Z = (\overline{X}_1 - \overline{X}_2)/\mathrm{SE}[\overline{X}_1 - \overline{X}_2]$, where $\mathrm{SE}[\overline{X}_1 - \overline{X}_2] = \sigma^2\{1/n_1 + 1/n_2\}$ is the standard error of the difference. Here n_1, n_2 are the respective sample sizes for the two groups. Under the null hypothesis, Z has the standard normal distribution with cumulative distribution $P(Z \le z) = \Phi(z)$. A large observed value $Z = Z_{obs}$ corresponds to a small tail area probability $P(Z \ge Z_{obs}) = \Phi(-Z_{obs})$.The smaller this probability the more the evidence against the null in the direction of the alternative $\mu_1 > \mu_2$. For a two-sided alternative $\mu_1 \ne \mu_2$, a test statistic is $|Z|$ and the evidence against the null is measured by the smallness of $P(|Z| \ge |Z_{obs}|) = 2\Phi(-|Z_{obs}|)$. These tail-error probabilities are examples of p-values for one- and two-sided tests.

To carry out a significance test then one needs, first, a *statistic* $S(X)$ (real function of the data X) that orders the outcomes X of a study so that larger values of $S(X)$ cast more doubt on the null hypothesis than smaller ones; and second, the probability distribution P_0 of $S(X)$ when the null hypothesis is true. One may be interested in simply assessing the evidence in the value obtained for the statistic S in an experiment, the Fisherian approach, or in making a decision to reject the null hypothesis in favor of an alternative hypothesis, the Neyman–Pearson approach.

Significance Tests for Assessing Evidence

By far the most prevalent concept for assessing evidence in S is the p-value, promoted by the influential scientist R.A. Fisher through his many articles and books, see the collection Fisher (1990).

The p-Value

Having observed data $X = x$, and hence $S(x) = S_{\mathrm{obs}}$, the *p-value* is defined by $p = P_0(S \ge S_{\mathrm{obs}})$. It is the probability of obtaining as much or more evidence against the null hypothesis as just observed with S_{obs}, assuming the null hypothesis is true. The p-value is decreasing with increasing S_{obs}, which means that smaller ▶p-values are indicative of a more significant result. Fisher (1973, pp. 80, 82, and 122), offered some rough guidelines for interpreting the strength of evidence measured by the p-value, based on his experience with agricultural experiments. He suggested that a p-value larger than 0.1 was not small enough to be significant, a p-value as small as 0.05 could seldom be disregarded, and a p-value less then 0.01 was clearly significant. Thus according to Fisher "significance testing" is the conducting of an experiment that will give the data a chance to provide evidence S_{obs} against the null hypothesis. Very small values of the p-value correspond to significant evidence, where "significant" is somewhat arbitrarily defined. It is a matter of history that Fisher's rough guideline "a value as small as 0.05 could seldom be disregarded" became a *de facto* necessity for publication of experimental results in many scientific fields. However, despite its usefulness for filtering out many inconsequential results, the p-value is often confused with fixed significance levels (see section ▶"Significance Tests for Making Decisions").

Finding the Null Distribution

It is not always easy to find the null distribution of a test statistic. It must be chosen carefully. For example, in the Z-test example of section ▶"Introduction", three assumptions were made, normality of the observations, equality of the group variances and knowledge of the common variance σ^2. If the first two assumptions hold, but the latter is relaxed to $\sigma^2 > 0$, then the distribution of the Z-test statistic depends on the unknown *nuisance parameter* σ^2, so one does not have a unique null distribution. An appropriate test statistic is the *two-sample pooled t-statistic*, which is just the Z-test statistic with σ replaced by s_{pooled}, where $s^2_{pooled} = \{(n_1-1)s_1^2 + (n_2-1)s_2^2\}/(n_1+n_1-2)$, and s_1^2, s_2^2 are the the respective sample variances. This t statistic has, under the null $\mu_1 = \mu_2$ a Student-t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom, which allows for computation of p-values.

If the assumption of normality of the groups is retained, but their variances are not assumed equal, then one can estimate them separately using the respective sample variances. An approximating t distribution for the resulting standardized mean difference is known as the *Welch t-test* see Welch (1938). If the assumption of normality is relaxed to a continuous distribution then a comparison can be based on the sum S of the ranks of one sample within the ranking of the combined sets of observations. The null hypothesis is that each group has the same continuous distribution F and then S has a unique distribution. This test is known as the ▶*Wilcoxon–Mann–Whitney test*. It is an example of a *distribution-free test*, because F is unspecified.

Another way of computing a p-value when the null hypothesis distribution is not uniquely specified is to sample repeatedly from the empirical distribution of the data and for each sample compute the value of the test statistic; the proportion of values greater than the original S_{obs} is a *bootstrap estimate* of the p-value.

Significance Tests for Making Decisions

Neyman and Pearson (1928), Neyman (1933) formulated the significance testing problem as one of decision making. The data X are assumed to have distributions P_θ indexed by the parameter θ known to lie in one of two mutually exclusive sets Θ_0, Θ_1, and one must choose between them, using only X. The parameter sets Θ_0 and Θ_1 are called the *null* and *alternative hypotheses*, respectively. Each may be *simple*, containing only a single value, or *composite*. If $X \sim P_\theta$ for some $\theta \in \Theta_0$, and one chooses Θ_1 a *Type I error*, (or, error of the first kind), is committed. If $X \sim P_\theta$ for some $\theta \in \Theta_1$, and one chooses Θ_0 a *Type II error*, (or, error of the second kind), is committed. Because the consequences of Type I and Type II errors are often incommensurate, see Neyman (1950), the Neyman–Pearson framework places a bound α on Type I error probabilities, called the *level* of the test, and subject to this constraint seeks a decision rule that in some sense minimizes the Type II error probabilities, $\beta(\theta_1)$ for $\theta_1 \in \Theta_1$.

A *decision rule* equals 1 or 0 depending on whether Θ_1 or Θ_0 is chosen, after observing $X = x$. It is by definition the indicator function $I_C(x)$ of the *critical region* C, which is the set of values of X for which Θ_1 is chosen. This region is critical in the sense that if $X \in C$, one rejects the null hypothesis and risks making a Type I error. The *size* of a critical region is $\sup_{\theta\in\Theta_0} P_\theta(X \in C)$. One seeks a critical region (test) for which the size is no greater than the level α and which has large power of detecting alternatives. The size may be set equal to the desired level α by choice of C when the distributions P_θ are continuous, but in the case of discrete P_θ, the size will often be less than α, unless some form of ▶randomization is employed, see Lehmann (1986).

Power Function of a Test and Optimal Test Statistics

The *power* of a test for detecting an alternative $\theta_1 \in \Theta_1$ is defined by $\Pi(\theta_1) = P_{\theta_1}(X \in C) = 1 - \beta(\theta_1)$. It is the probability of making the right decision (rejecting Θ_0) when $\theta_1 \in \Theta_1$; and as indicated, it is also 1 minus the probability of making a Type II error for this θ_1. The *power function* is defined by $\Pi(\theta_1)$, for each $\theta_1 \in \Theta_1$. Let f_θ be the density of P_θ with respect to a dominating measure for the distributions of X. Neyman and Pearson showed that for a simple hypothesis θ_0 and simple alternative θ_1, there exists a most powerful level-α test which rejects the null when the *likelihood ratio* $\lambda(x) = f_{\theta_1}(x)/f_{\theta_0}(x)$ is large. That is, the critical region is of the form $C = \{x : \lambda(x) \geq c\}$, where the *critical value* c defining the boundary of the critical region is chosen so $P_{\theta_0}\{\lambda(X) \geq c\} = \alpha$. For composite hypotheses, the *likelihood test statistic* defined by $\lambda(x) = \sup_\Theta f_\theta(x)/\sup_{\Theta_0} f_\theta(x)$ is the basis for many tests, because its large sample distribution is known. A *uniformly most powerful* level-α test maximizes the power for each value of the alternative amongst all level-α tests. Uniformly most powerful tests for composite alternatives are desirable, but such tests do not usually exist. See Lehmann (1986) for a comprehensive development of the theory of hypothesis testing.

Inversion of a Family of Tests to Obtain Confidence Regions

A *confidence region* of level $1-\alpha$ for a parameter θ is a random set $R(X)$ for which $P_\theta\{\theta \in R(X)\} \geq 1-\alpha$ for all $\theta \in \Theta$. When Θ is a subset of the real line, the region is

usually in the form of a random *confidence interval* $[L, U]$, where $L = L(X), U = U(X)$. The inversion procedure, due to Neyman (1935), supposes that for each $\theta_0 \in \Theta$ there is a level–α test with critical region $C_\alpha(\theta_0)$ for testing the simple null hypothesis $\Theta_0 = \{\theta_0\}$ against its complement $\Theta_0^c = \{\theta \in \Theta : \theta \neq \theta_0\}$. This family of tests can be converted into a level $1-\alpha$ confidence region for θ, given by $R(X) = \{\theta_0 \in \Theta : X \notin C_\alpha(\theta_0)\}$. Thus a parameter θ_0 belongs to the confidence region if and only if it is not rejected by the level α test of $\theta = \theta_0$ against $\theta \neq \theta_0$.

On p-Values and Fixed Significance Levels

The purpose of choosing a fixed level α as a prior upper bound on the probability of Type I errors is to avoid making decisions that are influenced by the observed data x. The p-value, on the other hand, requires knowledge of x for its computation, and subsequent interpretation as evidence against the null hypothesis. Thus when used for the separate purposes for which they were designed, there is no confusion. However, having observed $S(x) = S_{\text{obs}}$, the p-value is equal to the level α for which $S_{\text{obs}} = c_\alpha$; that is, the smallest fixed level for which the test rejects the null. For this reason, it is sometimes called the *observed significance level*. One rejects the null at level α if and only if the p-value $\leq \alpha$. It is widespread practice to use the Neyman–Pearson framework to obtain a powerful test of level $\alpha = 0.05$, and then to report the p-value. Thus there has evolved in practice a combination of concepts that can prove confusing to the uninitiated, see Berger (2003) Hubbard and Bayarri (2003) and Lehmann (1993).

Bayesian Hypothesis Testing

The Bayesian framework for significance testing assumes a *prior* probability measure $\pi(\theta)$ over the parameter space $\Theta = \Theta_0 \cup \Theta_1$. This yields prior probabilities $\pi_0 = \pi(\Theta_0)$, $1 - \pi_0$ on the null and alternative hypotheses Θ_0, Θ_1, respectively, and the *prior odds* $\pi_0/(1-\pi_0)$ in favor of the null. It is further assumed that for each θ, the data X has a conditional distribution $f(x|\theta)$ for X, given θ. The *posterior probability of the null* is then $P(\Theta_0|x) = \int_{\Theta_0} f(x|\theta)d\pi(\theta)/f_X(x)$, where $f_X(x) = \int_\Theta f(x|\theta)d\pi(\theta)$. One can, if a decision is required, reject the null in favor of the alternative when $P(\Theta_0|x)$ is less than some preassigned level, as in NP testing; or, one can simply choose to interpret it as a measure of support for Θ_0.

Bayes Factor

It turns out that the posterior odds for Θ_0 are related to its prior odds by $P(\Theta_0|x)/(1-P(\Theta_0|x)) = B_{01}(x)\,\pi_0/(1-\pi_0)$, The *Bayes factor* $B_{01}(x) = f_{\Theta_0}(x)/f_{\Theta_1}(x)$, where $f_{\Theta_i}(x) = \int_{\Theta_i} f(x|\theta)d\pi(\theta)/\pi(\Theta_i)$, $i = 0,1$. The Bayes factor measures the change in odds for the null hypothesis Θ_0 after observation of $X = x$. It is also often interpreted as a measure of support for Θ_0, but this interpretation is not without controversy; for further discussion see Kass (1995) and Lavine and Schervish (1999).

Significance Tests for Special Purposes

When one wants to adopt the model $X \sim \{P_\theta : \theta \in \Theta\}$ for inference, be it testing or estimation, a *goodness-of-fit test* rejects the entire model if a suitable test statistic $S(X)$ has small p-value. Thus if the data do not cast doubt on the model, the statistician happily proceeds to adopt it. This procedure is informal in that many other models might equally pass such a test, but are not considered. Tests for submodel selection in regression have the same feature; one "backs into" acceptance of a submodel because an F-test does not reject it. All such significance tests are simply informal guides to ►model selection, with little regard for Type II errors, or the subsequent effects on inference with the chosen model. *Equivalence tests*, on the other hand, place great emphasis on formal testing, and do provide evidence for a null hypothesis of no effect. They do this by interchanging the traditional roles of null and alternative hypotheses. For example, if θ represents the mean difference in effects of two drugs, one might be interested in evidence for $|\theta| \leq \theta_0$, where θ_0 defines a region of "equivalence." This is taken as the alternative hypothesis, to a null $|\theta| \geq \theta_1$, where $\theta_1 > \theta_0$ is large, say. One also simultaneously tests the null $\theta \leq -\theta_1$ against the alternative of equivalence. If one rejects both these null hypotheses in favor of the alternative, evidence for equivalence is found. See Wellek (2003) for a complete development.

Final Remarks and Additional Literature

Statistical significance of a test, meaning a null hypothesis is rejected at a pre-specified level such as 0.05, is not evidence for a result which has practical or scientific significance. This has led many practitioners to move away from the simple reporting of p-values to reporting of confidence intervals for effects; see Krantz (1999) for example. A measure of *evidence for* a positive effect that leads to confidence intervals for effects is developed in Kulinskaya et al. (2008). Fuzzy hypothesis tests and confidence intervals are introduced in Dollinger et al. (1996) and explored in Geyer and Meeden (2006).

About the Author

For biography *see* the entry ►Meta-Analysis.

Cross References

►Accelerated Lifetime Testing
►Anderson-Darling Tests of Goodness-of-Fit
►Bartlett's Test

►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Test: Analysis of Contingency Tables
►Chi-Square Tests
►Dickey-Fuller Tests
►Durbin–Watson Test
►Effect Size
►Equivalence Testing
►Equivalence Testing
►Fisher Exact Test
►Frequentist Hypothesis Testing: A Defense
►Full Bayesian Significant Test (FBST)
►Jarque-Bera Test
►Kolmogorov-Smirnov Test
►Mood Test
►Most Powerful Test
►Multiple Comparisons Testing from a Bayesian Perspective
►Neyman-Pearson Lemma
►Nonparametric Rank Tests
►Null-Hypothesis Significance Testing: Misconceptions
►Omnibus Test for Departures from Normality
►Parametric Versus Nonparametric Tests
►Permutation Tests
►Presentation of Statistical Testimony
►Psychological Testing Theory
►Psychology, Statistics in
►P-Values
►Randomization Tests
►Rank Transformations
►Scales of Measurement and Choice of Statistical Methods
►Sequential Probability Ratio Test
►Sign Test
►Significance Tests, History and Logic of
►Significance Tests: A Critique
►Simes' Test in Multiple Testing
►Statistical Evidence
►Statistical Inference
►Statistical Inference: An Overview
►Statistical Significance
►Step-Stress Accelerated Life Tests
►Student's *t*-Tests
►Testing Exponentiality of Distribution
►Testing Variance Components in Mixed Linear Models
►Tests for Discriminating Separate or Non-Nested Models
►Tests for Homogeneity of Variance
►Tests of Fit Based on The Empirical Distribution Function
►Tests of Independence
►Wilcoxon–Mann–Whitney Test
►Wilcoxon-Signed-Rank Test

References and Further Reading

Berger JO (2003) Could Fisher, Jeffreys and Neyman have agreed on testing? Stat Sci 18(1):1–32, With discussion

Dollinger MB, Kulinskaya E, Staudte RG (1996) Fuzzy hypothesis tests and confidence intervals. In: Dowe DL, Korb KB, Oliver JJ (eds) Information, statistics and induction in science. World Scientific, Singapore, pp 119–128

Fisher RA (1990) Statistical methods, experimental design and scientific inference. Oxford University Press, Oxford. Reprints of Fisher's main books, first printed in 1925, 1935 and 1956, respectively. The 14th edition of the first book was printed in 1973

Geyer C, Meeden G (2006) Fuzzy confidence intervals and p-values (with discussion). Stat Sci 20:258–387

Hubbard R, Bayarri MJ (2003) Confusion over measures of evidence (*p*'s) versus errors (*a*'s) in classical statistical testing. Am Stat 57(3):171–182, with discussion

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90: 773–795

Krantz D (1999) The null hypothesis testing controversy in psychology. J Am Stat Assoc 94(448):1372–1381

Kulinskaya E, Morgenthaler S, Staudte RG (2008) Meta analysis: a guide to calibrating and combining statistical evidence. Wiley, Chichester, www.wiley.com/go/meta_analysis

Lavine M, Schervish M (1999) Bayes factors: what they are and what they are not. Am Stat 53:119–122

Lehmann EL (1986) Testing statistical hypotheses, 2nd edn. Wiley, New York

Lehmann EL (1993) The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? J Am Stat Assoc 88:1242–1249

Neyman J (1935) On the problem of confidence intervals. Ann Math Stat 6:111–116

Neyman J (1950) First course in probability and statistics. Henry Holt, New York

Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20A:175–240 and 263–294

Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc A 231:289–337

Welch BL (1938) The significance of the difference between two means when the variances are unequal. Biometrika 29:350–361

Wellek S (2003) Testing statistical hypotheses of equivalence. Chapman & Hall/CRC Press, New York

Significance Tests, History and Logic of

Henrik Olsson, Mirta Galesic
Max Planck Institute for Human Development, Berlin, Germany

By most accounts, the first significance test was published in 1710 by the Scottish mathematician, physician, and author John Arbuthnot. He believed that, because males were subject to more external accidents than females, they

enjoyed an advantage of a higher birthrate. Arbuthnot calculated the expectation, or the probability, of the data from 82 years of birth records in London given a chance hypothesis of equal birthrates for both sexes. Because this expectation was very low he concluded "that it is Art, not Chance, that governs" (p. 189), and that this result constituted a proof of existence of an active god. Although he never used the terms *significance* or *significant* – these terms were first used at the end of the nineteenth century by Francis Ysidro Edgeworth (1885) and John Venn (1888) – his argument is strikingly similar to the logic underlying modern null hypothesis testing as implemented in Ronald Fisher's significance testing approach (e.g., 1925, 1935).

The beginning of the twentieth century saw the development of the first modern significance tests: Karl Pearson's (1900) *chi-squared test* and William Sealy Gosset's (or Student's 1908) *t-test* (although the term t-test appeared only later, in 1932 in the fourth edition of Fisher's *Statistical Methods for Research Workers*). Both are examples of tail-area significance tests, in which a hypothesis is rejected if the tail of the null distribution beyond the observed value is less than a prescribed small number. Gosset's article was also the beginning of the field of small sample statistics, where the earlier asymptotics ($n \to \infty$) were replaced by exact probabilities.

The use of significance tests really took root among applied researchers after the publication of Fisher's influential books, *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935). Fisher rejected the (older) methods of inverse probability (of hypothesis given data) and proposed a method of *inductive inference*, a formal way of getting from data to hypothesis. His approach can be summarized as follows: The researcher sets up a null hypothesis that a sample statistic comes from a hypothetical infinite population with a known sampling distribution. The null hypothesis is rejected or, as Fisher called it, "disproved," if the sample statistic deviates from the mean of the sampling distribution by more than a specified criterion. This criterion – or *level of significance* – is typically set to 5%, although Fisher later recommended reporting the exact probability. In this approach, no claims about the validity of alternative hypotheses are possible. It is nevertheless tempting to view the complement of the null hypothesis as an alternative hypothesis and argue, as Arbouthnot did, that the rejection of the null hypothesis gives credit to an unspecified alternative hypothesis. Fisher's approach is also associated with an epistemic interpretation of significance: A Fisherian *p-value* is thought to measure the strength of evidence against the null hypothesis and to allow the researcher to learn about the truth or falsehood of a specific hypothesis from a single experiment.

The major rival to Fisher's approach was Jerzy Neyman and Egon Pearson's (1928a, 1928b, 1933) approach to hypothesis testing, originally viewed as an extension and improvement of Fisher's ideas. Neyman and Pearson rejected the idea of *inductive inference* and replaced it with the concept of *inductive behavior*. They sought to establish rules for making decisions between different hypotheses regardless of researcher's beliefs about the truth of those hypotheses. They argued for specifying both a null hypothesis and an alternative hypothesis, which allows for the calculation of two error probabilities, *Type I* error and *Type 2* error, based on considerations regarding decision criteria, sample size and effect size. *Type I* error occurs when the null hypothesis is rejected although it is true. The probability of a *Type I* error is called α. *Type II* error occurs when the alternative hypothesis is rejected although it is true. The probability of a *Type II* error is called β and 1-β is called the *power* of the test or the long run frequency of accepting the alternative hypothesis if it is true. The decision to accept or reject hypotheses in the Neyman–Pearson approach depends on the costs associated with *Type I* and *Type II* errors. The cost considerations lie outside of the formal statistical theory and must be based on context-dependent pragmatic personal judgment. The goal, then, for a researcher is to design an experiment that controls for α and β and use a test that minimizes β given a bound on α. In contrast to the data dependent ►*p-values* in Fisher's approach, α is specified before collecting the data. Despite the different conceptual foundations of Fisher's approach and Neyman-Pearson's approach, classical statistical inference, as commonly presented, is essentially an incoherent hybrid of the two approaches (Hubbard and Bayarri 2003; Gigerenzer 1993), although there exist attempts to reconcile them (Lehmann 1993). There is a considerable literature discussing the pros and cons of classical statistical inference, especially null hypothesis significance testing in the Fisherian tradition (e.g., Berger and Wolpert 1988; Royall 1997; Morrison and Henkel 1970). The major alternative to classical significance and hypothesis testing is Bayesian hypothesis testing (Jeffreys 1961; Kass and Raftery 1995).

Cross References

►Effect Size
►Frequentist Hypothesis Testing: A Defense
►Significance Testing: An Overview
►Statistical Significance

References and Further Reading

Arbuthnot J (1710) An argument for divine providence, taken from the constant regularity observed in the births of both sexes. Philos Tr R Soc 27:186–190

Berger JO, Wolpert RL (1988) The likelihood principle, 2nd edn. Institute of Mathematical Statistics, Hayward, CA
Edgeworth FY (1885) Methods of statistics. Jubilee Volume of the Statistical Society. E. Stanford, London, pp 181–217
Fisher RA (1925) Statistical methods for research workers. Oliver and Boyd, Edinburgh
Fisher RA (1935) The design of experiments. Oliver and Boyd, Edinburgh
Gigerenzer G (1993) The Superego, the Ego, and the Id in statistical reasoning. In: Keren G, Lewis C (eds) A handbook for data analysis in the behavioral sciences: methodological issues. Erlbaum, Hillsdale, NJ, pp 311–339
Hubbard R, Bayarri M-J (2003) Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. Am Stat 57:171–182
Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, Oxford
Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:377–395
Lehmann EL (1993) The Fisher, Neyman–Pearson theories of testing hypotheses: one theory or Two? J Am Stat Assoc 88:1242–1249
Morrison DE, Henkel RE (eds) (1970) The significance test controversy: a reader. Aldine, Chicago
Neyman J, Pearson ES (1928a) On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. Biometrika 20A:175–240
Neyman J, Pearson ES (1928b) On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. Biometrika 20A:263–294
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Tr R Soc S-A 231:289–337
Royall RM (1997) Statistical evidence: a likelihood paradigm. Chapman and Hall, London
Venn J (1888) The logic of chance: an essay on the foundations and province of the theory of probability, 3rd edn. Macmillan, London

Significance Tests: A Critique

Bruno Lecoutre
ERIS, Laboratoire de Mathématiques Raphaël Salem, C.N.R.S. and Université de Rouen, Mont Saint Aignan, France

> It is very bad practice to summarise an important investigation solely by a value of P.
>
> (Cox 1982, p 327)

In spite of some recent changes, significance tests are again conventionally used in most scientific experimental publications. According to this publication practice, each experimental result is dichotomized: significant vs. non-significant. But scientists cannot in this way find appropriate answers to their precise questions, especially in terms of effect size evaluation. It is not surprising that, from the outset (e.g., Boring 1919), significance tests have been subject to intense criticism. Their use has been explicitly denounced by the most eminent and most experienced scientists, both on theoretical and methodological grounds, not to mention the sharp controversies on the very foundations of statistical inference that opposed Fisher to Neyman and Pearson, and continue to oppose frequentists to Bayesians. In the 1960s there was more and more criticism, especially in the behavioral and social sciences, denouncing the shortcomings of significance tests: *the significance test controversy* (Morrison and Henkel 1970).

Significance Test Are Not a Good Scientific Practice

> It is foolish to ask 'Are the effects of A and B different?' They are always different - for some decimal place.
>
> (Tukey 1991, p 100)

In most applications, no one can seriously believe that the different treatments have produced no effect: the point null hypothesis is only a *straw man* and a significant result is an evidence against an hypothesis known to be false before the data are collected, but not an evidence in favor of the alternative hypothesis. It is certainly not a good scientific practice, where one is expected to present arguments that support the hypothesis in which one is really interested. The real problem is to obtain estimates of the sizes of the differences.

The innumerable misuses of significance tests

> The psychological literature is filled with misinterpretations of the nature of the tests of significance.
>
> (Bakan 1967, in Morrison and Henkel 1970, p 239)

Due to their inadequacy in experimental data analysis, the practice of significance tests entails considerable distortions in the designing and monitoring of experiments. It leads to innumerable misuses in the selection and interpretation of results. The consequence is the existence of publication biases denounced by many authors: while non-significant results are – theoretically – only statements of ignorance, only the significant results would really deserve publication.

The evidence of distortions is the use of the symbols *NS*, *, **, and *** in scientific journals, as if the degree of significance was correlated with the meaningfulness of research results. Many researchers and journal editors appear to be "star worshippers": see Guttman (1983), who openly attacked the fact that some scientific journals, and *Science* in particular, consider the significance test as a criterion of scientificness. A consequence of this overeliance

on significant effects is that most users of statistics overestimate the probability of replicating a significant result (Lecoutre et al. 2010).

The Considerable Difficulties Due to the Frequentist Approach

> What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure.
>
> (Jeffreys 1998/1939, Sect. 7.2)

Since the *p*-value is the proportion of samples that are "at least as extreme" as the observed data (under the null hypothesis), the rejection of the null hypothesis is based on the probability of the samples that *have not been observed*, what Jeffreys ironically expressed in the above terms. This mysterious and unrealistic use of the sampling distribution for justifying null hypothesis significance tests is for the least highly counterintuitive. This is revealed by questions frequently asked by students and statistical users: "why one considers the probability of samples outcomes that are more extreme than the one observed?"

Actually, due to their frequentist conception, significance tests involve considerable difficulties in practice. In particular, many statistical users misinterpret the *p*-values as inverse (Bayesian) probabilities: $1 - p$ is "the probability that the alternative hypothesis is true." All the attempts to rectify this misinterpretation have been a loosing battle.

Significance Tests Users' Dissatisfaction

> Neither Fisher's null hypothesis testing nor Neyman-Pearson decision theory can answer most scientific problems.
>
> (Gigerenzer 2004, p 599)

Several empirical studies emphasized the widespread existence of common misinterpretations of significance tests among students and scientists (for a review, see Lecoutre et al. 2001). Many methodology instructors who teach statistics, including professors who work in the area of statistics, appear to share their students' misinterpretations. Moreover, even professional applied statisticians are not immune to misinterpretations of significance tests, especially if the test is nonsignificant. It is hard to interpret these finding as an individual's lack of mastery: they reveal that significance test do not address the questions that are of primary interest for the scientific research.

In particular, the dichotomous significant/non significant outcome of significance tests strongly suggests binary research decisions: "reject/accept the null hypothesis." "But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested" (Rozeboom, in Morrison and Henkel 1970, p. 221). The "reject/accept" attitude is obviously a poor and unfortunate decision practice.

- A statistically significant test provides no information about the departure from the null hypothesis. When the sample is large a descriptively small departure may be significant.
- A nonsignificant test is not evidence favoring the null hypothesis. In particular, a descriptively large departure from the null hypothesis may be nonsignificant if the experiment is insufficiently sensitive.

In fact, in order to interpret their data in a reasonable way, users must resort to a more or less naive mixture of significance tests outcomes and other information. But this is not an easy task! This leads users to make *adaptative distortions*, designed to make an ill-suited tool fit their true needs. Actually, many users explicitly appear to have a real consciousness of the stranglehold of significance tests: in many cases they use them only because they know no other alternative.

Concluding Remarks

> Inevitably, students (and essentially everyone else) give an inverse or Bayesian twist to frequentist measures such as confidence intervals and P values.
>
> (Berry 1997, p 241)

It is not acceptable that statistical inference methods users will continue using nonappropriate procedures because they know no other alternative. Nowadays, proposals for changes in reporting experimental results are constantly made. In all fields these changes, especially in presenting and interpreting effect sizes, are more and more enforced within editorial policies. Unfortunately, academic debates continue and give a discouraging feeling of *déjà-vu*. Rather than stimulating the interest of experimental scientists, this endless controversy is without doubt detrimental to the impact of new proposals, if not to the image of statistical inference.

The majority official trend is to advocate the use of confidence intervals, in addition to or instead of significance tests. However, reporting confidence intervals appears to have very little impact on the way the authors interpret their data. Most of them continue to focus on the statistical significance of the results. They only wonder whether the

interval includes the null hypothesis value, rather than on the full implications of confidence intervals: the steamroller of significance tests cannot be escaped.

Furthermore, for many reasons due to their frequentist conception, confidence intervals can hardly be seen as the ultimate method. We then naturally have to ask ourselves whether the "Bayesian choice" will not, sooner or later, be unavoidable. It can be argued that an *objective Bayes theory* is by no means a speculative viewpoint but on the contrary is perfectly feasible (Rouanet et al. 2000; Lecoutre et al. 2001; Lecoutre 2008).

Cross References

►Frequentist Hypothesis Testing: A Defense
►Null-Hypothesis Significance Testing: Misconceptions
►Presentation of Statistical Testimony
►Psychology, Statistics in
►P-Values
►Significance Testing: An Overview
►Significance Tests, History and Logic of
►Statistical Evidence
►Statistical Inference: An Overview
►Statistical Significance

References and Further Reading

Berry DA (1997) Teaching elementary Bayesian statistics with real applications in science. Am Stat 51:241–246
Boring EG (1919) Mathematical versus scientific significance. Psychol Bull 16:335–338
Cox DR (1982) Statistical significance tests. Br J Clin Pharmacol 16:325–331
Gigerenzer G (2004) Mindless statistics. J Socio-Economics 33:587–606
Guttman L (1983) What is not what in statistics? Statistician 26:81–107
Jeffreys H (1961) Theory of probability, 3rd edn (1st edn: 1939). Clarendon, Oxford
Lecoutre B (2008) Bayesian methods for experimental data analysis. In: Rao CR, Miller J, Rao DC (eds) Handbook of statistics: epidemiology and medical statistics, vol 27. Elsevier, Amsterdam, pp 775–812
Lecoutre B, Lecoutre M-P, Poitevineau J (2001) Uses, abuses and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? Int Stat Rev 69:399–418
Lecoutre B, Lecoutre M-P, Poitevineau J (2010) Killeen's probability of replication and predictive probabilities: How to compute, use and interpret them. Psychol Methods 15:158–171
Morrison DE, Henkel RE (eds) (1970) The Significance test controversy - a reader. Butterworths, London
Rouanet H, Bernard J-M, Bert M-C, Lecoutre B, Lecoutre M-P, Le Roux B (2000) New ways in statistical methodology: from significance tests to Bayesian inference, 2nd edn. Peter Lang, Bern, CH
Tukey JW (1991) The philosophy of multiple comparisons. Stat Sci 6:100–116

Simes' Test in Multiple Testing

Sanat K. Sarkar
Professor
Temple University, Philadelphia, PA, USA

Over the past decade there has been a revival of interest in the field of multiple testing due to its increased relevance in modern scientific investigations, such as DNA microarray and functional magnetic resonance imaging (fMRI) studies. Simes' (1986) test plays an important role in the developments of a number of multiple testing methods. Given a family of null hypotheses $H_1, \ldots, H_n$ and the corresponding p-values $P_1, \ldots, P_n$, it is a global test of the intersection null hypothesis $H_0 : \bigcap_{i=1}^{n} H_i$ based on these p-values. It rejects H_0 at a significance level α if $P_{(i)} \le i\alpha/n$ for at least one $i = 1, \ldots, n$, where $P_{(1)} \le \cdots \le P_{(n)}$ are the ordered p-values.

Simes' test is more powerful than the Bonferroni test. However, to control the Type I error rate at the desired level, it requires certain assumptions about dependence structure of the p-values under H_0, unlike the Bonferroni test. For instance, if p-values are either independent or positively dependent in the following sense:

$$E_{H_0}\{\phi(P_1, \ldots, P_n) | P_i = u\} \text{ is non-decreasing in } u \quad (1)$$

for each $i = 1, \ldots, n$, and any coordinatewise non-decreasing function $\varphi(P_1, \ldots, P_n)$ of $P_1, \ldots, P_n$, then Simes' test controls the Type I error rate at α; that is, the following inequality holds:

$$\Pr_{H_0}\{\text{Rejecting } H_0\} = \Pr_{H_0}\left\{\bigcup_{i=1}^{n}\left(P_{(i)} \le i\alpha/n_0\right)\right\} \le \alpha.$$

Such positive dependence is exhibited by p-values in some commonly encountered multiple testing situations. For instance, p-values generated from (I) dependent standard normal variates with non-negative correlations, (II) absolute values of dependent standard normal variates with a correlation matrix R such that the off-diagonal entries of $-DR^{-1}D$ are non-negative for some diagonal matrix D with diagonal entries ± 1, (III) multivariate t with the associated normal variates having non-negative correlations (under a minor restriction on the range of values of u), and (IV) absolute values of multivariate t with the associated normal variates having a correlation matrix as in (II), satisfy (1) (Sarkar 1998, 2008a; Sarkar and Chang 1997).

For simultaneous testing of $H_1, \ldots, H_n$, the family-wise error rate (FWER), which is the probability of falsely

rejecting at least one null hypothesis, is often used as a measure of overall Type I error. Methods strongly controlling the FWER, that is, with this probability not exceeding a pre-specified value α under any configuration of true and false null hypotheses, have been proposed. Hochberg (1988) suggested such a method. It rejects H_i if $P_i \le P_{(\hat{i})}$, where

$$\hat{i} = \max\left\{i : P_{(i)} \le \alpha/(n-i+1)\right\}$$

provided the maximum exists, otherwise accepts all null hypotheses. This is a stepup method with the critical values $\alpha_i = \alpha/(n-i+1)$, $i = 1,\ldots,n$. For any stepup method with critical values $\alpha_1 \le \cdots \le \alpha_n$, the FWER is 0 if n_0, the number of true null hypotheses, is 0, otherwise it satisfies the following inequality:

$$FWER \le \Pr\left\{\bigcup_{i=1}^{n_0}\left(\hat{P}_{(i)} \le \alpha_{n-n_0+i}\right)\right\},$$

where $\hat{P}_{(1)} \le \cdots \le \hat{P}_{(n_0)}$ are the ordered versions of the p-values corresponding to the n_0 true null hypotheses (Romano and Shaikh 2006). For the Hochberg method, since

$$\alpha_{n-n_0+i} = \alpha/(n_0-i+1) \le i\alpha/n_0 \text{ for } i = 1,\ldots,n_0,$$

its FWER is bounded above by the Type I error rate of the level α Simes' test for the intersection of n_0 null hypotheses based on their p-values. In other words, the Hochberg method controls its FWER in situations where Simes' global test controls its Type I error rate.

The closed testing method of Marcus et al. (1976) is often used to construct multiple testing method with a strong control of the FWER. It operates as follows. Given a finite family of null hypotheses $\{H_i, i = 1,\ldots,n\}$, form the closure of this family by considering all non-empty intersections $H_J = \bigcap_{i\in J} H_i$ for $J \subseteq \{1,\ldots,n\}$. Suppose a level-$\alpha$ global test is available for each H_J. Then, a closed testing method rejects H_J if and only if every H_K with $K \supseteq J$ is rejected by its level-α test. Hommel (1988) used Simes' global test in the closed testing method to construct an improvement of the Hochberg method. It finds

$$\hat{j} = \max\left\{i : P_{(n-i+k)} \ge k\alpha/i \text{ for all } k = 1,\ldots,i\right\},$$

and rejects H_i if $P_i \le \alpha/\hat{j}$, provided the maximum exists, otherwise rejects all null hypotheses.

Benjamini and Hochberg (1995) introduced the ►false discovery rate (FDR), which is a less conservative notion of error rate than the FWER. With R and V denoting the total number rejections and the total number of false rejections, respectively, of null hypotheses, it is defined as follows:

$$FDR = E(V/\max\{R,1\}).$$

The FDR is said to be strongly controlled at α by a multiple testing method if the above expectation does not exceed α, irrespective of the number of true null hypotheses. As noted in Hommel (1988), while making decisions on the individual null hypotheses using the stepup method based on the critical values in the Simes' test, which are $\alpha_i = i\alpha/n$, $i = 1,\ldots,n$, the FWER is not strongly controlled. However, the false discovery rate (FDR) is strongly controlled, again if the p-values are independent or positively dependent in the sense of (1), but with the P_i now representing the p-value corresponding to a null hypothesis (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001; Sarkar 2002). A proof of this result can be seen in Sarkar (2008b), who gave the following expression for the FDR of a stepup method with critical values $\alpha_1 \le \cdots \le \alpha_n$:

$$FDR = \sum_{i\in J_0} E\left[\frac{I\left(P_i \le \alpha_{R_{n-1}^{(-i)}+1}\right)}{R_{n-1}^{(-i)}+1}\right],$$

where I is the indicator function, J_0 is the set of indices corresponding to the true null hypotheses, $R_{n-1}^{(-i)}$ is the number of rejections in the stepup method based on the $n-1$ p-values other than P_i and the critical values $\alpha_2 \le \cdots \le \alpha_n$. Examples of p-values satisfying this positive dependence condition are those that are generated from test statistics in (I) and (III).

About the Author

Dr. Sanat K. Sarkar is Professor and Senior Research Fellow, Department of Statistics, Temple University. He is a Fellow of the Institute of Mathematical Statistics, a Fellow of the American Statistical Association, and an Elected member of the International Statistical Institute. He is Associate Editor of *Annals of Statistics*, *The American Statistician* and *Sankhya, B*. He has made significant contributions to the development of modern multiple testing techniques. He has received a number of awards and honors from his university for his research contributions.

Cross References

►False Discovery Rate
►Multiple Comparison
►Multiple Comparisons Testing from a Bayesian Perspective

References and Further Reading

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29: 1165–1188

Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800–802
Hommel G (1988) A stagewise rejective multiple test procedure based on a modifed Bonferroni test. Biometrika 75:783–786
Marcus R, Peritz E, Gabriel KR (1976) On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63:655–660
Romano JP, Shaikh AM (2006) Stepup procedures for control of generalizations of the familywise error rate. Ann Stat 34:1850–1873
Sarkar SK (1998) Some probability inequalities for ordered *MTP*2 random variables: a proof of the Simes conjecture. Ann Stat 26:494–504
Sarkar SK (2002) Some results on false discovery rate in stepwise multiple testing procedures. Ann Stat 30:239–257
Sarkar SK (2008a) On the Simes inequality and its generalization. IMS collections beyond parametrics. In: Interdisciplinary research: Festschrift in honor of professor Pranab K. Sen 1: 231–242
Sarkar SK (2008b) On methods controlling the false discovery rate. Sankhya A 70(Part 2):135–168
Sarkar SK, Chang CK (1997) The Simes method for multiple hypothesis testing with positively dependent test statistics. J Am Stat Assoc 92:1601–1608
Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73:751–754

Simple Linear Regression

Sung H. Park
Professor
Seoul National University, Seoul, Korea

Regression analysis is a collection of statistical modeling techniques that usually describes the behavior of a random variable of interest by using one or more quantitative variables. The variable of interest may be the crop yield, the price of oil in the world market, the tensile strength of metal wire, and so on. This variable of interest is called the *dependent variable*, or *response* variable and denoted with Y. Other variables that are thought to provide information on the dependent variable are incorporated into the model as *independent* variables. These variables are also called the *predictor*, or *regressor*, or *explanatory* variables, and are denoted by Xs. If the height of a son is affected by the height of his father, then the height of the father is X and the height of the son becomes Y.

The Xs are assumed to be known constants. In addition to the Xs, all models involve unknown constants, called *parameters*, which control the behavior of the model. In practical situations, the statistical models usually fall into the class of models that are *linear in the parameters*. That is, the parameters enter the model as simple coefficients on the independent variables. Such models are referred to as ►*linear regression* models. If there is only one independent variable X for the dependent variable of interest Y, and the functional relationship between Y and X is a straight line, this model is called the *simple linear regression* model.

In a nonstatistical context, the word *regression* means "to return to an earlier place or state." The term "*regression*" was first used by Francis Galton (1822–1911), who observed that children's heights tended to "revert" to the average height of the population rather than diverting from it. Galton applied "a regression line" to explain that the future generations of offspring who are taller than average are not progressively taller than their respective parents, and parents who are shorter than average do not beget successively smaller children. But the term is now applied to any linear or nonlinear functional relationships in general.

In the simple linear model, the true mean of Y changes at a constant rate as the value of X increases or decreases. Thus, the functional relationship between the true mean of Y, denoted by $E(Y)$, and X is the equation of a straight line

$$E(Y) = \beta_0 + \beta_1 X.$$

Here, β_0 is the intercept, the value of $E(Y)$ when $X = 0$, and β_1 is the slope of the line, the rate of change in $E(Y)$ per unit change in X. Suppose we have n observations on Y, say, $Y_1, Y_2, Y_3, \ldots, Y_n$ at $X_1, X_2, X_3, \ldots, X_n$, respectively. The i^{th} observation on the dependent variable Y_i at X_i is assumed to be a random observation with the random error ε_i to give the statistical model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \quad (1)$$

The random errors ε_i have zero mean and assumed to have common variance σ^2 and to be pairwise independent. The random error assumptions are frequently stated as

$$\varepsilon_i \sim NID(0, \sigma^2)$$

where *NID* stands for normally and independently distributed. The quantities in parentheses denote the mean and the variance, respectively, of the normal distribution.

Once β_0 and β_1 in Eq. 1 have been estimated from a given set of data on X and Y, the following prediction equation results:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X \text{ or } \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \quad (2)$$

The "hats" (as they are called) above β_0 and β_1 signify that those parameters are being estimated, but the hat above Y means that the dependent variable is being predicted. Point estimates of β_0 and β_1 are needed to obtain the fitted line given in Eq. 2. One way is to minimize the sum of the absolute values of the vertical distances with each distance measured from each point to the fitted line (see,

e.g., Birkes and Dodge 1993). These vertical distances are called ►residuals. The standard approach, however, is to minimize the sum of the squares of the vertical distances, and this is accomplished by using the *method of least squares*.

The starting point of the method of ►least squares is to write the estimated model as

$$e = \widehat{Y} - (\widehat{\beta}_0 + \widehat{\beta}_1 X)$$

since the residual e represents the vertical distance Y to the line. Then the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are chosen that minimize the sum of the squares of residuals

$$S = \Sigma e_i^2 = \Sigma \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right)^2.$$

To minimize S, we take the partial derivative of S with respect to each of the two estimates and set the resulting expressions equal to zero. Thus we obtain

$$\widehat{\beta}_0 n + \widehat{\beta}_1 \Sigma X_i = 0$$

$$\widehat{\beta}_0 \Sigma X_i + \widehat{\beta}_1 \Sigma X_i^2 = 0$$

which are called the *normal equations*. If we solve these equations for $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we obtain

$$\widehat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}.$$

The method of least squares, on which most methods of estimation for regression analysis are based was apparently first published by Legendre (1805), but the first treatment along the lines now familiar was given by Gauss (for the details regarding history of least squares *see* ►Gauss–Markov theorem). Gauss showed that the least squares method gives estimators of the unknown parameters with minimum variance among unbiased linear estimators. This basic result is now known as the Gauss–Markov theorem, and the least squares estimators as Gauss–Markov estimators. That is, there is no other choice of values for the two parameters β_0 and β_1 that provide a smaller $\sum e_i^2$. If a residual, e_i, is too large compared with the other residuals, the corresponding Y_i may be an outlier or may be an influential observation that influences the estimates of two parameters β_0 and β_1. Detection of an outlier or an influential observation is an important research area, and many books such as Belsley et al. (1980) and Cook and Weisberg (1982), deal with this topic. (see also ►Cook's distance, ►Regression diagnostics, ►Influential observations).

About the Author

Professor Sung Park is Past President of Korean Statistical Society (1997–1998), Korean Society for Quality Management, Vice President of International Society for Business and Industrial Statistics, and Academician of International Academy for Quality. In 2000, he received the prestigious gold medal from the President of the Korean Government for his contribution to quality management in Korea. Recently, he has served as the Dean of the College of Natural Sciences, Seoul National University. He has published more than 30 books on statistics and quality control including three books in English: *Robust Design and Analysis for Quality Engineering* (Chapman & Hall, 1996), and edited the text *Statistical Process Monitoring and Optimization* (with G. Geoffrey Vining, Marcel Dekker, 2000) and *Six Sigma for Quality and Productivity Promotion* (Asian Productivity Organization, Free eBook, 2003.)

Cross References

►Cook's Distance
►Gauss-Markov Theorem
►Influential Observations
►Least Squares
►Linear Regression Models
►Regression Diagnostics
►Residuals

References and Further Reading

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York

Birkes D, Dodge Y (1993) Alternative methods of regression. Wiley, New York

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall, London

Legendre AM (1805) Nouvelles méthodes pour la détermination des orbites des comètes. Firmin Didot, Paris

Simple Random Sample

Roger E. Kirk
Distinguished Professor of Psychology and Statistics
Baylor University, Waco, TX, USA

A ►census, surveying every element in a finite population, is used to discover characteristics of the population. If the population is large, a census can be costly, time consuming, or impracticable. Alternatively, a simple random sample can be used to obtain information and draw inferences about the population. It is customary to sample elements without replacement. That is, once an element has been

selected, it is removed from the population so that it cannot be selected a second time. A simple random sampling procedure is used to obtain a simple random sample. The procedure selects a sample of size n from a finite population of size $N < n$ such that each of the ${}_{N}C_{n} = N!/[n!(N-n)!]$ possible samples is equally likely to be selected. If sample elements are returned to the population after being selected – sampling with replacement – each of the ${}_{N+n-1}C_{n} = (N+n-1)!/\{n![(N+n-1)-n]!\}$ possible samples is equally likely to be selected.

Simple random sampling is a type of probability sampling. All probability sampling procedures have three characteristics in common: (a) the elements that compose the population are explicitly defined, (b) every potential sample of a given size that could be drawn from the population can be enumerated, and (c) the probability of selecting any potential sample can be specified. Non-probability sampling procedures do not satisfy one or more of the three characteristics. An example of a non-probability sampling procedure is convenience sampling-elements are selected because they are readily available. For simple random sampling without replacement, the probability of a particular sample being selected is $1/({}_{N}C_{n})$. For sampling with replacement, the probability of a particular sample being selected is $1/({}_{N+n-1}C_{n})$. When sampling with replacement the inclusions of the ith and jth $(i \neq j)$ members of the population are statistically independent. However, these events are not statistically independent when sampling without replacement. For this case, the probability of the inclusions of ith and jth population members is $n(n-1)/[N(N-1)]$ (McLeod 1988).

Simple random samples have two interrelated advantages over non-probability samples. First, randomness avoids bias, that is, a systematic or long-run misrepresentation of the population. Second, randomness enables researchers to apply the laws of probability in determining the likely error of sample statistics. A particular random sample rarely yields an estimate of the population characteristic that equals the population characteristic. However, the expected value of the sample estimate will over an indefinitely large number of samples equal the population characteristic. Furthermore, for any simple random sample, it is possible to estimate the magnitude of the error associated with the estimate. For large populations the error depends only on the sample size, a fact that is counterintuitive (Anderson 2001).

The first step in obtaining a simple random sample is to develop a sampling frame: a list of all of the elements in the population of interest. The sampling frame operationally defines the population from which the sample is drawn and to which the sample results can be generalized. Once the sampling frame has been developed, a simple random sample can be obtained in a variety of ways. For example, a researcher can record on a slip of paper the identifying code for each element in the sampling frame. The slips of paper are placed in a container and thoroughly shuffled. The first n unique slips drawn without bias from the container compose the sample. The most common method of obtaining a simple random sample uses random numbers. Tables of random numbers are available in many statistics textbooks. The tables contain a sequence of random digits whose terms are chosen so that each digit is equally likely to be $0, 1, \ldots, 9$ and the choices at any two different places in the sequence are independent. For ease in reading the digits in a random number table, the digits are often grouped with two digits in a group, four digits in a group, and so on. To use a table to select a simple random sample of size, say, $n = 50$ from a population of size $N = 988$, assign the numbers $000, 002, \ldots, 987$ to the elements in the sampling frame. Select a starting point in the table by dropping a pointed object on the table. Choose three-digit numbers beginning at the starting point until 50 distinct numbers between 000 and 987 are obtained. The sample consists of the elements corresponding to the 50 numbers selected. This procedure illustrates sampling without replacement because once a number has been selected, the number is ignored if it is encountered again. Computer packages such as SAS, SPSS, and MINITAB and many hand calculators have routines that produce numbers that in every observable way appear to be random. For an in-depth discussion of sampling procedures, see Schaeffer et al. (2006).

About the Author

Professor Kirk is a Fellow of the American Psychological Association, Association for Psychological Science, American Educational Research Association, and the American Association of Applied and Preventive Psychology. He is the 2005 recipient of the American Psychological Association's Jacob Cohen Award for Distinguished Contributions to Teaching and Mentoring. He is a Founding Associate editor of the *Journal of Educational Statistics*, 42nd president of the Southwestern Psychological Association (1995–1996), and 46th president of Division 5 (Evaluation, Measurement, and Statistics) of the American Psychological Association (1992–1993). Professor Kirk was President of the Society for Applied Multivariate Research (1984–1985).

Cross References

►Handling with Missing Observations in Simple Random Sampling and Ranked Set Sampling
►Non-probability Sampling Survey Methods

►Proportions, Inferences, and Comparisons
►Ranked Set Sampling
►Representative Samples
►Sample Size Determination
►Sampling From Finite Populations
►Uniform Random Number Generators

References and Further Reading

Anderson NH (2001) Empirical directions in design and analysis. Erlbaum, Mahwah

McLeod I (1988) Simple random sampling. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 8. Wiley, New York, pp 478–479

Schaeffer RL, Ott RL, Mendenhall W (2006) Elementary survey sampling 6th edn. Thompson Learning, Belmont

Simpson's Paradox

Zhi Geng
Professor, Director of the Institute of Mathematical Statistics of Peking University
Peking University, Beijing, China

An association measurement between two variables X and Y may be dramatically changed from positive to negative by omitting a third variable Z, which is called Simpson's paradox or the Yule-Simpson paradox (Yule, 1903; Simpson, 1951). A numerical example is shown in Table 1. The risk difference (RD) is defined as the difference between the recovery proportion in the treated group and that in the placebo group, $RD = (80/200) - (100/200) = -0.10$. If the population is split into two populations of male and female, a dramatic change can be seen from Table 2. The risk differences for male and female are both changed to 0.10. Thus we obtain a self-contradictory conclusion that the new drug is effective for both male and female but it is ineffective for the whole population. Should patients in the population take the new drug or not? Should the correct answer depend on whether the doctor know the gender of patients?

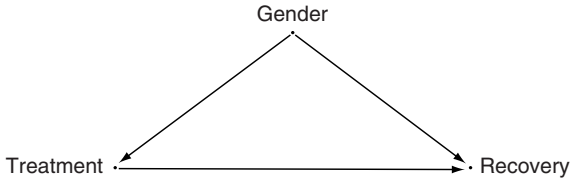
From Table 2, we can see that most males took placebo, but most females took the new drug. As depicted in Fig. 1, there may be a spurious association between treatment and response because gender associates with both treatment and response. Such a factor that is associated with both treatment and response is called a confounding factor or a confounder. If a confounder is known and observed, the bias due to the confounder can be removed by stratification or standardization. If there are unknown or unobserved confounders, in order to remove the confounding bias, we can randomize the treatment assignment such that the association path between the confounders and the treatment is broken. In Fig. 2, we depict a randomized experiment for this example, where 200 males (M) and

Simpson's Paradox. Table 1 Recovery proportions in treatment and placebo groups

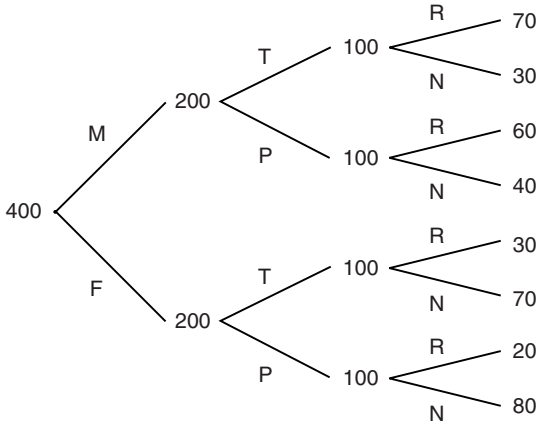
Treatment	Recovery	Non-recovery	Total
New drug	80	120	200
Placebo	100	100	200
	$RD = \frac{80}{200} - \frac{100}{200} = -0.10$		

Simpson's Paradox. Table 2 Populations stratified by gender

	Male		Female	
Treatment	Recovery	Non-recovery	Recovery	Non-recovery
New drug	35	15	45	105
Placebo	90	60	10	40
	$RD_M = 0.10$		$RD_F = 0.10$	



Simpson's Paradox. Fig. 1 A confounding factor: gender



Simpson's Paradox. Fig. 2 Randomized experiment

Simpson's Paradox. Table 3 Subscription renewal rates in 1979

Month	Source of current subscription					Overall
	Gift	Previous renewal	Direct mail	Subscription service	Catalog agent	
January						
Total	3,594	18,364	2,986	20,862	149	45,955
Renewals	2,918	14,488	1,783	4,343	13	23,545
Rate	0.812	0.789	0.597	0.208	0.087	0.512
February						
Total	884	5,140	2,224	864	45	9,157
Renewals	704	3,907	1,134	122	2	5,869
Rate	0.796	0.760	0.510	0.141	0.044	0.641

Simpson's Paradox. Table 4 Total income and total tax (in 10^3 dollars) and tax rate

Adjusted gross income	1974			1978		
	Income	Tax	Tax rate	Income	Tax	Tax rate
Under $5,000	41,651,643	2,244,467	0.054	19,879,622	689,318	0.035
$5,000 to $9,999	146,400,740	13,646,348	0.093	122,853,315	8,819,461	0.072
$10,000 to $14,999	192,688,922	21,449,597	0.111	171,858,024	17,155,758	0.100
$15,000 to $99,999	470,010,790	75,038,230	0.160	865,037,814	137,860,951	0.159
$100,000 or more	29,427,152	11,311,672	0.384	62,806,159	24,051,698	0.383
Total	880,179,247	123,690,314	0.141	1,242,434,934	188,577,186	0.152

200 females (F) are randomly assigned into the new drug group (T) and the placebo group (M). The recovery proportion is 35/50 in the new drug group of males, and thus 70 of 100 treated males recover (R) and the other 30 do not recover (N). From Fig. 2, the total number of recovered people is 70+30=100 and the recovery proportion is 100/200 in the new drug group; the total number is 60+20=80 and the recovery proportion is 80/200 in the placebo group. Thus, we conclude on that the new drug increases recovery proportion by 10%, which is consistent with that shown in Table 2.

Two real-life examples of Simpson's paradox were showed by Wagner (1982). The first example, as shown in Table 3, illustrates that the overall renewal rate of *American History Illustrated* magazine increased from 51.2 percent in January 1979 to 64.1 percent in February 1979, but the renewal rates actually declined in every subscription category. The second example, as shown in Table 4, illustrates that the overall income tax rate increased from 14.1 percent in 1974 to 15.2 percent in 1978, but the tax rate decreased in each income category. Reintjes et al. (2000) gave the following example from hospital epidemiology: 3519 gynecology patients from eight hospitals in a nonexperimental study were used to study the association between antibiotic prophylaxis (AB-proph.) and urinary tract infections (UTI). The eight hospitals were stratified into two groups with a low incidence percentage (< 2.5%) and a high percentage (> 2.5%) of UTI. By Table 5, the relative risk (RR) was (42/1279)/(104/2240) = 0.7 for the overall eight hospitals, which means that AB-proph. had a protective effect on UTI. But the RRs were 2.6 and 2.0 for the low and the high incidence groups, respectively, which means that

Simpson's Paradox. Table 5 Data on UTI and AB-proph. stratified by incidence of UTI per hospital

	Hospitals with low UTI		Hospitals with high UTI		All hospitals	
AB-proph.	UTI	no-UTI	UTI	no-UTI	UTI	no-UTI
Yes	20	1093	22	144	42	1237
No	5	715	99	1421	104	2136
	$RR_L = 2.6$		$RR_H = 2.0$		$RR = 0.7$	

AB-proph. had a risk effect on UTI for both groups. The real effect of AB-proph. on UTI has been shown to be protective in randomized clinical trials, which is consistent with the crude analysis rather than the stratified analysis. This result explains that there were more unidentified confounders which canceled their effects each other out in the crude analysis.

There are many topics related to Simpson's paradox. Collapsibility of association measurements deals with conditions under which association measurements are unchanged by omitting other variables (Cox and Wermuth, 2003; Ma et al. 2006). From the viewpoint of causality, Simpson's paradox occurs because there are confounders such that association measurement is biased from causal effects (Pearl, 2000; Geng et al. 2002). A variation of Simpson's paradox is a surrogate paradox, which means that a treatment has a positive effect on an intermediate variable called a surrogate, which in turn has a positive effect on the true endpoint, but the treatment has a negative effect on the true endpoint (Chen et al. 2007; Ju and Geng, 2010). Moore (1995) describes a real trial of antiarrhythmic drugs in which an irregular heartbeat is a risk factor of early mortality but correction of the heartbeat increased mortality.

About the Author

Dr. Zhi Geng is Professor, School of Mathematical Sciences, Peking University. He obtained his PhD (1989) from Kyushu University, Japan. He has been a member of International Statistical Institute since 1996. Professor Geng is Associate Editor of *Computational Statistics and Data Analysis*.

Cross References

►Causation and Causal Inference
►Collapsibility
►Confounding and Confounder Control
►Statistical Fallacies

References and Further Reading

Chen H, Geng Z, Jia J (2007) Criteria for surrogate endpoints. J R Stat Soc B 69:919–932
Cox DR, Wermuth N (2003) A general condition for avoiding effect reversal after marginalization. J R Stat Soc B 65:937–941
Geng Z, Guo J, Fung WK (2002) Criteria for confounders in epidemiological studies. J R Stat Soc B 64:3–15
Ju C, Geng Z (2010) Criteria for surrogate endpoints based on causal distributions. J R Stat Soc B 72:129–142
Ma ZM, Xie XC, Geng Z (2006) Collapsibility of distribution dependence. J R Stat Soc B 68:127–133
Moore T (1995) Deadly medicine: why tens of thousands of patients died in America's worst drug disaster. Simon & Shuster, New York
Pearl J (2000) Causality: models, reasoning, and inference. University Press, Cambridge
Reintjes R, de Boer A, van Pelt W, Mintjes-de Groot J (2000) Simpson's paradox: an example from hospital epidemiology. Epidemiology 11:81–83
Simpson EH (1951) The interpretation of interaction in contingency tables. J R Stat Soc B 13:238–241
Wagner CH (1982) Simpson's paradox in real life. Am Stat 36:46–48
Yule GU (1903) Notes on the theory of association of attributes in statistics. Biometrika 2:121–134

Simulation Based Bayes Procedures for Model Structures with Non-Elliptical Posteriors

Lennart Hoogerheide[1], Herman K. van Dijk[2]
[1]Econometric and Tinbergen Institutes, Erasmus University Rotterdam, Rotterdam, The Netherlands
[2]Professor, Director of the Tinbergen Institute Econometric and Tinbergen Institutes, Erasmus University Rotterdam, Rotterdam, The Netherlands

The financial market turmoil has been shocking the world since early 2008. As is aptly stated by the president of the European Central Bank, Trichet (2008), the widespread

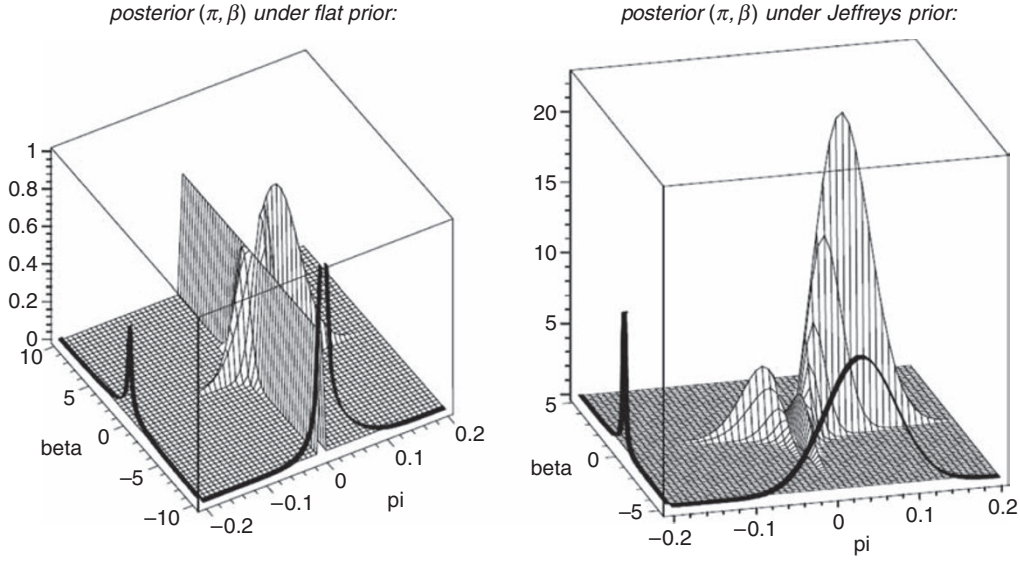
undervaluation of risk is one of the most important issues in this context and appropriate operational risk management is a crucial issue to be investigated. A seemingly unrelated issue is to measure and predict the *treatment effect of education on income*. This issue is crucial for any country that increasingly relies on the "knowledge economy." In recent research by the authors it is stressed that ***these seemingly unrelated issues pose similar questions and have common components from a modeling and statistical viewpoint***.

There exist connections between dynamic time series models used in the first issue and treatment effect models. This common problem structure is explained in research by the authors as follows: the restricted reduced form of the instrumental variable (IV) model and the Vector Error Correction Model (VECM) under cointegration are both instances of the general reduced rank regression model with different variables and parameters playing the same roles, as summarized in the Table 1.

In these models with near reduced rank one may encounter non-elliptical posteriors. In the Bayesian analysis of treatment effects, for example in the instrumental variable (IV) model, we often encounter posterior distributions that display these shapes. The reason for this is

Simulation Based Bayes Procedures for Model Structures with Non-Elliptical Posteriors. Table 1 Common model structures

Model	Restricted reduced form (RRF) of instrumental variable (IV) model	Vector Error Correction Model (VECM) under cointegration
Endogenous variables	Dependent variable and (possibly) endogenous regressors	Vector of variables' changes (= current − previous values)
Predetermined variables corresponding to parameter matrix with *reduced rank*	Instrumental variables (having no *direct* effect on the dependent variable, only an *indirect* effect via the (possibly) endogenous regressors)	Vector of previous values
Predetermined variables corresponding to *unrestricted* parameter matrix	Control variables (having a *direct* effect on both the dependent variable and the (possibly) endogenous regressors)	Vector of other explanatory variables and past variables' changes



Simulation Based Bayes Procedures for Model Structures with Non-Elliptical Posteriors. Fig. 1 Posterior density of π (expected difference in years of education between children born in April-December and January-March) and β (treatment effect of education on income) for 29,015 data (used by Angrist and Krueger (1991)) from men born in the state of New York

local non-identification: if some of the model parameters (reflecting the strength of the instruments) tend to 0, i.e., the case of weak instruments, other model parameters (corresponding to the relevant treatment effect) become unidentified.

Angrist and Krueger (1991) consider the estimation of the treatment effect β of education on income, which is non-trivial due to unobserved (intellectual) capabilities that not only influence education but also directly affect income, and due to measurement errors in the reported education level. Angrist and Krueger (1991) use American data and suggest using quarter of birth to form ►instrumental variables. These instruments exploit that students born in different quarters have different average education. This results since most school districts require students to have turned age six by a certain date, a so-called "birthday cutoff" which is typically near the end of the year, in the year they enter school, whereas compulsory schooling laws compel students to remain at school until their 16th, 17th or 18th birthday. This asymmetry between school-entry requirements and compulsory schooling laws compels students born in certain months to attend school longer than students born in other months: students born earlier in the year enter school at an older age and reach the legal dropout age after less education. Hence, for students who leave school as soon as the schooling laws allow for it, those born in the first quarter have on average attended school for three quarters less than those born in the fourth quarter. Suppose we use as a single instrument a 0/1 indicator variable with value 0 indicating birth in the first quarter; the strength of this instrument is given by its effect on education, parameter π. The left panel of Fig. 1 shows the posterior density of π and β (under a flat prior) for 29,015 data from men born in the state of New York in 1930–1939. This shows a clear "ridge" around $\pi = 0$, indicating that for π tending to 0 a wide range of values of β becomes possible. An alternative prior, the Jeffreys prior, regularizes the posterior shapes in the sense that it eliminates the asymptote around $\pi = 0$ for the marginal posterior of π, yet the joint posterior shapes in the right panel of Fig. 1 are still far from elliptical. This example illustrates that the weakness of the instruments may imply that even for large data sets posterior distributions may be highly non-elliptical.

Thus for the Bayesian analysis of (non-linear) extensions of the IV model, we need flexible simulation methods. The use of neural network based simulation is then particularly useful. A Bayesian optimal information processing procedure using advanced simulation techniques based on artificial neural networks (ANN) is recently developed and it can be used as a powerful tool for forecasting and policy advice. These simulation methods have already been successfully applied to evaluate risk measures (Value-at-Risk, Expected Shortfall) for a single asset. The procedures proposed by the authors are just one step forward on the path of understanding these issues and these involve a novel manner of processing the information flow on these issues. It is – of course – the intention of this research that its results improve forecasting of risk and uncertainty that influence the effectiveness of interventions and treatments.

About the Author

Herman K. van Dijk is director of the Tinbergen Institute and professor of Econometrics with a Personal Chair at Erasmus University Rotterdam. He is a former Director of the Econometric Institute and Honorary Fellow of the Tinbergen Institute. He has been a visiting fellow and a visiting professor at Cambridge University, the Catholic University of Louvain, Harvard University, Duke University, Cornell University, and the University of New South Wales. He received the Savage Prize for his Ph.D. dissertation and is listed in the journal Econometric Theory in the Econometricians Hall of Fame amongst the top ten European econometricians. He serves on the Editorial Board of major journals in econometrics. His publications consist of several books and more than 160 international scientific journal papers and reports.

Cross References

►Instrumental Variables
►Neural Networks
►Quantitative Risk Management

References and Further Reading

Angrist JD, Krueger AB (1991) Does compulsory school attendance affect schooling and earnings? Quart J Econom 106:979–1014

Ardia D, Hoogerheide LF, Van Dijk HK (2009) To bridge, to warp or to wrap? A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihoods. TI discusion paper 09-017/4

Ardia D, Hoogerheide LF, Van Dijk HK (2009b) AdMit: adaptive mixtures of student-*t* distributions. The R Journal 1(1):25–30

Ardia D, Hoogerheide LF, Van Dijk HK (2009c) Adaptive mixture of Student-*t* distributions as a flexible candidate distribution for efficient simulation: the R package AdMit. J Stat Softw 29(3): 1–32

Hoogerheide LF (2006) Essays on neural network sampling methods and instrumental variables. Ph.D. thesis, Book nr. 379 of the Tinbergen Institute Research Series, Erasmus University Rotterdam

Hoogerheide LF, Van Dijk HK (2006) A reconsideration of the Angrist-Krueger analysis on returns to education. Report 2006-15 of the Econometric Institute, p 35

Hoogerheide LF, Van Dijk HK (2007) Note on neural network sampling for Bayesian inference of mixture processes. Bulletin of the

International Statistical Institute, Proceedings of the Biennial Sessions in Lisbon 2007, p 8
Hoogerheide LF, Van Dijk HK (2010) Bayesian forecasting of value at risk and expected shortfall using adaptive importance sampling. Int J Forecasting, 26:231–247
Hoogerheide LF, Kaashoek JF, Van Dijk HK (2003) Neural network approximations to posterior densities: an analytical approach. In: Proceedings of the section on Bayesian statistical science, American Statistical Association, 2003, p 5
Hoogerheide LF, Kaashoek JF, Van Dijk HK (2007a) On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks. J Econom 139(1): 154–180
Hoogerheide LF, Kleibergen F, Van Dijk HK (2007b) Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data. J Econom 138(1):63–103
Hoogerheide LF, Kleijn R, Ravazzolo F, Van Dijk HK, Verbeek M (2010) Forecast accuracy and economic gains from Bayesian model averaging using time varying weights. J Forecast, 29:251–269
Hoogerheide LF, Van Dijk HK, Van Oest RD (2009) Simulation based Bayesian econometric inference: principles and some recent computational advances. Chapter 7 in Handbook of Computational Econometrics, Wiley, pp 215–280
Trichet JC (2008) Macroeconomic policy is essential to stability. Financial Times, November 13, 2008, p 13

Singular Spectrum Analysis for Time Series

Anatoly Zhigljavsky
Professor, Chair in Statistics
Cardiff University, Cardiff, UK

Singular spectrum analysis (SSA) is a technique of time series analysis and forecasting. It combines elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. SSA aims at decomposing the original series into a sum of a small number of interpretable components such as a slowly varying trend, oscillatory components and a "structureless" noise. It is based on the singular-value decomposition of a specific matrix constructed upon time series. Neither a parametric model nor stationarity-type conditions have to be assumed for the time series; this makes SSA a model-free technique.

The commencement of SSA is usually associated with publication of the papers (Broomhead and King 1986a, b) by Broomhead and King. Nowadays SSA is becoming more and more popular, especially in applications. There are several hundred papers published on methodological aspects and applications of SSA, see Golyandina et al. (2001), Vautard et al. (1992), Vautard and Ghil (1989), Allen and Smith (1996), and Zhigljavsky (2010) and references therein. SSA has proved to be very successful, and has already become a standard tool in the analysis of climatic, meteorological and geophysical time series; see, for example, Vautard et al. (1992), Vautard and Ghil (1989), and Allen and Smith (1996). More recent areas of application of SSA include engineering, medicine, econometrics and many other fields. Most recent developments in the theory and methodology of SSA can be found in Zhigljavsky (2010). We start with 'Basic SSA', which is the most common version of SSA.

Basic SSA

Let $x_1, \ldots, x_N$ be a time series of length N. Given a window length L $(1<L<N)$, we construct the L-lagged vectors $X_i = (x_i, \ldots, x_{i+L-1})^T$, $i = 1, 2, \ldots, K = N–L+1$, and compose these vectors into the matrix $\mathbf{X} = (x_{i+j-1})_{i,j=1}^{L,K} = [X_1 : \ldots : X_K]$. This matrix has size $L \times K$ and is often called "trajectory matrix." It is a Hankel matrix, which means that all the elements along the diagonal $i+j$=const are equal.

The columns X_j of $\mathbf{X}$, considered as vectors, lie in the L-dimensional space $\mathbb{R}^L$. The singular-value decomposition of the matrix $\mathbf{X}\mathbf{X}^T$ yields a collection of L eigenvalues and eigenvectors. A particular combination of a certain number $l < L$ of these eigenvectors determines an l-dimensional subspace in $\mathbb{R}^L$. The L-dimensional data $\{X_1, \ldots, X_K\}$ is then projected onto this l-dimensional subspace and the subsequent averaging over the diagonals gives us some Hankel matrix $\tilde{\mathbf{X}}$ which is considered as an approximation to $\mathbf{X}$. The series reconstructed from $\tilde{\mathbf{X}}$ satisfies some linear recurrent formula which may be used for forecasting.

In addition to forecasting, the Basic SSA can be used for smoothing, filtration, noise reduction, extraction of trends of different resolution, extraction of periodicities in the form of modulated harmonics, gap-filling (Kondrashov and Ghil 2006; Golyandina and Osipov 2007) and other tasks, see Golyandina et al. (2001). Also, the Basic SSA can be modified and extended in many different ways some of which are discussed below.

Extensions of the Basic SSA

SSA for analyzing stationary series (Vautard and Ghil 1989). Under the assumption that the series $x_1, \ldots, x_N$ is stationary, the matrix $\mathbf{X}\mathbf{X}^T$ of the Basic SSA is replaced with the so-called lag-covariance matrix $\mathbf{C}$ whose elements are $c_{ij} = \frac{1}{N-k}\sum_{t=1}^{N-k} x_t x_{t+k}$ with $i,j = 1, \ldots, L$ and $k = |i-j|$. In the terminology of Golyandina et al. (2001), this is "Toeplitz SSA."

Monte-Carlo SSA (Allen and Smith 1996). In the Basic SSA we implicitly associate the "structureless" component of the resulting SSA decomposition with "white noise" (this noise may not necessarily be random). In some applications, however, it is more natural to assume that the noise is "colored". In this case, special tests based on Monte Carlo simulations may be used to test the hypothesis of the presence of a signal.

Improvement or replacement of the singular-value decomposition (SVD) procedure. There are two main reasons why it may be worthwhile to replace the SVD operation in the Basic SSA with some another operation. The first reason is simplicity: in problems where the dimensions of the trajectory matrix is large, SVD may simply be too costly to perform; substitutions of SVD are available, see Golub and van Loan (1996) and Moskvina and Schmidt (2003). The second reason is the analysis of the accuracy of SSA procedures based on the perturbation theory (Zhigljavsky 2010). For example, in the problems of separating signal from noise, some parts of the noise are often found in SVD components corresponding to the signal. As a result, a small adjustment of the eigenvalues and eigenvectors is advisable to diminish this effect. The simplest version of the Basic SSA with a constant adjustment in all eigenvalues was suggested in Van Huffel (1993) and is sometimes called the minimum-variance SSA.

Low-rank matrix approximations, Cadzow iterations, connections with signal processing. As an approximation to the trajectory matrix $\mathbf{X}$, the Basic SSA yields the Hankel matrix $\tilde{\mathbf{X}}$. This matrix is obtained as a result of the diagonal averaging of a matrix of rank l. Hence $\tilde{\mathbf{X}}$ is typically a matrix of full rank. However, in many signal processing applications, when a parametric form of an approximation is of prime importance, one may wish to find a Hankel matrix of size $L \times K$ and rank l which gives the best approximation to $\mathbf{X}$; this is a problem of the structured low-rank approximation (Markovsky et al. 2006). The simplest procedure of finding a solution to this problem (not necessarily the globally optimal one though) is the so-called Cadzow iterations (Cadzow 1988) which are the repeated alternating projections of the matrices (starting at $\mathbf{X}$) to the set of matrices of rank l (by performing the singular-value decompositions) and to the set of Hankel matrices (by making the diagonal averaging). That is, Cadzow iterations are simply the repeats of the Basic SSA. It is not guaranteed however that Cadzow iterations lead to more accurate forecasting formulas than the Basic SSA (Zhigljavsky 2010).

SSA for change-point detection and subspace tracking (Moskvina and Zhigljavsky 2003). Assume that the observations $x_1, x_2, \ldots$ of the series arrive sequentially in time and we apply the Basic SSA to the observations at hand. Then we can monitor the distances from the sequence of the trajectory matrices to the l-dimensional subspaces we construct and also the distances between these l-dimensional subspaces. Significant changes in any of these distances may indicate on a change in the mechanism generating the time series. Note that this change in the mechanism does not have to affect the whole structure of the series but rather only a few of its components.

SSA for multivariate time series. Multivariate (or multichannel) SSA (shortly, MSSA) is a direct extension of the standard SSA for simultaneous analysis of several time series. Assume that we have two series, $X = \{x_1, \ldots, x_N\}$ and $Y = \{y_1, \ldots, y_N\}$. The (joint) trajectory matrix of the two-variate series (X, Y) can be defined as either $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ or $\mathbf{Z} = (X, Y)^T$, where $\mathbf{X}$ and $\mathbf{Y}$ are the trajectory matrices of the individual series X and Y. Matrix $\mathbf{Z}$ is block-Hankel rather than simply Hankel. Other stages of MSSA are identical to the stages of the univariate SSA except that we build a block-Hankel (rather than ordinary Hankel) approximation $\tilde{\mathbf{Z}}$ to the trajectory matrix $\mathbf{Z}$.

MSSA may be very useful for analyzing several series with common structure. MSSA may also be used for establishing a causality between two series. Indeed, the absence of causality of Y on X implies that the knowledge of Y does not improve the quality of forecasts of X. Hence an improvement in the quality of forecasts for X which we obtain using MSSA against univariate SSA forecasts for X gives us a family of SSA-causality tests, see Hassani et al. (2010).

Cross References

►Forecasting: An Overview
►Monte Carlo Methods in Statistics
►Statistical Signal Processing
►Time Series

References and Further Reading

Allen MR, Smith LA (1996) Monte Carlo SSA: detecting irregular oscillations in the presence of colored noise. J Clim 9:3373–3404

Broomhead DS, King GP (1986a) Extracting qualitative dynamics from experimental data. Physica D 20:217–236

Broomhead DS, King GP (1986b) On the qualitative analysis of experimental dynamical systems. In: Sarkar S (ed) Nonlinear phenomena and chaos. Adam Hilger, Bristol, pp 113–144

Cadzow JA (1988) Signal enhancement a composite property mapping algorithm. IEEE Trans Acoust Speech Signal Process 36: 49–62

Golub G, van Loan C (1996) Matrix computations. Johns Hopkins University Press, London

Golyandina N, Osipov E (2007) The Caterpillar-SSA method for analysis of time series with missing values. J Stat Plan Infer 137:2642–2653

Golyandina N, Nekrutkin V, Zhigljavsky A (2001) Analysis of time series structure: SSA and related techniques. Chapman & Hall/CRC Press, New York/London

Hassani H, Zhigljavsky A, Patterson K, Soofi A (2010) A comprehensive causality test based on the singular spectrum analysis, causality in science. In: McKay Illary P, Russo F, Williamson J (eds) Causality in Sciences, Oxford University Press

Kondrashov D, Ghil M (2006) Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Proc Geoph 13: 151–159

Markovsky I, Willems JC, Van Huffel S, De Moor B (2006) Exact and approximate modeling of linear systems: a behavioral approach. SIAM, Philadelphia

Moskvina V, Schmidt KM (2003) Approximate projectors in singular spectrum analysis. SIAM J Matrix Anal Appl 24:932–942

Moskvina VG, Zhigljavsky A (2003) An algorithm based on singular spectrum analysis for change-point detection. Commun Stat Simul Comput 32:319–352

Van Huffel S (1993) Enhanced resolution based on minimum variance estimation and exponential data modeling. Signal process 33:333–355

Vautard R, Ghil M (1989) Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. Physica D 35:395–424

Vautard R, Yiou P, Ghil M (1992) Singular spectrum analysis: a toolkit for short, noisy and chaotic series. Physica D 58:95–126

Zhigljavsky A (ed) (2010) Statistics and its interface, special issue on the singular spectrun analysis for time series. Springer, New York

SIPOC and COPIS: Business Flow – Business Optimization Connection in a Six Sigma Context

Rick L. Edgeman
Professor, Chair & Six Sigma Black Belt
University of Idaho, Moscow, ID, USA

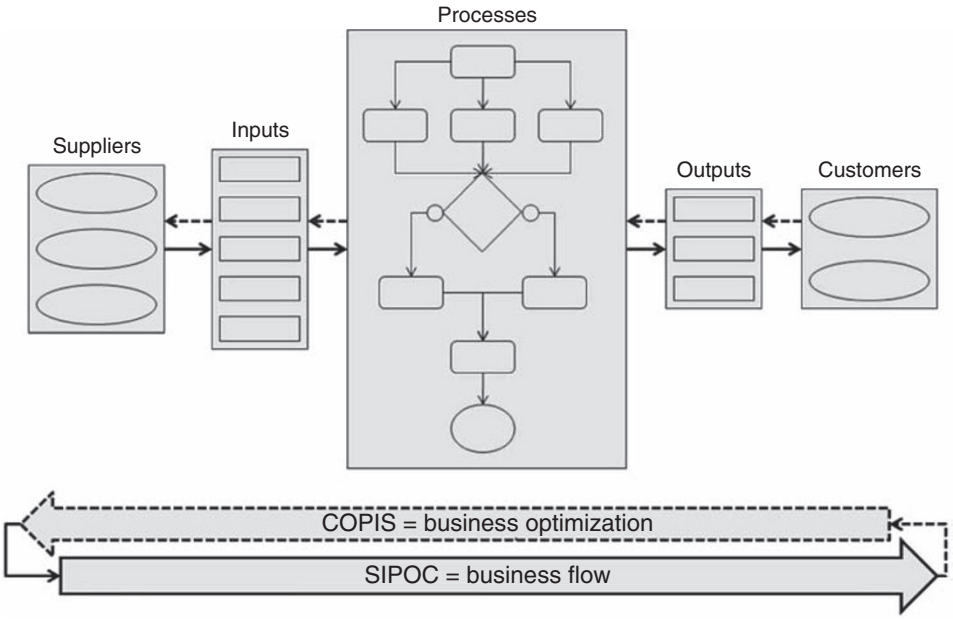
►Six Sigma can be defined as *a highly structured strategy for acquiring, assessing, and applying customer, competitor, and enterprise intelligence in order to produce superior product, system or enterprise innovation and designs* (Klefsjö et al. 2006). Focal to this definition is the customer and indeed the customer functions as the pivot point for this contribution as customer needs and wants drive change in most organizations.

Six Sigma originated at Motorola approximately 3 decades ago as a means of generating near-perfect products via focus on associated manufacturing processes and while initially applied almost exclusively in manufacturing environments, its inherent sensibilities and organization facilitated migration to service operations. Similarly, while Six Sigma was at the outset used to generate significant innovation in and improvement of existing products, those same sensibilities led to its adaptation to new product and process design environments. In statistical terms a process operating at a "true" six sigma level produces an average of only 3.4 defects per million opportunities (DPMO) for defects where this figure is associated with a process with a 12 standard deviation spread between lower and upper specification limits, but wherein the 3.4 DPMO figure is based on allowance for a 1.5 standard deviation non-centrality factor or shift away from "perfect centering" so that, in essence, one specification limit is 4.5 standard deviations away from the targeted or ideal performance level whereas the other specification limit is 7.5 standard deviations away from that performance level.

Within the context of a structured problem-solving context Six Sigma integrates various strategies and tools from Statistics, Quality, Business, and Engineering with the adoption of new ones likely as its use expands to more business sectors and areas of application. Its focus divides into two significant and related branches that share a number of tools, techniques and objectives, but often apply these tools and techniques differently and its use has added multiple billions in any currency to the financial bottom lines of numerous organizations across many sectors of the economy, including financial, healthcare, military, and general manufacturing. Six Sigma's branches are ones that focus on significant innovation/redesign in or of existing products, processes, and systems and a second that is directed at design of new products, processes or systems. Included among the leading companies emphasizing Six Sigma are GE, 3M, Raytheon, Sun Microsystems, DuPont, Bank of America, American Express, Motorola, Rolls Royce, and Boeing.

Central to business flow is the familiar SIPOC model (Suppliers → Inputs → Processes → Outputs → Customers) indicating that, commonly, suppliers provide inputs that are transformed by internal processes into outputs that are in turn provided to customers. While this flow is common and logical, its optimization is far less so, but can be approached application of Stephen Covey's familiar "habit" of "beginning with the end in mind" (Covey 1989), a manifestation of which in the present case is COPIS (Customers → Outputs → Processes → Inputs → Suppliers).

Organizations that practice COPIS – often as part of a quality management or six sigma culture – do so by first carefully elaborating who their customers are as well as the needs and wants of those customers (called the "Voice of the Customer" or "VOC"). Customer-driven organizations will ensure that these needs and wants are reflected in and fulfilled by the outputs of processes that must be optimally



SIPOC and COPIS: Business Flow – Business Optimization Connection in a Six Sigma Context. Fig. 1

configured in order to deliver these outputs by transforming the most appropriate inputs that have been provided by the most apt suppliers. It can be seen from this that, consistent with Covey, "see the end from the beginning," that is, to be customer-driven. In a continuous improvement culture this occurs not once, but cyclically. These ideas are portrayed in Fig. 1.

Statistical and other quantitatively oriented methods that can be brought to bear throughout the COPIS-SIPOC flow include the use of sample survey methods to elicit the VOC and numerous additional analytical techniques from across the statistical spectrum can be used to assess the VOC. Optimal process configuration is not merely a matter of work flow and equipment, but also of ensuring that however those are assembled, that the outputs themselves are optimized. While many tools can be employed, generally outputs can be regarded as response variables, Y, where

$$Y = f(X_1, X_2, \ldots, X_P) + \varepsilon,$$

where $X_1, X_2, \ldots, X_P$ are controllable variables, the optimal combination of settings of which can be determined using response surface methods, steepest ascent methods, and evolutionary operations or EVOP (Myers et al. 2009). In a similar way, such methods can be used to assist in selection of inputs and subsequently the suppliers from whom these should be obtained.

In all, what we see is that as best practice, business is conceived of as COPIS to yield optimal results as determined by the VOC, but subsequently deployed as SIPOC. While SIPOC is common to most business environments, employment of COPIS is practiced far less often and then typically only in customer-driven environments. Practice of COPIS offers rich opportunities for application of statistical methods as well as subsequent rewards.

About the Author

For biography *see* the entry ►Design for Six Sigma.

Cross References

►Business Statistics
►Design for Six Sigma
►Industrial Statistics
►Six Sigma

References and Further Reading

Covey SR (1989) The seven habits of highly effective people. Free, New York

Klefsjö B, Bergquist B, Edgeman, R (2006) Six sigma and total quality management: different day, same soup? Six Sigma and Competitive Advantage 2(2):162–178

Myers RH, Montgomery DC, Anderson-Cook CM (2009) Response surface methodology: process and product optimization using designed experiments, 3rd edn. Wiley, New York

Six Sigma

David M. Levine
Professor Emeritus of Statistics and Computer Information Systems
Baruch College, City University of New York, New York, NY, USA

Six Sigma is a quality improvement system originally developed by Motorola in the mid-1980s. After seeing the huge financial successes at Motorola, GE, and other early adopters of Six Sigma, many companies worldwide have now instituted Six Sigma to improve efficiency, cut costs, eliminate defects, and reduce product variation (see Arndt 2002; Cyger 2006; Hahn et al. 2000; Snee 2000). Six Sigma offers a more prescriptive and systematic approach to process improvement than TQM. It is also distinguished from other quality improvement systems by its clear focus on achieving bottom-line results in a relatively short 3- to 6-month period of time.

The name Six Sigma comes from the fact that it is a managerial approach designed to create processes that result in no more than 3.4 defects per million. The Six Sigma approach assumes that processes are designed so that the upper and lower specification limits are six standard deviations away from the mean. Then, if the processes are monitored correctly with ▶control charts, the worst possible scenario is for the mean to shift to within 4.5 standard deviations from the nearest specification limit. The area under the normal curve less than 4.5 standard deviations below the mean is approximately 3.4 out of a million.

The DMAIC Model

To guide managers in their task of improving short- and long-term results, Six Sigma uses a five-step process known as the *DMAIC model* – named for the five steps in the process:

- *Define*. The problem is defined, along with the costs, the benefits, and the impact on the customer.
- *Measure*. Operational definitions for each critical-to-quality (CTQ) variable are developed. In addition, the measurement procedure is verified so that it is consistent over repeated measurements.
- *Analyze*. The root causes of why defects occur are determined, and variables in the process causing the defects are identified. Data are collected to determine benchmark values for each process variable. This analysis often uses control charts.
- *Improve*. The importance of each process variable on the CTQ variable is studied using designed experiments. The objective is to determine the best level for each variable.
- *Control*. The objective is to maintain the benefits for the long term by avoiding potential problems that can occur when a process is changed.

The Define phase of a Six Sigma project consists of the development of a project charter, performing a SIPOC analysis, and identifying the customers for the output of the process. The development of a project charter involves forming a table of business objectives and indicators for all potential Six Sigma projects. Importance ratings are assigned by top management, projects are prioritized, and the most important project is selected. A SIPOC analysis is used to identify the *S*uppliers to the process, list the *I*nput provided to the suppliers, flowchart the *P*rocess, list the process *O*utputs, and identify the *C*ustomers of the process. This is followed by a Voice of the Customer analysis that involves market segmentation in which different types of users of the process are identified and the circumstances of their use of the process are identified. Statistical methods used in the Define phase include tables and charts, descriptive statistics, and control charts.

In the Measure phase of a Six Sigma project, members of a team first develop operational definitions of each CTQ variable. This is done so that everyone will have a firm understanding of the CTQ. Then studies are undertaken to ensure that there is a valid measurement system for the CTQ that is consistent across measurements. Finally, baseline data are collected to determine the capability and stability of the current process. Statistical methods used in the Measure phase include tables and charts, descriptive statistics, the normal distribution, the Analysis of Variance, and control charts.

The Analyze phase of a Six Sigma project focuses on the factors that affect the central tendency, variation, and shape of each CTQ variable. Factors are identified, related to each CTQ, have operational definitions developed, and have measurement systems established. Statistical methods used in the Analyze phase include tables and charts, descriptive statistics, the ▶Analysis of Variance, regression analysis, and control charts.

In the Improve phase of a Six Sigma project, team members carry out designed experiments to actively intervene in a process. The objective of the experimental design is to determine the settings of the factors that will optimize the central tendency, variation, and shape of each CTQ variable. Statistical methods used in the Improve phase include tables and charts, descriptive statistics, regression

analysis, hypothesis testing, the Analysis of Variance, and designed experiments.

The Control phase of a Six Sigma project focuses on the maintenance of improvements that have been made in the Improve phase. A risk abatement plan is developed to identify elements that can cause damage to a process. Statistical methods used in the Control phase include tables and charts, descriptive statistics, and control charts.

About the Author

Dr. David M. Levine is Professor Emeritus of Statistics and Computer Information Systems at Baruch College, City University of New York (CUNY). David has won a number of awards for his teaching including Teacher of the Year and the Baruch College Dean's Award for Continued Excellence in Teaching. He has also earned the Robert Pearson Award from the Northeast Region of the Decision Science Institute for his development of an innovative Statistical Process Control Course for Business Students and an Honorable Mention for the Decision Science Institute Innovation Teaching Award. He is a co author of the well known introductory books on statistics: *Basic Business Statistics* (with M.L. Berenson and T.C. Krehbiel, 11th edition, Prentice Hall, 2008) and *Statistics for Managers Using Microsoft Excel* (with D.F. Stephan, T.C. Krehbiel and M.L. Berenson, 5th edition, Prentice Hall, 2007). David's most recent specialty is Six Sigma and he is the author of well-received *Statistics for Six Sigma for Green Belts and Champions* and coauthor of *Quality Management*, and *Business Statistics for Quality and Productivity*.

Cross References

►Business Statistics
►Design for Six Sigma
►Industrial Statistics
►SIPOC and COPIS: Business Flow–Business Optimization Connection in a Six Sigma Context

References and Further Reading

Arndt M (2002) Quality isn't just for widgets. BusinessWeek July 22:72–73
Automotive Industry Action Group (AIAG) (1995) Statistical Process Control Reference Manual (Chrysler, Ford, and General Motors Quality and Supplier Assessment Staff
Bothe DR (1997) Measuring process capability. McGraw-Hill, New York
Cyger M (November/December 2006) The last word – riding the bandwagon. iSixSigma Magazine
Davis RB, Krehbiel TC (2002) Shewhart and zone control charts under linear trend. Commun Stat Simulat 31(1):91–96
Deming WE (1986) Out of the crisis. MIT Center for Advanced Engineering Study, Cambridge, MA
Deming WE (1993) The new economics for business, industry, and government. MIT Center for Advanced Engineering Study, Cambridge, MA
Gabor A (1990) The man who discovered quality. Time Books, New York
Gitlow H, Levine D (2005) Six sigma for green belts and champions. Financial Times/Prentice Hall, Upper Saddle River, NJ
Gitlow H, Levine D, Popovich E (2006) Design for six sigma for green belts and champions. Financial Times/Prentice Hall, Upper Saddle River, NJ
Hahn GJ, Doganaksoy N, Hoerl R (2000) The evolution of six sigma. Qual Eng 12:317–326
Lemak DL, Mero NP, Reed R (2002) When quality works: a premature post-mortem on TQM. J Bus Manage 8:391–407
Levine DM (2006) Statistics for six sigma for green belts with minitab and JMP. Financial Times/Prentice Hall, Upper Saddle River, NJ
Microsoft Excel 2007 (Redmond, WA: Microsoft Corp., 2007)
Scherkenbach WW (1987) The deming route to quality and productivity: road maps and roadblocks. CEEP, Washington, DC
Shewhart WA, Economic Control of the Quality of Manufactured Product (New York: Van Nostrand-Reinhard, 1931, reprinted by the American Society for Quality Control, Milwaukee, 1980)
Snee RD (2000) Impact of six sigma on quality. Qual Eng 12:ix–xiv
Vardeman SB, Jobe JM (2009) Statistical methods for quality assurance: basics, measurement, control, capability and improvement. Springer, New York
Walton M (1986) The Deming management method. Perigee Books, New York

Skewness

Paul von Hippel
Assistant Professor
University of Texas, Austin, TX, USA

Skewness is a measure of distributional asymmetry. Conceptually, skewness describes which side of a distribution has a longer tail. If the long tail is on the right, then the skewness is rightward or positive; if the long tail is on the left, then the skewness is leftward or negative. Right skewness is common when a variable is bounded on the left but unbounded on the right. For example, durations (response time, time to failure) typically have right skewness since they cannot take values less than zero; many financial variables (income, wealth, prices) typically have right skewness since they rarely take values less than zero; and adult body weight has right skewness since most people are closer to the lower limit than to the upper limit of viable body weight. Left skewness is less common in practice, but it can occur when a variable tends to be closer to its maximum than its minimum value. For example, scores on an easy

exam are likely to have left skewness, with most scores close to 100% and lower scores tailing off to the left. Well-known right-skewed distributions include the Poisson, chi-square, exponential, lognormal, and gamma distributions. I am not aware of any widely used distributions that always have left skewness, but there are several distributions that can have either right or left skew depending on their parameters. Such ambidextrous distributions include the binomial and the beta.

Mathematically, skewness is usually measured by the third standardized moment $E((X - \mu)/\sigma)^3)$, where X is a random variable with mean μ and standard deviation σ. The third standardized moment can take any positive or negative value, although in practical settings it rarely exceeds 2 or 3 in absolute value. Because it involves cubed values, the third standardized moment is sensitive to ►outliers (Kim and White 2004), and it can even be undefined for heavy-tailed distributions such as the Cauchy density or the Pareto density with a shape parameter of 3. When the third standardized moment is finite, it is zero for symmetric distributions, although a value of zero does not necessarily mean that the distribution is symmetric (Ord 1968; Johnson and Kotz 1970, p. 253). To estimate the third standardized moment from a sample of n observations, a biased but simple estimator is the third sample moment $1/n \sum((x - \bar{x})/s)^3$, where $\bar{x}$ is the sample mean and s is the sample standard deviation. An unbiased estimator is the third k statistic, which is obtained by taking the third sample moment and replacing $1/n$ with the quantity $n/((n-1)(n-2))$ (Rose and Smith 2002).

Although the third standardized moment is far and away the most popular definition of skewness, alternative definitions have been proposed (MacGillivray 1986). The leading alternatives are bounded by −1 and +1, and are zero for symmetric distributions, although again a value of zero does not guarantee symmetry. One alternative is Bowley's (1920) quartile formula for skew: $((q_3 - m) - (m - q_1))/(q_3 - q_1)$, or more simply $(q_1 + q_3 - 2m)/(q_3 - q_1)$, where m is the median and q_1 and q_3 are the first (or left) and third (or right) quartiles. Bowley's skew focuses on the part of the distribution that fits in between the quartiles: if the right quartile is further from the median than is the left quartile, then Bowley's skew is positive; if the left quartile is further from the median than the right quartile, then Bowley's skew is negative. Because it doesn't cube any values and doesn't use any values more extreme than the quartiles, Bowley's skew is more robust to outliers than is the conventional third-moment formula (Kim and White 2004). But the quantities in Bowley's formula are arbitrary: instead of the left and right quartiles – i.e., the 25th and 75th percentiles – Bowley could just as plausibly have used the 20th and 80th percentiles, the 10th and 90th percentiles, or more generally the 100pth and 100$(1 - p)$th percentiles $F^{-1}(p)$ and $F^{-1}(1 - p)$. Substituting these last two expressions into Bowley's formula, Hinkley (1975) proposed the generalized skewness formula $(F^{-1}(1-p)+F^{-1}(p)-2m)/(F^{-1}(1-p)-F^{-1}(p))$, which is a function of high and low percentiles defined by p. Since it is not clear what value of p is most appropriate, Groeneveld and Meeden (1984) averaged Hinkley's formula across all ps from 0 to 0.5. Groeneveld and Meeden's average was $(\mu - m)/E|X - m|$, which is similar to an old skewness formula that is attributed to Pearson: $(\mu - m)/\sigma$ (Yule 1911).

The Pearson and Groeneveld–Meeden formulas are consistent with a widely taught rule of thumb claiming that the skewness determines the relative positions of the median and mean. According to this rule, in a distribution with positive skew the mean lies to the right of the median, and in a distribution with negative skew the mean lies to the left of the median. If we define skewness using the Pearson or Groeneveld–Meeden formulas, this rule is self-evident: since the numerator of both formulas is simply the difference between the mean and the median, both will give positive skew when the mean is greater than the median, and negative skew when the situation is reversed. But if we define skewness more conventionally, using the third standardized moment, the rule of thumb can fail. Violations of the rule are rare for continuous variables, but common for discrete variables (von Hippel 2005). A simple discrete violation is the ►binomial distribution with $n = 10$ and $\pi = 0.09$ (cf. Lesser 2005). In this distribution, the mean 0.9 is left of the median 1, but the skewness as defined by the third standardized moment is positive, at 0.906, and the distribution, with its long right tail, looks like a textbook example of positive skew. Examples like this one argue against using the Pearson, Groeneveld-Meeden, or Bowley formulas, all of which yield a negative value for this clearly right-skewed distribution. Most versions of Hinkley's skew also contradict intuition here: Hinkley's skew is negative for $0.5 > p > 0.225$, zero for $0.225 \geq p > 0.054$, and doesn't become positive until $p \leq 0.054$.

Since many statistical inferences assume that variables are symmetrically or even normally distributed, those inferences can be inaccurate if applied to a variable that is skewed. Inferences grow more accurate as the sample size grows, with the required sample size depending on the amount of skew and the desired level of accuracy. A useful rule states that, if you are using the normal or t distribution to calculate a nominal 95% confidence interval for the mean of a skewed variable, the interval will have at least 94% coverage if the sample size is at least 25 times the absolute value of the (third-moment) skew (Cochran

1977; Boos and Hughes-Oliver 2000). For example, a sample of 50 observations should be plenty even if the skew is as large as 2 (or −2).

In order to use statistical techniques that assume symmetry, researchers sometimes transform a variable to reduce its skew (von Hippel 2003). The most common transformations for reducing positive skew are the logarithm and the square root, and a much broader family of skew-reducing transformations has been defined (Box and Cox 1964). But reducing skew has costs as well as benefits. A transformed variable can be hard to interpret, and conclusions about the transformed variable may not apply to the original variable before transformation (Levin et al. 1996). In addition, transformation can change the shape of relationships among variables; for example, if X is right-skewed and has a linear relationship with Y, then the square root of X, although less skewed, will have a curved relationship with Y (von Hippel 2010). In short, skew reduction is rarely by itself a sufficient reason to transform a variable. Skew should be treated as an important characteristic of the variable, not just a nuisance to be eliminated.

Cross References

►Box–Cox Transformation
►Heavy-Tailed Distributions
►Mean Median and Mode
►Mean, Median, Mode: An Introduction
►Normality Tests
►Omnibus Test for Departures from Normality

References and Further Reading

Boos DD, Hughes-Oliver JM (2000) How large does n have to be for Z and t intervals? Am Stat 54(2):121–128
Bowley AL (1920) Elements of statistics. Scribner, New York
Box GEP, Cox D (1964) An analysis of transformations. J R Stat Soc B 26(2):211–252
Cochran WG (1977) Sampling techniques. Wiley, New York
Groeneveld RA (1986) Skewness for the Weibull family. Stat Neerl 40:135–140
Groeneveld RA, Meeden G (1984) Measuring skewness and kurtosis. Stat 33:391–399
Hinkley DV (1975) On power transformations to symmetry. Biometrika 62:101–111
Johnson NL, Kotz S (1970) Continuous univariate distributions 1. Houghton Mifflin, Boston
Kim TH, White H (2004) On more robust estimation of skewness and kurtosis. Finance Res Lett 1(1):56–73
Lesser LM (2005) Letter to the editor [comment on von Hippel (2005)]. J Stat Educ 13(2) http://www.amstat.org/publications/jse/v13n3/lesser_letter.html
Levin A, Liukkonen J, Levine DW (1996) Equivalent inference using transformations. Commun Stat Theor Meth 25(5):1059–1072
MacGillivray HL (1986) Skewness and asymmetry: measures and orderings. Ann Stat 14(3):994–1011
Ord JK (1968) The discrete student's t distribution. Ann Math Stat 39:1513–1516
Rose C, Smith M (2002) Mathematical statistics with mathematica. Springer, New York
Sato M (1997) Some remarks on the mean, median, mode and skewness. Aust J Stat 39(2):219–224
von Hippel PT (2003) Normalization. In: Lewis-Beck M, Bryman A, Liao TF (eds) Encyclopedia of social science research methods. Sage, Thousand Oaks
von Hippel PT (2005) Mean, median, and skew: correcting a textbook rule. J Stat Edu 13 (2) www.amstat.org/publications/jse/v13n2/vonhippel.html
von Hippel PT (2010) How to impute skewed variables under a normal model. Unpublished manuscript, under review
Yule GU (1911) Introduction to the theory of statistics. Griffith, London

Skew-Normal Distribution

Adelchi Azzalini
Professor of Statistics
University of Padua, Padua, Italy

In its simplest reading, the term "skew-normal" refers to a family of continuous probability distributions on the real line having density function of form

$$\varphi(z;\alpha) = 2\,\varphi(z)\,\Phi(\alpha z), \qquad (-\infty < z < \infty), \qquad (1)$$

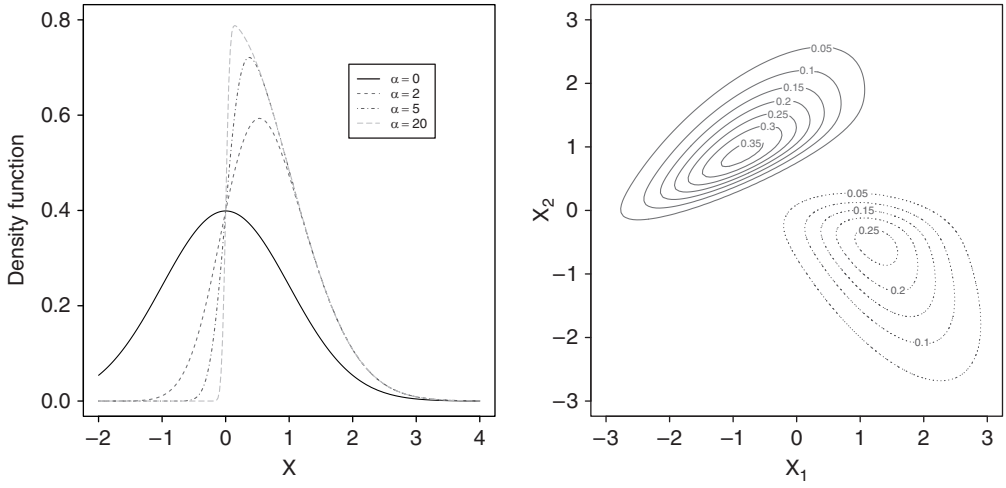
where $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the $\mathrm{N}(0,1)$ density and cumulative distribution function, respectively, and α is a real parameter which regulates the shape of the density. The fact that (1) integrates to 1 holds by a more general result, given by Azzalini (1985), where φ and Φ are replaced by analogous functions for any choice of two distributions symmetric around 0.

It is immediate that the choice $\alpha = 0$ lends the $\mathrm{N}(0,1)$ distribution, and that, if Z is a random variable with density (1), denoted $Z \sim SN(\alpha)$, then $-Z \sim SN(-\alpha)$. Figure 1a displays $\varphi(z;\alpha)$ for a few choices of α; only positive values of this parameter are considered, because of the property just stated.

An interesting property is that $Z^2 \sim \chi^2_1$, if $Z \sim SN(\alpha)$, irrespectively of α. The ►moment generating function of Z is

$$M(t) = 2\exp(t^2/2)\,\Phi(\delta t), \quad \delta = \alpha/\sqrt{1+\alpha^2}, \qquad (2)$$

and from $M(t)$ it is simple to obtain the mean, the variance, the index of skewness and the index of kurtosis,



Skew-Normal Distribution. Fig. 1 Some examples of skew-normal density function, for the scalar case (*left*) and for the bivariate case in the form of contour level plots (*right*)

which are

$$\mu_\alpha = \sqrt{\frac{2}{\pi}}\,\delta, \quad \sigma_\alpha^2 = 1 - \mu_\alpha^2,$$
$$\gamma_1 = \frac{4-\pi}{2}\frac{\mu_\alpha^3}{\sigma_\alpha^3}, \quad \gamma_2 = 2(\pi-3)\,\frac{\mu_\alpha^4}{\sigma_\alpha^4} \tag{3}$$

respectively. Multiplication of $M(t)$ by $\exp(t^2/2)$ shows another interesting property: if $U \sim \mathrm{N}(0,1)$ independent of Z, then $(Z+U)/\sqrt{2} \sim SN\left(\alpha/\sqrt{2+\alpha^2}\right)$. Additional facts about this distribution are given by Azzalini; Azzalini (1985; 1986), Henze (1986) and Chiogna (1998).

For practical statistical work, we need to consider the three-parameter distribution of $Y = \xi + \omega Z$, where ξ and ω are a location and a scale parameter, respectively ($\omega > 0$). Extension of the above results to the distribution of Y is immediate.

For the d-dimensional version of (1) we introduce directly a location parameter $\xi \in \mathbb{R}^d$ and a scale $d \times d$ matrix Ω which is symmetric and positive definite, and we denote by ω a $d \times d$ diagonal matrix formed by the square roots of the diagonal elements of Ω. The density function of the multivariate skew-normal distribution at x is

$$2\,\varphi_d(x-\xi;\Omega)\,\Phi\left(\alpha^\top \omega^{-1}(x-\xi)\right), \quad (x \in \mathbb{R}^d), \tag{4}$$

where $\varphi_d(x;\Omega)$ denotes the $\mathrm{N}_d(0,\Omega)$ density function, and the shape parameter α is a vector in $\mathbb{R}^d$. Figure 1b displays function 4 for two choices of the parameter set (ξ, Ω, α). Initial results on this distribution have been obtained by Azzalini and Dalla Valle (1996) and by Azzalini and Capitanio (1999).

The multivariate skew-normal distribution enjoys a number of formal properties. If Y is a d-dimensional random variable with density (4), its moment generating function is

$$M(t) = \exp\left(\xi^\top t + \frac{1}{2}t^\top \Omega t\right)\Phi(\delta^\top \omega t),$$
$$\delta = \frac{1}{(1+\alpha^\top \bar{\Omega}\alpha)^{1/2}}\bar{\Omega}\alpha \tag{5}$$

where $\bar{\Omega} = \omega^{-1}\Omega\omega^{-1}$ is the correlation matrix associated to Ω. From $M(t)$ one obtains that

$$\mathbb{E}\{Y\} = \xi + \sqrt{\frac{2}{\pi}}\,\omega\delta, \quad var\{Y\} = \Omega - \frac{2}{\pi}\omega\delta\delta^\top\omega,$$

while the marginal indices of skewness and kurtosis are computed by applying expressions γ_1 and γ_2 in (3) to each component of δ. Another result derived from (5) is that an affine transformation $a + A\,Y$, where $a \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times d}$, is still of type (4), with suitably modified dimension and parameters. This fact implies closure of this family of distributions with respect to marginalization. Closure of the class under conditioning holds if one extends the class by inserting an additional parameter in the argument of Φ in (4), and adapting the normalizing constant correspondingly; for details on this extended class, see Arnold and Beaver (2000) and Capitanio et al. (2003).

The *chi-square distribution* property stated for the scalar case extends substantially in the multivariate case. If Y has density (4) with $\xi = 0$, then a quadratic form $Y^\top AY$, where A is a symmetric $d \times d$ matrix, has the same distribution of $X^\top AX$ where $X \sim \mathrm{N}_d(0,\Omega)$; for instance

$Y^\top \Omega^{-1} Y \sim \chi^2_d$. This distributional result can be obtained from first principles, but it is mosty simply derived as a special case of the distributional invariance property of the family of *skew-symmetric distributions*, of which the skew-normal distribution is a special instance. According to this property, the distribution of $T(Y)$ is the same of $T(X)$ for any function T, possibly multi-valued, such that $T(x) = T(-x)$ for all $x \in \mathbb{R}^d$.

An attractive feature of this distribution is that it admits various *stochastics representations*, which are relevant for random number generation and also for supporting the adoption of this distribution in statistical modelling work. Here we restrict ourselves to one of these representations, which is related to a *selective sampling* mechanism: if

$$\begin{pmatrix} X_0 \\ X \end{pmatrix} \sim N_{1+d}(0, \Omega^*), \quad \Omega^* = \begin{pmatrix} 1 & \delta^\top \omega \\ \omega\delta & \Omega \end{pmatrix} > 0,$$

where X_0 and X have dimension 1 and d, respectively, then

$$Y = \xi + \begin{cases} X & \text{if} \quad X_0 > 0, \\ -X & \text{otherwise} \end{cases}$$

has density function (4) where $\alpha = (1 - \delta^\top \bar{\Omega}^{-1} \delta)^{-1/2} \bar{\Omega}^{-1} \delta$.

Additional information on the skew-normal distribution and related areas is presented in the review paper of Azzalini (2005), followed by a set of comments of Marc Genton, and rejoinder of the author. Themes considered include: additional properties and types of stochastic representation, aspects of statistical inference, historical development, extensions to skew-ellitical and skew-symmetric type of distributions, and connections with various application areas.

About the Author

Professor Azzalini is an elected fellow of the International Statistical Institute, and a member of various scientific societies. As a member of the "Bernoulli Society", he served as a member of the Council of the Society (1991–94) and as a chairman of the European Regional Committee (2006–2008). He also served as an associate editor for some scholarly journals (*Applied Statisics*, *Scandinavian J. Statistics*, *Metron*). Currently he is on the Advisory Board of *Metron*.

Editor's note: Professor Azzalini was the first to thoroughly set the foundations and provided systematic treatment of skew-normal distribution (in 1985 and 1986) and introduced the multivariate skew-normal distribution, with Dalla Valle, in 1996.

Cross References

►Chi-Square Distribution
►Normal Distribution, Univariate
►Skewness
►Skew-Symmetric Families of Distributions

References and Further Reading

Arnold BC, Beaver RJ (2000) Hidden truncation models. Sankhyā A 62(1):22–35
Azzalini A (1985) A class of distributions which includes the normal ones. Scand J Stat 12:171–178
Azzalini A (1986) Further results on a class of distributions which includes the normal ones. Statistica 46(2):199–208
Azzalini A (2005) The skew-normal distribution and related multivariate families (with discussion) Scand J Stat 32:159–188 (C/R 189–200)
Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. J R Stat SocB 61(3):579–602 Full version of the paper at http://arXiv.org (No. 0911.2093)
Azzalini A, Dalla Valle A (1996) The multivariate skew-normal distribution. Biometrika 83:715–726
Capitanio A, Azzalini A, Stanghellini E (2003) Graphical models for skew-normal variates. Scand J Statist 30:129–144
Chiogna M (1998) Some results on the scalar skew-normal distribution. J Ital Stat Soc 7:1–13
Henze N (1986) A probabilistic representation of the 'skew-normal' distribution. Scand J Stat 13:271–275

Skew-Symmetric Families of Distributions

Adelchi Azzalini
Professor of Statistics
University of Padua, Padua, Italy

The term 'skew-symmetric distributions' refers to the construction of a continuous probability distribution obtained by applying a certain form of perturbation to a symmetric density function.

To be more specific, a concept of *symmetric distribution* must be adopted first, since in the multivariate setting various forms of symmetry have been introduced. The variant used in this context is the one of *central symmetry*, a natural extension of the traditional one-dimensional form to the d-dimensional case: if f_0 is a density function on $\mathbb{R}^d$ and ξ is a point of $\mathbb{R}^d$, central symmetry around ξ requires that $f_0(t - \xi) = f_0(-t - \xi)$ for all $t \in \mathbb{R}^d$, ignoring sets of 0 probability. To avoid notational complications, we shall concentrate on the case with $\xi = 0$; it is immediate to rephrase what follows in the case of general ξ, which simply amounts to a shift of the location of the distribution.

If f_0 is a probability density function on $\mathbb{R}^d$ centrally symmetric around 0, there are two largely equivalent expressions to build skew-symmetric densities. For the first one, introduce a one-dimensional continuous distribution function G such that $G(-x) = 1 - G(x)$ for all $x \in \mathbb{R}$, and $w(\cdot)$ a real-valued function on $\mathbb{R}^d$ such that $w(-t) = -w(t)$ for all $t \in \mathbb{R}^d$. Then it can be shown that

$$f(t) = 2 f_0(t)\, G\{w(t)\} \tag{1}$$

is a density function on $\mathbb{R}^d$. Notice that in general $G\{w(t)\}$ is not a probability distribution. In the second type of formulation, consider a function $\pi(t)$ such that $0 \le \pi(t) \le 1$ and $\pi(t) + \pi(-t) = 1$ for all $t \in \mathbb{R}^d$, which leads to the density function

$$f(t) = 2 f_0(t)\, \pi(t)\,. \tag{2}$$

Formulations (1) and (2) have been obtained independently by Azzalini and Capitanio (2003) and by Wang et al. (2004), who adopted the term 'skew-symmetric distribution'. Each of the two forms has its advantages. Any expression of type $G\{w(t)\}$ in (1) automatically satisfies the requirements for $\pi(t)$ in (2), but it is not unique: there are several forms $G\{w(t)\}$ corresponding to the same $\pi(t)$. On the other hand, any $\pi(t)$ can be written in the form $G\{w(t)\}$. Hence the two sets of distributions coincide.

The proof that (1) and (2) are proper density functions is exceptionally simple. The argument below refers to (1) in the univariate case; the multivariate case is essentially the same with only a minor technical complication. If Y is a random variable with density function f_0 and X is an independent variable with distribution function G, then $w(Y)$ is symmetrically distributed around 0 and

$$\begin{aligned}\frac{1}{2} &= \mathbb{P}\{X - w(Y) \le 0\} = \mathbb{E}_Y\{\mathbb{P}\{X \le w(y) | Y = y\}\} \\ &= \int_{\mathbb{R}^d} G\{w(y)\} f_0(y) \mathrm{d}y\,.\end{aligned}$$

This proof also shows the intimate connection of this formulation with a selective sampling mechanism where a value Y sampled from f_0 is retained with probability $G\{w(t)\}$, and it is otherwise rejected. A refinement of this scheme says that

$$Z = \begin{cases} Y & \text{if } X \le w(Y), \\ -Y & \text{otherwise} \end{cases} \tag{3}$$

has density (1). Since (3) avoids rejection of samples, it is well suited for random numbers generation.

In spite of their name, skew-symmetric distributions are not *per se* linked to any idea of ▶skewness. The name is due to the historical connection with the ▶*skew-normal distribution*, which has been the first construction of this type. The skew-normal density function is

$$2\, \varphi_d(y; \Omega)\, \Phi(\eta^\top y), \qquad (y \in \mathbb{R}^d), \tag{4}$$

where $\varphi_d(y; \Omega)$ is the $\mathrm{N}_d(0, \Omega)$ density function, Φ is the $\mathrm{N}(0,1)$ distribution function and η is a vector parameter. This density is of type (1) with $f_0(y) = \varphi_d(y; \Omega)$ and $G\{w(y)\} = \Phi(\eta^\top y)$. In this case the perturbation of the original density φ_d does indeed lead to an asymmetric density, as it typically occurs when $w(y)$ is a linear function.

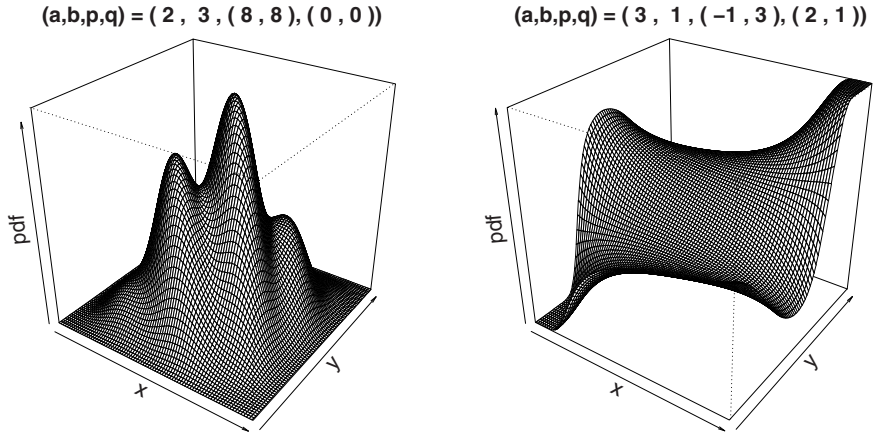
To illustrate visually the flexibility which can be achieved by the perturbation mechanism, consider f_0 to be the product of two symmetric Beta densities of parameters (a, a) and (b, b), say, both shifted and scaled to the interval $(-1, 1)$, G equal to the standard logistic distribution function and

$$w(y) = \frac{\sin(p_1 y_1 + p_2 y_2)}{1 + \cos(q_1 y_1 + q_2 y_2)}, \qquad y = (y_1, y_2)^\top \in (-1, 1)^2$$

where $p = (p_1, p_2)$ and $q = (q_1, q_2)$ are additional parameters. Figure 1 displays a few of the shapes produced with various choices of the parameters a, b, p, q. These skew-symmetric densities do not exhibit any obvious sign of skewness.

An important implication of representation (3) is the following property of distributional invariance: if Y has density f_0 and Z has density (1), then $T(Z)$ and $T(Y)$ have the same distribution for any function $T(\cdot)$ from $\mathbb{R}^d$ to $\mathbb{R}^q$ which is even, in the sense that $T(z) = T(-z)$ for all $z \in \mathbb{R}^d$. For instance, if Z has skew-normal distribution (4), then a quadratic form $T(Z) = Z^\top A Z$ has the same distribution of $T(Y) = Y^\top A Y$ when $Y \sim N_d(0, \Omega)$, for any symmetric matrix A; a further specialization says that $Z^\top \Omega^{-1} Z \sim \chi^2_d$. Other results on skew-elliptical distributions have been given by Arellano-Valle et al. (2006) and Umbach (2008).

An important subset of the skew-symmetric distributions occurs if f_0 in (1) or (2) is an *elliptically contoured density*, or briefly an *elliptical density*, in which case we obtain a skew-elliptical distribution. In fact, this subset was the first considered, in chronological order, starting from the skew-normal distribution, and the formulation evolved via a sequence of successive generalizations. This development is visible in the following sequence of papers, to be complemented with those already quoted: Azzalini and Capitanio (1999), Branco and Dey (2001), Genton and



Skew-Symmetric Families of Distributions. Fig. 1 Densities obtained by perturbation of the product of two symmetric Beta densities for some choices of the parameters a,b,p,q

Loperfido (2005) and the collection of papers in the book edited by Genton (2004).

About the Author

For biography *see* the entry ►Skew-Normal Distribution.

Cross References

►Beta Distribution
►Logistic Distribution
►Skewness
►Skew-Normal Distribution

References and Further Reading

Arellano-Valle RB, Branco MD, Genton MG (2006) A unified view on skewed distributions arising from selections. Canad J Stat 34:581–601

Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew normal distribution. J R Stat Soc B 61(3):579–602. Full version of the paper at http://arXiv.org (No. 0911.2093)

Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t* distribution. J R Stat Soc B 65(2):367–389. Full version of the paper at http://arXiv.org (No. 0911.2342)

Branco MD, Dey DK (2001) A general class of multivariate skew-elliptical distributions. J Multivariate Anal 79(1):99–113

Genton MG (ed) (2004) Skew-elliptical distributions and their applications: a journey beyond normality. Chapman & Hall/CRC Press, Boca Raton, FL

Genton MG, Loperfido N (2005) Generalized skew-elliptical distributions and their quadratic forms. Ann Inst Stat Math 57: 389–401

Umbach D (2008) Some moment relationships for multivariate skew-symmetric distributions. Stat Probab Lett 78(12): 1619–1623

Wang J, Boyer J, Genton MG (2004) A skew-symmetric representation of multivariate distributions. Stat Sinica 14:1259–1270

Small Area Estimation

Danny Pfeffermann
Professor Emeritus
Hebrew University of Jerusalem, Jerusalem, Israel
Professor
University of Southampton, Southampton, UK

Introduction

Over the past three decades there is a growing demand in many countries for reliable estimates of small domain parameters such as means, counts, proportions or quantiles. Common examples include the estimation of unemployment rates, proportions of people under poverty, disease incidence and use of illicit drugs. These estimates are used for fund allocations, new social or health programs, and more generally, for short and long term planning. Recently, small area estimates are employed for testing, the administrative records used for modern censuses (see ►Census). Although commonly known as "small area estimation" (SAE), the domain of studies may actually consist of socio-demographic subgroups as defined, for example, by gender, age and race, or the intersection of such domains with geographical location.

The problem of SAE is that the sample sizes in at least some of the domains of study are very small, and often there are no samples available for many or even most of these domains. As a result, the direct estimates obtained from the survey are unreliable (large, unacceptable variances), and no direct survey estimates can be computed for areas with no samples. SAE methodology addresses therefore the following two major problems:

1. How to obtain reliable estimates for each of these areas.
2. How to assess the error of the estimators (MSE, confidence intervals, etc.).

Notice in this regard that even if direct survey estimates can be used for areas with samples, no design-based methodology exists for estimating the quantities of interest in areas with no samples. The term "Design-based inference" refers to inference based on the randomization distribution over all the samples possibly selected from the finite population under study, with the population values considered as fixed numbers. Note also that the sample sizes in the various areas are random, unless when some of the domains of study are defined as strata and samples of fixed sizes are taken in these domains.

In what follows I describe briefly some of the basic methods used for SAE, assuming, for simplicity, that the sample is selected by simple random sampling. More advanced methods and related theory, with many examples and references can be found in the book of Rao (2003) and the review papers by Ghosh and Rao (1994), Rao (1999), Pfeffermann (2002), and Rao (2005). See also Chaps. 31 and 32 in the new Handbook of Statistics, 29B (eds. Pfeffermann and Rao 2009).

Design-Based Methods

Let Y define the characteristic of interest and denote by y_{ij} the outcome value for unit j belonging to area i, $i = 1,\ldots,M$; $j = 1,\ldots,N_i$, where N_i is the area size. Let $s = s_1 \cup \cdots \cup s_m$ denote the sample, where s_i of size n_i is the sample observed for area i. Suppose that it is required to estimate the true area mean $\overline{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i$. If no auxiliary information is available, the *direct* design unbiased estimator and its design variance over the *randomization distribution* (the distribution induced by the random selection of the sample with the population values held fixed), are given by

$$\hat{\overline{Y}}_i = \sum_{j=1}^{n_i} y_{ij}/n_i; \quad Var_D\left[\hat{\overline{Y}}_i | n_i\right] = \left(S_i^2/n_i\right)\left[1-(n_i/N_i)\right] = S_i^{*2}, \tag{1}$$

where $S_i^2 = \sum_{k=1}^{N_i}(y_{ik}-\overline{Y}_i)^2/(N_i-1)$. Clearly, for small n_i the variance will be large, unless the variability of the *y-values* is sufficiently small. Suppose, however, that values x_{ij} of p concomitant variables $\mathbf{x}_1,\ldots,\mathbf{x}_p$ are measured for each of the sample units and that the area means $\overline{X}_i = \sum_{k=1}^{N_i} x_{ik}/N_i$ are likewise known. Such information may be obtained from a recent census or some other administrative records. In this case, a more efficient design-based estimator is the regression estimator,

$$\hat{\overline{Y}}_{i,reg} = \overline{y}_i + (\overline{X}_i - \overline{x}_i)'\beta_i; \quad Var_D(\overline{y}_{reg,i}|n_i) = S_i^{*2}\left(1-R_i^2\right), \tag{2}$$

where $\overline{y}_i$ and $\overline{x}_i$ are the sample means of Y and X in area i, and β_i and R_i are correspondingly the vector of regression coefficients and the multiple correlation coefficient between Y and $\mathbf{x}_1,\ldots,\mathbf{x}_p$ in area i. Thus, by use of the concomitant variables, the variance is reduced by the factor $\left(1-R_i^2\right)$, illustrating the importance of using auxiliary information with good prediction power for SAE.

In practice, the coefficients β_i are unknown. Replacing β_i by its ordinary least square estimator from the sample s_i may not be effective in the case of a small sample size. If, however, the regression relationships are "similar" across the areas and assuming $x_{ij,1} = 1$ for all (i,j), a more stable estimator is the *synthetic regression* estimator,

$$\hat{\overline{Y}}_{i,syn} = \sum_{j=1}^{N_i} \hat{y}_{ik}/N_i = \overline{X}_i'\hat{B}, \tag{3}$$

where $\hat{y}_{ik} = x_{ik}'\hat{B}$ and $\hat{B} = \left[\sum_{i,j\in S} x_{ij}x_{ij}'\right]^{-1}\sum_{i,j\in S} x_{ij}y_{ij}$ is the ordinary least squares estimator computed from all the sample data. The prominent advantage of synthetic estimation is the substantial variance reduction since the estimator uses all the sample data, but it can lead to severe biases if the regression relationships differ between the areas.

An approximately design-unbiased estimator is obtained by replacing the synthetic estimator by the GREC estimator,

$$\hat{\overline{Y}}_{i,greg} = \sum_{k=1}^{N_i} \hat{y}_{ik}/N_i + \sum_{j\in S_i}(y_{ij}-\hat{y}_{ij})/n_i. \tag{4}$$

However, this estimator may again be very unstable in small samples. The choice between the synthetic estimator and the GREG is therefore a trade off between bias and variance. A compromise is achieved by using a composite estimator of the form,

$$\hat{\overline{Y}}_{i,com} = \alpha_i\hat{\overline{Y}}_{i,greg} + (1-\alpha_i)\hat{\overline{Y}}_{i,syn}, \tag{5}$$

but there is no principled theory of how to determine the coefficients α_i.

Design-based estimators are basically model free but the requirement for approximate design-unbiasedness generally yields estimators with large variance due to the small sample sizes. The construction of confidence intervals requires large sample normality assumptions, which do not generally hold in SAE problems. No design-based theory exists for estimation in areas with no samples.

Model-Dependent Estimators

In view of the problems underlying the use of design-based methods, it is common practice in many applications to use instead statistical models that define how to "borrow strength" from other areas and/or over time in case of repeated surveys. Let θ_i define the parameter of interest in area i, $i = 1, \ldots, M$, and let y_i, x_i denote the data observed for this area. When the only available information is at the area level, y_i is typically the direct estimator of θ_i and $\mathbf{x}_i$ is a vector of area level covariates. When unit level information is available, y_i is a vector of individual outcomes and $\mathbf{x}_i$ is the corresponding matrix of individual covariate information.

A typical small area model consists of two parts: The first part models the distribution of $y_i|\theta_i; \Psi_{(1)}$. The second part models the distribution of $\theta_i|x_i; \Psi_{(2)}$ linking θ_i to the parameters in other areas and to the covariates. The (vector) parameters $\Psi_{(1)}$ and $\Psi_{(2)}$ are typically unknown and are estimated from all the available data $D(s) = \{y_i, x_i; 1, \ldots, m\}$. In what follows I define and discuss briefly three models in common use.

"Unit Level Random Effects Model"

The model, employed originally by Battese et al. (1988), assumes,

$$y_{ij} = x'_{ij}\beta + u_i + \varepsilon_{ij}, \tag{6}$$

where u_i and ε_{ij} are mutually independent error terms with zero means and variances σ_u^2 and σ_ε^2 respectively. The random term u_i represents the joint effect of area characteristics not accounted for by the concomitant variables. Under the model, the true small area means are $\overline{Y}_i = \overline{X}'_i\beta + u_i + \overline{\varepsilon}_i$, but since $\overline{\varepsilon}_i = \sum_{k=1}^{N_i} \varepsilon_{ik}/N_i \cong 0$ for large N_i, the target parameters are often defined as $\theta_i = \overline{X}'_i\beta + u_i$. For known variances $\left(\sigma_u^2, \sigma_\varepsilon^2\right)$, the Best Linear Unbiased Predictor (BLUP) of θ_i is,

$$\hat{\theta}_i = \gamma_i[\,\overline{y}_i + (\overline{X}_i - \overline{x}_i)'\hat{\beta}_{GLS}] + (1-\gamma_i)\overline{X}'_i\hat{\beta}_{GLS}, \tag{7}$$

where $\hat{\beta}_{GLS}$ is the generalized least square (GLS) estimator of β computed from all the observed data and $\gamma_i = \sigma_u^2/\left(\sigma_u^2 + \sigma_\varepsilon^2/n_i\right)$. For areas l with no samples, $\hat{\theta}_l = \overline{X}'_l\hat{\beta}_{GLS}$. Notice that unlike under the randomization distribution, the synthetic estimator $\overline{X}'_i\hat{\beta}_{GLS}$ is unbiased for θ_i under the model in the sense that $E(\overline{X}'_i\hat{\beta}_{GLS} - \theta_i) = 0$.

The BLUP $\hat{\theta}_i$ is also the Bayesian predictor (posterior mean) under normality of the error terms and a diffuse prior for β. In practice, however, the variances σ_u^2 and σ_ε^2 are seldom known. A Bayesian solution to this problem is to set prior distributions for the unknown variances and then compute the corresponding posterior mean and variance of $\theta_i|\{y_k, x_k; k \in s\}$ by aid of Markov Chain Monte Carlo (MCMC) simulations (see ▶Markov Chain Monte Carlo). The common procedure under the frequentist approach is to replace the unknown variances in the BLUP formula by standard variance components estimates like Maximum Likelihood Estimators (MLE), Restricted MLE (REML) or Analysis of Variance (ANOVA) estimators. The resulting predictors are known as the Empirical BLUP (EBLUP). See the references listed in the introduction for estimation of the MSE of the EBLUP under different methods of variance estimation.

"Area Level Random Effects Model"

This model is in broad use when the concomitant information is only at the area level. It was used originally by Fay and Herriot (1979) for predicting the mean income in geographical areas of less than 500 inhabitants. Denote by $\tilde{\theta}_i$ the direct sample estimator of θ_i. The model assumes that,

$$\tilde{\theta}_i = \theta_i + e_i; \quad \theta_i = x'_i\beta + u_i, \tag{8}$$

such that e_i represents the sampling error, assumed to have zero mean and known design variance $Var_D(e_i) = \sigma_{Di}^2$, $(= S_i^{*2}$ if $\tilde{\theta}_i = \overline{y}_i$, see Eq. 1). The model integrates therefore a model dependent random effect u_i and a sampling error e_i with the two errors being independent. The BLUP under this model is,

$$\hat{\theta}_i = \gamma_i\tilde{\theta}_i + (1-\gamma_i)x'_i\hat{\beta}_{GLS} = x'_i\hat{\beta}_{GLS} + \gamma_i\left(\tilde{\theta}_i - x'_i\hat{\beta}_{GLS}\right), \tag{9}$$

which again is a composite estimator with coefficient $\gamma_i = \sigma_u^2/\left(\sigma_{Di}^2 + \sigma_u^2\right)$. As with the unit level model, the variance σ_u^2 is usually unknown and is either assigned a prior distribution under the Bayesian approach, or is replaced by a sample estimate in (9), yielding the corresponding EBLUP predictor.

Unit Level Random Effects Model for Binary Data

The previous two models are for continuous measurements. Suppose now that y_{ij} is a binary variable taking the values 0 or 1. For example, $y_{ij} = 1$ if individual j in area i is unemployed (or suffers from a certain disease), and $y_{ij} = 0$ otherwise, such that $p_i = N_i^{-1}\sum_{k=1}^{N_i} y_{ik}$ is the true unemployment rate (true disease incidence). The following model is often used for predicting the proportions p_i:

$$\begin{aligned} y_{ij}|p_{ij} &\overset{indep.}{\sim} Bernoulli(p_{ij}) \\ \text{logit}(p_{ij}) &= \log[\,p_{ij}/(1-p_{ij})] = x'_{ij}\beta + u_i; \\ u_i &\overset{indep.}{\sim} N\left(0, \sigma_u^2\right), \end{aligned} \tag{10}$$

where as in (6), $\mathbf{x}_{ij}$ is a vector of concomitant values, β is a vector of fixed regression coefficients and u_i is a random effect representing the unexplained variability of the individual probabilities between the areas.

For this model there is no explicit expression for the predictor $\hat{p}_i$. Writing $p_i = N_i^{-1}\left[\sum_{j\in s_i} y_{ij} + \sum_{l\notin s_i} y_{il}\right]$, predicting p_i by its best predictor is equivalent to the prediction of the sum $\sum_{l\notin s_i} y_{il}$ of the missing observations. See Jiang et al. (2002) for the computation of the empirical best predictor and estimation of its MSE.

About the Author

Danny Pfeffermann is Professor of statistics at the Hebrew University of Jerusalem, Israel, and at the Southampton Statistical Sciences Research Institute, University of Southampton, UK. He has presented and published many articles on small area estimation in leading statistical journals and is teaching this topic regularly. He is consulting the Office for National Statistics in the UK and the Bureau of Labor Statistics in the USA on related problems. Professor Pfeffermann is past president of the Israel Statistical Association (2005–2007), an Elected Fellow of the American Statistical Association (1990), and an Elected member of the International Statistical Institute. He was Associate Editor of *Biometrika* and the *Journal of Statistical Planning and Inference* and is currently Associate Editor for *Survey Methodology*. Professor Pfeffermann has recently completed jointly with Professor C.R. Rao editing the new two-volume Handbook of Statistics on *Survey Samples*, published by North Holland (2009). He is the recipient of the Waksberg award for 2011.

Cross References

►Best Linear Unbiased Estimation in Linear Models
►Census
►Estimation
►Estimation: An Overview
►Inference Under Informative Probability Sampling
►Markov Chain Monte Carlo
►Non-probability Sampling Survey Methods
►Sample Survey Methods
►Social Statistics
►Superpopulation Models in Survey Sampling

References and Further Reading

Battese GE, Harter RM, Fuller WA (1988) An error component model for prediction of county crop areas using survey and satellite data. J Am Stat Assoc 83:28–36

Fay RE, Herriot R (1979) Estimates of income for small places: an application of James Stein procedures to census data. J Am Stat Assoc 74:269–277

Ghosh M, Rao JNK (1994) Small area estimation: an appraisal (with discussion). Stat Sci 9:65–93

Jiang J, Lahiri P, Wan SM (2002) A unified jackknife theory for empirical best prediction with M-estimation. Ann Stat 30: 1782–1810

Pfeffermann D (2002) Small area estimation – new developments and directions. Int Stat Rev 70:125–143

Pfeffermann D, Rao CR (eds) (2009) Handbook of statistics 29B. Sample surveys: inference and analysis. Elsevier, North Holland

Rao JNK (1999) Some recent advances in model-based small area estimation. Survey Methodol 25:175–186

Rao JNK (2003) Small area estimation. Wiley, New York

Rao JNK (2005) Inferential issues in small area estimation: some new developments. Stat Transit 7:513–526

Smoothing Splines

Grace Wahba
IJ Schoenberg-Hilldale Professor of Statistics, Professor of Biostatistics and Medical Informatics
University of Wisconsin, Madison, WI, USA

Univariate Smoothing Splines

Univariate smoothing splines were introduced by I.J. Schoenberg in the 40s, an early paper is (Schoenberg 1964). Given data $y_i = f(x(i)) + \epsilon_i, i = 1,\cdots,n$, where the ϵ_i are i.i.d samples from a zero mean Gaussian distribution and $0 < x(1) < \cdots < x(n) < 1$, the (univariate) polynomial smoothing spline is the solution to: find f in W_2^m to minimize

$$\frac{1}{n}\sum_{i=0}^{1}(y_i - f(x(i))^2 + \lambda \int_0^1 (f^{(m)}(x))^2 dx,$$

where W_2^m is the Sobolev space of functions with square integral mth derivative. The solution is well known to be a piecewise polynomial of degree $2m-1$ between each pair $\{x(j+1), x(j)\}, j = 1,\cdots,n-1$ and of degree $m-1$ in $[0, x(1)]$ and $[x(n), 1]$, and the pieces are joined so that the function has $2m-1$ continuous derivatives. Figure 1 illustrates the cubic smoothing spline ($m = 2$) and how it depends on the smoothing parameter λ. The dashed line in each of the three panels is the underlying function $f(x)$ used to generate the data. The observations y_i were generated as $y_i = f(x_i) + \epsilon_i$ where the ϵ_i were samples from a zero mean Gaussian distribution with common variance. The wiggly solid line in the top panel was obtained with a λ that is too small. The solid line in the middle panel has λ too large.

If λ had been even larger, the solid line would have tended to flatten out towards the least squares straight line best fitting the data. Note that linear functions are in the *null space* of the penalty functional $\int (f'')^2$, that is, their second derivatives are 0. In the third panel, λ has been chosen by the GCV (Generalized Cross Validation) method (Craven and Wahba 1979; Golub et al. 1979). Generalizations of the univariate smoothing spline include penalties that replace $(f^{(m)})^2$ with $(Lf)^2$, where Lf is a linear differential operator of order m, see Kimeldorf and Wahba (1971) and Ramsay and Silverman (1997). Code for smoothing splines is available in the R library http://cran.r-project.org, for example `pspline` and elsewhere. Other generalizations include replacing the residual sum of squares by the negative log likelihood for Bernoulli, Poisson or other members of the exponential family, by robust or quantile functionals, or by the so-called hinge function to get a Support Vector Machine (Cristianini and Shawe-Taylor 2000). In each case the solution will be a piecewise polynomial of the same form as before as a consequence of the so called representer theorems in Kimeldorf and Wahba (1971). Other tuning criteria are appropriate for the other functionals, for example the GACV (Xiang and Wahba 1996) for Bernoulli data.

Thin Plate Splines

Thin Plate Splines (TPS) appeared in French in 1975 (Duchon 1975) and were combined with the GCV for tuning in Wahba and Wendelberger (1980). The TPS of order 2 in two dimensions is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_1(i), x_2(i))^2 + \lambda J_{2,2}(f)$$

where $J_{2,2}$ is given by

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x_1x_1}^2 + 2f_{x_1x_2}^2 + f_{x_2x_2}^2 dx_1 dx_2.$$

In this case f is known to have a representation

$$f(x) = d_0 + d_1x_1 + d_2x_2 + \sum_{i=1}^{n} c_i E(x, x(i))$$

where

$$E(x, x(i)) = \|x - x(i)\|^2 log\|x - x(i)\|,$$

where $\|\cdot\|$ is the Euclidean norm.

There is no penalty on linear functions of the components (x_1, x_2) of the attribute vector (the "null space" of $J_{2,2}$). It is known that the c_i for the solution satisfy $\sum_{i=1}^{n} c_i = 0$, $\sum_{i=1}^{n} c_i x_1(i) = 0$ and $\sum_{i=1}^{n} c_i x_2(i) = 0$, and furthermore,

$$J_{2,2}(f) = \sum_{i,j=1,\cdots,n} c_i c_j E(x(i), x(j)).$$

The TPS is available for general d and for any m with $2m - d > 0$. The general TPS penalty functional in d dimensions and m derivatives is

$$J_{d,m} = \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\cdots\alpha_d!} \int_{-\infty}^{\infty}\cdots \int_{-\infty}^{\infty}\left(\frac{\partial^m f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}\right)^2 \prod_j dx_j.$$

See Wahba (1990). Note that there is no penalty on polynomials of degree less than m, so that the TPS with d greater than 3 or 4 is rarely attempted because of the very high dimensional null space of $J_{d,m}$. As λ tends to infinity, the solution tends to its best fit in the unpenalized space, and as λ tends to 0, the solution attempts to interpolate the data. Public codes in R containing TPS codes include `assist`, `fields`, `gss`, `mgcv`. Again, the residual sum of squares may be replaced by other functionals as in the univariate spline and the form of the solution will be the same.

Splines on the Sphere

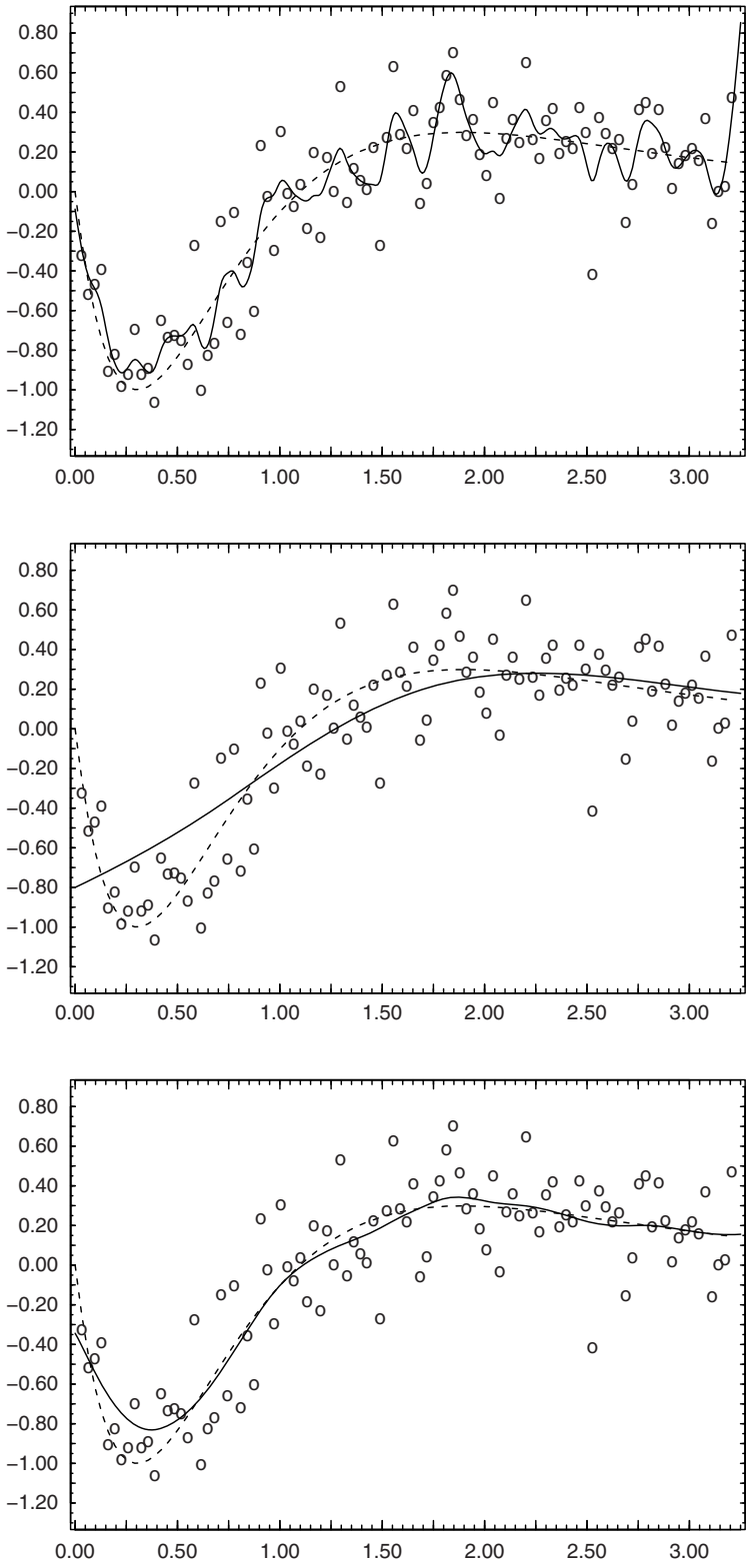
Splines on the sphere were proposed in Wahba; Wahba (1981; 1982). The penalty functional $J(f)$ for splines on the sphere is $J(f) = \int (\Delta)^{m/2} f$ where Δ is the (surface) Laplacian on the the (unit) sphere given by

$$\Delta f = \frac{1}{\cos^2\phi} f_{\theta\theta} + \frac{1}{\cos\phi}(\cos\phi f_\phi)_\phi$$

where θ is the longitude, $(0 \le \theta \le 2\pi)$ and ϕ is the latitude $\left(-\frac{\pi}{2} \le \phi \le \frac{\pi}{2}\right)$. Here we are using subscripts θ and ϕ to indicate derivatives with respect to θ and ϕ. Closed form expressions for the minimizer f are not in general available, but closed form expressions for a close approximation are, see Wahba; Wahba (1981; 1982).

Splines on Riemannian Manifolds

The splines we have mentioned above have penalty functionals associated with the Laplacian (note the form is different for the compact domain cases of splines on the unit interval and splines on the sphere, as opposed to the thin plate spline on the infinite plane). Splines on arbitrary compact Riemannian manifolds can be defined, implicitly or explicitly involving the eigenfunctions and eigenvalues of the m-iterated Laplacian, see Kim (1999), Pesenson (2004), Belkin and Niyogi (2004, Sect. 5.2).



Smoothing Splines. Fig. 1 Cubic smoothing spline with three different tuning parameters

Smoothing Spline ANOVA Models

Let $x = (x_1, \cdots, x_d)$, where $x_\alpha \in \mathcal{T}^{(\alpha)}, \alpha = 1, \cdots, d$ and $y_i = f(x(i)) + \epsilon_i, i = 1, \cdots, n$. where the ϵ_i are as before. The $\mathcal{T}^{(\alpha)}$ can be quite arbitrary domains. It is desired to estimate $f(x)$ for x in some region of interest contained in $\mathcal{T} = \mathcal{T}^{(1)} \otimes \cdots \otimes \mathcal{T}^{(d)}$. f is expanded as $f(x) = C + \sum_\alpha f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, t_\beta) + \cdots$, where the terms satisfy side conditions analogous to those in ordinary ANOVA which guarantee identifiability, and the decomposition is usually truncated at some point. The model is fit by minimizing the residual sum of squares plus

$$J_\lambda(f) = \sum_\alpha \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha<\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots.$$

The $J_\alpha, J_{\alpha\beta}, \cdots$ are composites of penalty functionals on the individual components and closed form expressions are available when they are available for the components. As before, the residual sum of squares may be replaced by the negative log likelihood and other functionals depending on y_i and $f(x(i))$. Details may be found in Wahba et al. (1995) and Gu (2002), and the R codes `assist` and `gss` are available to fit these models.

About the Author

Grace Wahba is IJ Schoenberg–Hilldale Professor of Statistics at the University of Wisconsin, where she has been a faculty member since 1967 after receiving her Ph.D. from Stanford in 1966. She is a Fellow of the International Statistical Institute, Institute of Mathematical Statistics, American Statistical Association, Society for Industrial and Applied Mathematics, and American Association for the Advancement of Science. She is also a Member of the National Academy of Sciences (2000). Dr. Wahba has an international reputation as an innovator in research on the theory and applications of "Spline Models for Observational Data." She was named the "Statistician of the Year" by the Chicago Chapter of ASA in 2004. Among many awards, she has been awarded the First Emanuel and Carol Parzen Prize for Statistical Innovation (1994), Committee of Presidents of Statistical Societies Elizabeth Scott Award (1996), Hilldale Award in the Physical Sciences, University of Wisconsin-Madison (2003). She has authored/coauthored about 130 papers and the book: *Spline Models for Observational Data* (SIAM, 1990). In 2007 she received an Honorary Doctorate from the University of Chicago. In 2009, Professor Wahba received the Gottfried E. Noether Senior Researcher Award for "outstanding contributions to the theory and applications of nonparametric statistics," and became the inaugural recipient of the Distinguished Alumni Award at Cornell University.

"Grace Wahba, the I. J. Schoenberg professor of statistics, University of Wisconsin-Madison, represents the very best of the modern synthesis of applied statistical, mathematical and computational science. Her most influential work has concerned problems in the estimation of curves and surfaces from large, high-dimensional data sets, such as occur frequently in geophysics." (Convocation Session, the University of Chicago Chronicle, Vol. 26, No 18, 2007).

Cross References

►Nonparametric Estimation
►Nonparametric Regression Using Kernel and Spline Methods
►Semiparametric Regression Models
►Smoothing Techniques

References and Further Reading

Belkin M, Niyogi P (2004) Semi-supervised learning on Riemannian manifolds. Mach Learn 56:209–239

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:377–403

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines. Cambridge University Press, Cambridge

Duchon J (1975) Fonctions splines et vecteurs aleatoires. Technical Report 213, Seminaire d'analyse numerique, universite scientifique et medicale, Grenoble

Golub G, Heath M, Wahba G (1979) Generalized cross validation as a method for choosing a good ridge parameter. Technometrics 21:215–224

Gu C (2002) Smoothing spline ANOVA models. Springer, New York

Kim P (1999) Splines on Riemannian manifolds and a proof of a conjecture by Wahba. Report, Department of Mathematics and Statistics, University of Guelph

Kimeldorf G, Wahba G (1971) Some results on Tchebycheffian spline functions. J Math Anal Appl 33:82–95

Pesenson I (2004) Variational splines on Riemannian manifolds with aplications to integral geometry. Adv Appl Math 33:548–572

Ramsay J, Silverman B (1997) Functional data analysis. Springer, New York

Schoenberg I (1964) Spline functions and the problem of graduation. In: Proceedings of the national academy sciences, vol 52. USA, pp 947–950

Wahba G (1981) Spline interpolation and smoothing on the sphere. SIAM J Sci Stat Comput 2:5–16

Wahba G (1982) Erratum: Spline interpolation and smoothing on the sphere. SIAM J Sci Stat Comput 3:385–386

Wahba G (1990) Spline models for observational data. In: CBMS-NSF regional conference series in applied mathematics, vol 59. SIAM, Philadelphia

Wahba G, Wendelberger J (1980) Some new mathematical methods for variational objective analysis using splines and cross-validation. Mon Weather Rev 108:1122–1145

Wahba G, Wang Y, Gu C, Klein R, Klein B (1995) Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. Ann Stat 23:1865–1895, Neyman Lecture

Xiang D, Wahba G (1996) A generalized approximate cross validation for smoothing splines with non-Gaussian data. Stat Sinica 6:675–692

Smoothing Techniques

Adrian W. Bowman
Professor
The University of Glasgow, Glasgow, UK

The idea of smoothing techniques is to identify trends, patterns, relationships and shapes in data without adopting strong assumptions about the specific nature of these. The one assumption that *is* made is that any trends and patterns are smooth. The term *nonparametric* is often used in the context of smoothing techniques to distinguish the methods from *parametric* modelling where specific distributional shapes (such as normal) or trends (such as linear) are adopted, leaving only some parameters to be estimated.

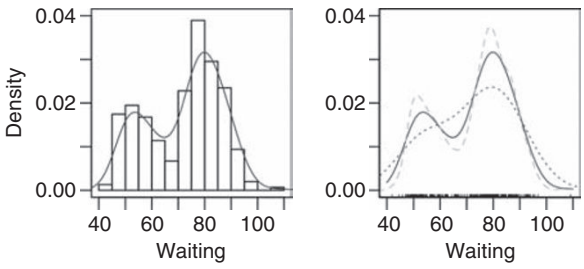
There are many situations where smoothing can be applied and many ways in which it can be implemented. This short article will give some simple examples in just two areas, namely density estimation and regression, and show how the latter techniques can be used in the context of wider regression modelling.

Density Estimation

The histogram is a time-honored way of presenting the shape of the variation in a set of data in graphical form. In fact, when the histogram is scaled to have area 1 it can be viewed as an estimate of the underlying density function $f(y)$. However, from that perspective it can be criticized because of its sharp edges. Instead of building the estimate from rectangular blocks, a kernel density estimate uses smooth functions, called kernels, in the estimate

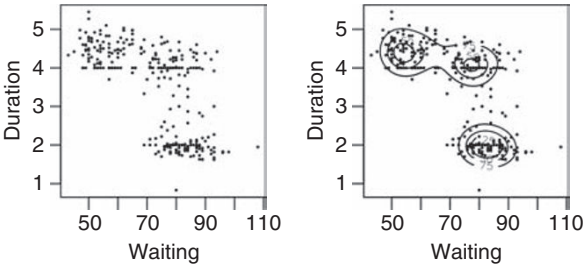
$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} w(y - y_i; h)$$

constructed from a sample of data $\{y_1, \ldots, y_n\}$. The kernel $w(.; h)$ might conveniently be chosen as a normal density function with mean 0 and standard deviation h. It remains to make a choice of the bandwidth, or smoothing parameter, h which is the equivalent of the bin width in a histogram. One effective means of doing this is to estimate the optimal value produced by a theoretical analysis. However, a very simple choice, which can also be very effective, is to use the optimal value associated with a normal distribution. That is the solution used in the examples below.



The left panel of the figure above shows a histogram of data on the waiting times between eruptions of the Old Faithful geyser in Yellowstone National Park. A kernel density estimate has been superimposed for comparison. The right panel shows the same density estimate along with estimates produced with larger (short dashed line) and smaller (long dashed line) degrees of smoothing.

These simple principles extend without difficulty to other types of data, simply by adopting a suitable form of kernel function. For example, the left hand panel below shows a plot of waiting time and the subsequent eruption time. The right panel shows the same plot with the contours of a density estimate superimposed. The kernel function here is simply a two-dimensional normal density function, with two smoothing parameters, one for each dimension. Although the scatterplot clearly shows a cluster of eruptions with shorter durations, the density estimate draws attention to the presence of two clusters in the eruptions with longer durations. In general, smoothing techniques such as density estimation can be helpful in identifying structure which is sometimes obscured by the variation in the data.



Silverman (1986) gave one of the first discussions of density estimation, with Scott (1992) focussing on the multivariate case. Wand and Jones (1995) is a source of very

useful theoretical analysis while Simonoff (1996) is particularly helpful in its broad coverage and extensive references.

Nonparametric Regression

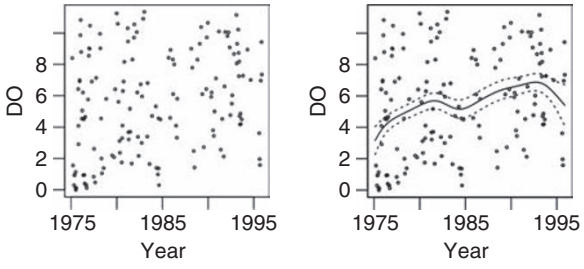
In the case of regression with a single covariate, smoothing techniques assume the model

$$y_i = m(x_i) + \varepsilon_i$$

for observed data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where the ε_i denote errors terms. The smooth function m can be estimated in a wide variety of ways. A kernel approach fits a standard model, such as a linear regression, but does so locally by solving the weighted least squares problem

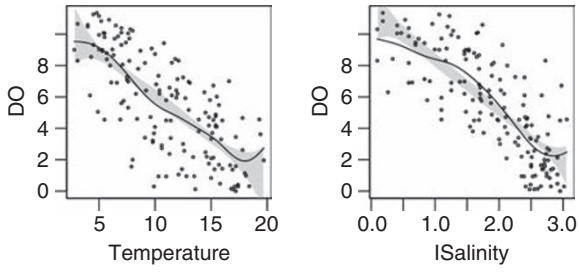
$$\min_{\alpha,\beta} \sum_{i=1}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 w(x_i - x; h).$$

The solution $\hat{\alpha}$ provides the estimate. However, there are many other approaches, many of these based on splines. For example, ▶smoothing splines arise as the solution of the problem $\min_m \sum_{i=1}^{n} \{y_i - m(x_i)\} + \lambda \int_a^b m''(x)dx$. Regression splines fit a model which is constructed as a linear combination of a set of basis functions while penalized splines place a smoothness penalty on these coefficients. This is a research topic with a large literature. Fan and Gijbels (1996) and Bowman and Azzalini (1997) describe the theory and applications of the kernel approach while Green and Silverman (1994) and Ruppert et al. (2003) focus on spline representations. In broad terms, these different methods have different approaches but a common aim. The method chosen for a particular problem can be a matter of convenience.



The panels above illustrate local linear smoothing on water quality data, expressed in dissolved oxygen (DO) at a particular sampling station on the River Clyde near Glasgow. The left hand panel shows DO against time in years, with little evidence of trend. The right hand plot adds a nonparametric regression curve which suggests that some trend may in fact be present, obscured by the large degree of variation in the data. The vector of fitted values from local linear, and indeed most other, forms of regression smoothing can be represented in vector–matrix form as $\hat{m} = Sy$, where S is an $n \times n$ smoothing matrix. This linear structure gives relatively easy access to standard errors and to the quantification of the level of smoothing through approximate degrees of freedom, by analogy with standard linear models. The right hand panel above has added two standard errors on either side of the nonparametric regression line, to indicate the precision of estimation. Bias is an inevitable consequence of smoothing so this cannot be strictly interpreted as a confidence band.

The two panels below show DO against temperature and Salinity on a log scale. Here the patterns are close to linear and the suitability of this model can be assessed by displaying a reference band around the linear model, based on two standard errors of the difference between a linear and a nonparametric model. Linearity looks reasonable for temperature but less so for Salinity.



The plots above were created by specifying the level of smoothing through the approximate number of degrees of freedom (6). The level of smoothing can also be chosen in a data-adaptive manner, through principles such as cross-validation or AIC.

These methods of nonparametric smoothing can be adapted to a wide variety of situations, such as more than one covariate or other types of response data.

Additive Models

Smoothing techniques can be built into wider models, particularly where several covariates are involved. An attractive framework is provided by additive models, described by Hastie and Tibshirani (1990) with an updated treatment by Wood (2006). Here, the regression model is defined as

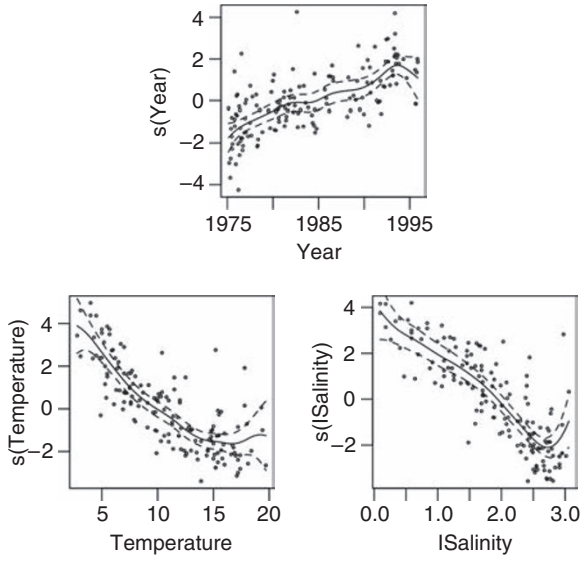
$$y_i = \alpha + m_1(x_{1i}) + \ldots + m_p(x_{pi}) + \varepsilon_i$$

for covariates $x_1, \ldots, x_p$. Each covariate x_j is allowed to influence the response variable through its own regression function m_j, which may be nonparametric but could in fact be linear or some other standard form. The backfitting

algorithm provides a means of fitting this type of model through the iterations defined by

$$\hat{m}_j^{(r+1)} = S_j\left(y - \hat{\alpha}\mathbf{1} - \sum_{k<j}\hat{m}_k^{(r+1)} - \sum_{k>j}\hat{m}_k^{(r)}\right).$$

At each stage, the regression function m_j is estimated by smoothing the partial residuals by S_j, the smoothing matrix associated with covariate j. For identifiability, the constraint that each component sums to 1 over the observed covariate values should also be added.



The panels above illustrate an additive model for the Clyde data. Instead of examining the effects of the covariates separately, they are combined into a single model which estimates the effects of one covariate while adjusting for the effects of the others. This much more powerful description now shows a much clearer time trend. The effects of temperature and salinity remain broadly linear but some unusual behavior is evident at high temperature and high salinity.

Bowman (2008) gives a more extended discussion of this example, using a different sampling station on the Clyde while McMullan et al. (2007) develop a more complex model for the whole river.

Acknowledgments

This work received partial support of grant PBCT-ADI13 of the Chilean Science and Technology Bicentennial Foundation.

About the Author

Adrian Bowman is a Professor of Statistics at the University of Glasgow. He is an elected Fellow of the International Statistical Institute and of the Royal Society of Edinburgh. He served as Joint Editor of *Applied Statistics* (J. Roy. Stat. Soc. Series C) for four years and has at various times served as associate editor for *Biometrika*, *J. Roy. Stat. Soc. Series B* and *Biometrics*. He is currently associate editor for *Biostatistics* and the *Journal of Statistical Software*. Prof. Bowman has acted as chair of the UK Committee of Professors of Statistics. He has also held a wide variety of responsibilities within the Royal Statistical Society, where he is currently an Honorary Officer.

Cross References

►Exponential and Holt-Winters Smoothing
►Median Filters and Extensions
►Moving Averages
►Nonparametric Density Estimation
►Nonparametric Estimation
►Nonparametric Models for ANOVA and ANCOVA Designs
►Nonparametric Regression Using Kernel and Spline Methods
►Smoothing Splines

References and Further Reading

Bowman A, Azzalini A (1997) Applied smoothing techniques for data analysis. Oxford University Press, Oxford
Bowman AW (2008) Smoothing techniques for visualisation. In: Chen C-H, Härdle W, Unwin A (eds) Handbook of data visualization. Springer, Heidelberg
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman & Hall, London
Green P, Silverman B (1994) Nonparametric regression and generalized linear models. Chapman & Hall, London
Hastie T, Tibshirani R (1990) Generalized additive models. Chapman & Hall, London
McMullan A, Bowman AW, Scott EM (2007) Water quality in the river clyde: a case study of additive and interaction models. Environmetrics 18:527–539
Ruppert D, Wand MP, Carroll R (2003) Semi- parametric regression. Cambridge University Press, London
Scott D (1992) Multivariate density estimation: theory, practice, and visualization. Wiley, New York
Silverman B (1986) Density estimation for statistics and data analysis. Chapman & Hall, London
Simonoff JS (1996) Smoothing methods in statistics. Springer, New York
Wand MP, Jones MC (1995) Kernel smoothing. Chapman & Hall, London
Wood S (2006) Generalized additive models: an introduction with R. Chapman & Hall/CRC Press, London

Social Network Analysis

Tom A. B. Snijders
Professor of Statistics
University of Oxford, UK
Professor of Methodology and Statistics, Faculty of Behavioral and Social Sciences
University of Groningen, Groningen, Netherlands

Social Networks

Social Network Analysis is concerned with the study of relations between social actors. Examples are friendship between persons, collaboration between employees in a firm, or trade between countries. The relation is regarded as a collection of dyadic ties, i.e., ties between pairs of actors. In most cases, data collection is either *sociocentric*, where a given group of actors is specified (in the examples this could be, e.g., a school class, a department of the firm, or all countries in the world), and all ties of the specific kind between actors in this group are considered; or *egocentric*, where a sample of actors is taken, and all ties of the sampled actors are considered. Other types of data collection exist, of which snowball sampling is the main example. The most interesting contributions of network analysis are made by considering indirect ties – in the sense that the way in which actors i and j are tied is better understood by considering the other ties of these two actors. Information about these is obtained much better from sociocentric than from egocentric approaches. Therefore, this article considers only statistical models for sociocentric network data.

The first step for the collection of sociocentric network data is to define the relation and the group of actors. This group will usually be treated as an isolated group, and any ties outside this group are disregarded. This is called the *network boundary problem*. An overview of methods for collecting network data is given by Marsden (2005).

Notation

The group of actors is denoted by $\mathcal{N} = \{1, \ldots, n\}$. Relations under study often are directed, which means that the tie $i \to j$ is distinct from the tie $j \to i$. The relation can then be represented by a nonreflexive directed graph (digraph) on $\mathcal{N}$ or, alternatively, by an $n \times n$ adjacency matrix with a structurally zero diagonal. The actors $i \in \mathcal{N}$ are the nodes of the graph. The adjacency matrix $\mathbf{y} = (y_{ij})$ indicates by $y_{ij} = 1$ or $y_{ij} = 0$, respectively, that there is a tie, or there is no tie, from actor i to actor j. The nonreflexivity means that self-ties are not considered, so that $y_{ii} = 0$ for all i. The variables y_{ij} are referred to as *tie variables*. If the network is nondirected, the representation is by a simple graph, or a symmetric adjacency matrix. Models for social networks in this article will be random graphs or digraphs and denoted by $\mathbf{Y}$.

Exponential Random Graph Models

Exponential families of probability distributions for graphs or digraphs are usually called *Exponential Random Graph Models* or ERGMs. The first model of this kind was the so-called p_1 model proposed by Holland and Leinhardt (1981). In this model the symmetrically positioned pairs (Y_{ij}, Y_{ji}) are assumed to be independent. This very restrictive assumption was lifted in the definition by Frank and Strauss (1986) of *Markov graphs*. This model can represent tendencies toward transitivity. It postulates that edge indicators Y_{ij} and Y_{hk}, when i, j, k, h are four distinct nodes, are independent conditional on the rest of the graph, i.e., conditional on the collection of tie indicators Y_{rs} for $(r,s) \neq (i,j)$, $(r,s) \neq (h,k)$. For non-directed networks with distributions not depending on the node labels, they proved that this property holds if and only if the probability distribution for $\mathbf{Y}$ can be expressed as

$$P_\theta\{\mathbf{Y} = \mathbf{y}\} = \exp\left(\sum_h \theta_h z_h(\mathbf{y}) - \psi(\theta)\right), \qquad (1)$$

where the $z_h(\mathbf{y})$ are functions of $\mathbf{y}$ each of which can be either the number of k-stars embedded in the graph $\mathbf{y}$ (for some $k, 1 \le k \le n-1$) or the number of triangles embedded in $\mathbf{y}$. These are the statistics S_k and T defined by

$$S_1(\mathbf{y}) = \sum_{1 \le i < j \le n} y_{ij} \quad \text{number of edges}$$

$$S_k(\mathbf{y}) = \sum_{1 \le i \le n} \binom{y_{i+}}{k} \quad \text{number of } k\text{-stars } (k \ge 2) \qquad (2)$$

$$T(\mathbf{y}) = \sum_{1 \le i < j < h \le n} y_{ij}\, y_{ih}\, y_{jh} \quad \text{number of triangles.}$$

The Markov model was generalized by Frank (1991) and Wasserman and Pattison (1996) to the Exponential Random Graph Model, in which the statistics $z_h(\mathbf{y})$ in (1) can be any functions of $\mathbf{y}$ and of covariates. Markov chain Monte Carlo (MCMC) methods (see ▶Markov Chain Monte Carlo) for parameter estimation for this model were proposed by Snijders (2002). Some interesting properties of this model are discussed by Robins et al. (2005). It appeared in applications, however, that in most cases the Markov model is not plausible as a model for transitivity. An model specification with more appropriate choices of the functions $z_h(\mathbf{y})$ was proposed in Snijders et al. (2006), and this has turned out to be a very useful model for representing empirically observed networks.

This model can represent dependencies between tie variables Y_{ij} in a reasonable manner. It can be used when the representation of these dependencies (transitivity, hierarchy, brokerage etc.) is an aim in itself; but also when the dependencies are a nuisance and the aim of the statistical analysis is the dependence of tie variables on covariates.

Latent Structure Models

Another way to represent dependencies between tie variables is to postulate a latent space of which the nodes are elements, and which probabilistically determines the ties. This is an application of the ideas of Latent Structure Analysis (Lazarsfeld and Henry 1986), and closely related to Latent Class Analysis. The tie variables Y_{ij} – or sometimes the dyads (Y_{ij}, Y_{ji}) – then are assumed to be conditionally independent given the latent structure.

Various latent space models have been proposed.

- A discrete (categorical) space, where the nodes have 'colors' and the distribution of the dyad (Y_{ij}, Y_{ji}) depends on the colors of i and j: see Nowicki and Snijders (2001).
- A general or Euclidean metric space, where the probability of a tie $Y_{ij} = 1$ depends on the distance between nodes i and j: see Hoff et al. (2002).
- An ultrametric space, where the probability of a tie $Y_{ij} = 1$ depends on the ultrametric distance between nodes i and j: see Schweinberger and Snijders (2003).
- A partially ordered space, where the probability of a tie $Y_{ij} = 1$ depends on how i and j are ordered: see Mogapi (2009).

Compared to Exponential Random Graph Models, these models have less flexibility to represent dependence structures between tie variables, so that they will usually achieve a less satisfactory goodness of fit. However, the representation of the nodes in the latent space can often provide an illuminating representation in itself and may be regarded as a helpful type of data reduction.

Longitudinal Models

Models for longitudinally observed networks were proposed by Snijders (2001). The most usual observational design is a panel design, where the observations of the network are $\mathbf{Y}(t_1), \ldots, \mathbf{Y}(t_M)$ for observation moments $t_1, \ldots, t_M$ ($M \geq 2$). A flexible class of models for panel data on networks can be obtained by assuming that the data are momentary observations of a continuous-time Markov process (see ►Markov Processes), in which each tie variable $X_{ij}(t)$ develops in stochastic dependence on the entire network $X(t)$. An actor-based model is often plausible, where tie changes are based on hypothetical choices of the actors. Such a model can be defined by the following steps, formulated in such a way that they can easily be represented by a computer simulation model. To obtain a parsimonious model, it is assumed that only one tie variable can change at any given moment. The model is characterized by so-called *rate functions* $\lambda_i(\mathbf{y})$ and *objective functions* $f_i(\mathbf{y})$, defined on the set of all digraphs.

1. The current state of the network is denoted $\mathbf{y}$.
2. The time until the next change is an exponentially distributed waiting time, with an expected duration of $1/\lambda_+(\mathbf{y})$ where $\lambda_+(\mathbf{y}) = \sum_i \lambda_i(\mathbf{y})$.
3. When this change occurs, the probability that an outgoing tie variable Y_{ij} of actor i can be changed, is $\lambda_i(\mathbf{y})/\lambda(\mathbf{y})$.
4. If actor i can change on outgoing tie variable, the set of new possible states of the network is

$$\mathcal{C}(\mathbf{y}) = \{\mathbf{y}' \mid y'_{hk} \neq y_{hk} \text{ only for } h = i, \text{ and for at most one } k\} .$$

The probability that the new state is $\mathbf{y}'$ is

$$\frac{\exp\left(f_i(\mathbf{y}')\right)}{\sum_{\mathbf{y}'' \in \mathcal{C}(\mathbf{y})} \exp\left(f_i(\mathbf{y}'')\right)} .$$

The model specification is done in the first place by the appropriate definition of the objective function. This is usually specified as a linear combination,

$$f_i(\beta, \mathbf{y}) = \sum_k \beta_k \, s_{ki}(\mathbf{y}) . \tag{3}$$

The functions $s_{ki}(\mathbf{y})$ represent ways in which the creation and maintenance of ties depend on currently existing ties, e.g.,

$$s_{ik}(, \mathbf{y}) = \sum_j y_{ij} \quad \text{(outdegree)}$$
$$\sum_j y_{ij} y_{ji} \quad \text{(reciprocated ties)}$$
$$\sum_{j,k} y_{ij} y_{jk} y_{ik} \quad \text{(transitive triplets),}$$

and they can also depend on combinations of network structure and covariates.

For this model, estimation procedures and algorithms according to a method of moments were proposed by Snijders (2001), Bayesian procedures by Koskinen and Snijders (2007), and an algorithm for maximum likelihood estimation by Snijders et al. (2010).

This model was generalized to a model for the simultaneous dynamics of networks and actor characteristics

("networks and behavior") by Snijders et al. (2007). Statistical procedures for this model are available in the R package RSiena.

About the Author

For biography *see* the entry ►Multilevel Analysis.

Cross References

►Graphical Markov Models
►Markov Chain Monte Carlo
►Markov Processes
►Methods of Moments Estimation
►Network Models in Probability and Statistics
►Network Sampling
►Panel Data
►Probabilistic Network Models

References and Further Reading

Basic information about social network analysis is in Wasserman and Faust (1994) and in Carrington et al (2005) A review of a variety of other statistical procedures and models for network analysis is given by Airoldi et al (2007)

Airoldi E, Blei DM, Fienberg SE, Goldenberg A, Xing EP, Zheng AX (2007) Statistical network analysis: models, issues and new directions (ICML 2006). Lecture notes in computer science, vol 4503. Springer, Berlin

Carrington PJ, Scott J, Wasserman S (eds) (2005) Models and methods in social network analysis. Cambridge University Press, Cambridge

Frank O (1991) Statistical analysis of change in networks. Stat Neerl 45:283–293

Frank O, Strauss D (1986) Markov graphs. J Am Stat Assoc 81: 832–842

Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. J Am Stat Assoc 97:1090–1098

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs (with discussion). J Am Stat Assoc 76:33–65

Koskinen JH, Snijders TAB (2007) Bayesian inference for dynamic network data. J Stat Plan Infer 13:3930–3938

Lazarsfeld PF, Henry NW (1968) Latent structure analysis. Houghton Mifflin, Boston

Marsden PV (2005) Recent developments in network measurement. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York, pp 8–30

Mogapi O (2009) A latent partial order model for social networks. D.Phil. thesis, Department of Statistics, University of Oxford

Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. J Am Stat Assoc 96:1077–1087

Robins GL, Woolcock J, Pattison P (2005) Small and other worlds: Global network structures from local processes. Am J Sociol 110:894–936

Schweinberger M, Snijders TAB (2003) Settings in social networks: a measurement model. In: Stolzenberg RM (ed) Sociological methodology, vol 23. Blackwell, Boston, pp 307–341

Snijders TAB (2001) The statistical evaluation of social network dynamics. In: Sobel ME, Becker MP (eds) Sociological methodology. Basil Blackwell, Boston and London, pp 361–395

Snijders TAB (2002) Markov chain Monte Carlo estimation of exponential random graph models. J Soc Struct 3:2

Snijders TAB, Pattison PE, Robins GL, Handcock MS (2006) New specifications for exponential random graph models. Sociol Methodol 36:99–153

Snijders TAB, Steglich CEG, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: van Montfort K, Oud H, Satorra A (eds) Longitudinal models in the behavioral and related sciences. Lawrence Erlbaum, Mahwah, pp 41–71

Snijders TAB, Koskinen JH, Schweinberger M (2010) Maximum likelihood estimation for social network dynamics. Ann Appl Stat, to be published

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Wasserman S, Pattison PE (1996) Logit models and logistic regression for social networks: I an introduction to Markov graphs and p^*. sychometrika 61:401–425

Social Statistics

Vassiliy Simchera
Director of Rosstat's Statistical Research Institute, Moscow, Russia

Social statistics is one of the largest domains of modern statistical science and practice, the subject of which is the exposure and study of regularity for formation and alteration of social phenomena with statistical techniques.

It has grown and developed at the borders of other sciences (►demography, economics, political science, philosophy, ethics, and psychology) as the discipline that integrates statistical resources and bases of humanitarian information studying human beings and society. It gained intensive development in the 19thand 20th centuries as a science studying social dynamics, which was initiated in the United States by Russian-American sociologist *Pitirim Sorokin*, although the first record of it one can find in ancient origins at the beginning of AD.

Social statistics operates with the branched system of indicators characterizing standards of life and human activities and further groups of people, public societies, nations, and civilizations, their evolution and structure, ways and standards of life, households, culture, education, moral, and human values, freedoms, rights, etc.

In contrast to many other statistical disciplines, its main emphasis is on the study of quantitatively immeasurable indicators as most common in social science.

It also scrutinizes and forecasts unobservable and non-registering "shadow," illegal, and informal social phenomena, by means of analysis techniques of social projects and doctrines, votes, and elections in particular.

In its work, along with the methods of sample surveys and ►public opinion polls, social statistics extensively applies *special methods*, among which are various methods of multivariate factor analysis, cluster analysis (see ►Cluster Analysis: An Introduction), and latent analysis. The particular classes are the methods of social modeling and managerial social analysis, on the basis of which a new section of modern statistics, called sociometrics evolved.

At present time, social statistics is positioned as an instrument of the application of its methods and information about social sciences, the main aim and product of which is qualitative measurement of social and widely spiritual aspects of material production and their integration as superior values and achievements of modern society into the socio-economic context.

There are an extensive collection of models, not only for common but also for applied social changes, in particular, the dynamics of climate change, epidemics, catastrophes, health care and diseases, crime, cloning, psychological and psychotropic conspiracies and wars, application of up-to-date and specialized computer and mathematical methods in demographics, medicine and sanitary statistics, as well as in biology, anthropology and other related sciences.

Social statistics also develops as *social groups statistics*, in particular poverty statistics, behavioral statistics, i.e., behavior of people in the exotic environment, statistics of crime, statistics of fair competition, and statistics on globalization and mass protests.

Another area is a statistics of interethnic conflicts and wars, terrorism, crisis and anthropogenic catastrophes, which threaten the existence of world civilizations.

Social statistics is formed on the basis of sampling surveys and public opinion polls; it actually relies upon opinions about facts rather than on the facts themselves, it characterizes mainly feedback, original responses to events in the surrounding world, rather than the events themselves. Without reliable criteria of estimation for data quality. Social statistics and its indicators, where applicable, require preliminary verification of their results and publications as they are least of all true and acceptable.

Main Social Statistics Centers:

- *Harvard Institute for Quantitative Social Science*
- *Inter-University Consortium for Political and Social Research*
- *Social Statistics Division, School of Social Sciences, University of Southampton, UK*
- *Social Statistics Research Group, University of Auckland, New Zealand*
- *UN Statistics Division - Demographic and Social Statistics*
- *Organization for Economic Co-operation and Development* (*OECD*)

Cross References

►Economic Statistics
►Public Opinion Polls
►Small Area Estimation
►Sociology, Statistics in

References and Further Reading

EuroStat (2006) European social statistics-social protection, Luxembourg
Irvine J, Miles I, Evans J (eds) (1979) Demystifying social statistics (1979). Pluto Press, London

Sociology, Statistics in

Gudmund R. Iversen
Professor Emeritus
Swarthmore College, Swarthmore,
PA, USA

Introduction

Statistics and sociology have a strong relationship that goes back several centuries. As new social theories and methods have been developed, statistics has responded by developing appropriate statistical methods. Also, sociologists have been quick adopting new statistical methods not necessarily developed with them in mind. The same is also the case with other social sciences such as political science, economics and psychology.

A few social sciences have relied more on statistics than others. Perhaps, the heaviest user of statistics has been economics, and the uses of statistics there have led to their own branch of statistics known as *econometrics*. With the abundance of economic data, econometrics has led to new uses of regression analysis. In turn, econometrics has been adopted by other social sciences, such as sociology and psychology.

Psychology is another social science where statistics has led to its own branch of statistics known as *psychometrics*. Psychology has an abundance of scores on

tests administered to college students and people seeking employment as well as psychiatry trying do diagnose people with suspected mental disorders. The most well known statistical methods in psychometrics is known as *factor analysis* of various kinds.

The abundance of survey analysis with the uses of questionnaires (see ►Questionnaire) in sociology would not have been possible without modern statistical *sampling* methods. Needs of sociology have led statisticians to develop sampling methods such as stratified sampling, ►cluster sampling and other sampling procedures. In turn, this has spilled over into the uses of sampling when the goal is obtain a complete ►census of some population. One of the leading organizations in the development of modern sampling methods for the collection of social science data has been the United States Bureau of the Census.

Sampling Theory

Sociologists, as well as others, have long collected data on individuals to study how people feel about issues of the day. In addition, political scientists have used sample surveys to try to predict outcomes of elections to be held sometime in the future. One of the most famous examples of such a prediction being wrong took place during the presidential election in the United States in 1948. On the night of the elections many surveys showed that Thomas E. Dewey had won and the incumbent Harry S. Truman had lost the election. Instead, Truman woke up the next day and found he had been elected president for the next four years. Another famous example took place during the US presidential election of 1936 when a well-known publication predicted on the basis of their poll that Governor Alf Landon would win the election. Instead, Franklin D. Roosevelt won almost two thirds of the popular vote that year and went on to win the next two elections as well.

What went wrong in both of these two cases was that statisticians had not stressed hard enough is that in order to generalize from a sample to a larger population, the sample must have been selected according to proper random statistical methods. In 1936 the sample was drawn from lists of people who owned cars. But this was in the middle of the economic depression years, and only reasonably wealthy people owned cars while most people without cars voted for Roosevelt. In 1948 George Gallup and others made use of the so-called quota sampling method. Each interviewer was told to go out and select respondents in such a way that the sample would reflect the population on characteristics such as gender and age. But that way interviewers would miss people who worked during off hours like a night shift at a factory and slept during the daytime when interviewers were seeking people with the right characteristic to satisfy the quotas they were given. An occasional survey still uses quota sampling for the selection of respondents, in spite of the well-known shortcomings of quota sampling. These days it is much more common to chose respondents by making a random selection of telephone numbers and dial those numbers.

Demography

For centuries, states have wanted to count the number of inhabitants for tax and military purposes. For this purpose, the German word *Statistik* was introduced more than two hundred and fifty years ago to denote matters of state, and the word probably comes from the Latin word *Statisticum*. In principle, a census does not require the use of statistical methods, but it is very difficult to take an accurate census without the use of sampling to count people who otherwise would be hard to include in the final count.

Simultaneous Structural Equations

The analysis of complex sociological models has led to generalizations of simple regressions models to models involving several regression equations where the parameters in all the equations are estimated at the same time. This formulation of a model has led both statisticians and sociologists to fruitful collaborations on how to estimate the parameters and how to interpret the estimates. The estimation procedure has moved from ordinary least squares estimation to what is known as two-stage and even three-stage estimation, depending upon the model. This is a case where theoretical work by economists have made major contributions to statistical theory and major uses in sociology.

Such models also go under the name of causal analysis or path analysis. Path analysis seems to have originated in biology around 1920, and it caught on in sociology in the 1960ies. A leading person in this field was the sociologist Hubert Blalock, perhaps best known for his famous textbook *Social Statistics* in addition to his writings on causal models. Causal modeling using path analysis has lost some of its attraction after people realized that establishing causality using statistical models did not necessarily lead to truly causal connections between variables.

Contingency Table Analysis

Much of the data in sociology consist of nominal (qualitative) variables such as gender (female, male), religious affiliation (protestant, catholic, Muslim, Jewish, etc.) and others. Because there are no meaningful numerical values attached to these categories, such data cannot be analyzed

by using means, standard deviations, single or multiple regression, etc. Instead, perhaps the best-known and oldest statistical method for the analysis of the relationship between two such variables is the chi-square analysis. It is based on the difference between the observed frequencies and expected frequencies computed as what the frequencies would have been if there were no relationship between the two variables.

A more recent development is the multivariate chi-square analysis for more than two categorical variables. This permits the study of interaction effects of the independent variables onto the dependent variable. Also, ►logistic regression has become popular for the case where the dependent variable has only two values. Finally, the use of ►dummy variables for quantitative variables have become possible using software so designed. Any quantitative variable with k different categories can be represented by $k - 1$ dummy variable, each having values of 0 and 1. With the data in this form it is possible to use ordinary linear regression for the study of the relationship between the dependent and the independent variables.

Conclusion

The empirical part of sociology could not exist without the use of statistics. Statistics has become an integral part of empirical sociological research. Any randomly chosen issue of a major sociological journal will have several articles making using of data analysis and statistics.

At one time it looked as if mathematics could play a similar role for sociology, but that effort has not paid off the way it was hoped. This takes us back to the importance of statistics for sociology. However, a major obstacle is that most sociologists lack the necessary background in statistics, partly due to the fact that they do not know enough mathematics to fully understand the statistical methods they are using. Similarly, most statisticians lack the knowledge of sociology needed to understand what statistical methods sociologists need. A few people have been able to bridge this gap, but most sociology students, even sociology graduate students, see the study of statistics as a hard task, perhaps mostly because statistics for sociologist has not been taught very well.

About the Author

For biography *see* the entry ►Analysis of Variance.

Cross References

►Chi-Square Test: Analysis of Contingency Tables
►Confounding and Confounder Control
►Demography
►Event History Analysis
►Factor Analysis and Latent Variable Modelling
►Non-probability Sampling Survey Methods
►Psychology, Statistics in
►Role of Statistics
►Social Statistics
►Structural Equation Models

References and Further Reading

Blalock H (1971) Causal models in the social sciences. Aldine, New York
Hald A (1998) A history of mathematical statistics from 1750 to 1930. Wiley, New York
Stiegler SM (1986) The history of statistics: the measurement of uncertainty before 1900. Belkap Press of Harvard University Press, Cambridge, MA and London

Spatial Point Pattern

Peter J. Diggle
Distinguished Univerisity Professor
Lancaster University, Lancaster, UK
Adjunct Professor
Johns Hopkins University School of Public Health, Baltimore, MD, USA
Adjunct Senior Researcher
Columbia University, New York, NY, USA

Introduction

A spatial point pattern is a set of data consisting of the locations, $x_i : i = 1, \ldots, n$, of all events of a particular kind within a designated spatial region A. Typically, the pattern is assumed to be the outcome of a stochastic point process (see ►Point Processes) whose properties are of scientific interest.

An example would be the locations x_i of all trees in a designated region within a naturally regenerated forest. The observed pattern could be the result of a complex mix of natural processes. For example: regeneration from seedlings around the base of a mature tree could produce clusters of young trees; variation in soil fertility could produce patches of relatively low and high intensity of regeneration; competition for limited nutrient or light could lead to a spatially regular pattern is which only the dominant member of a cluster of seedlings survives.

Complete Spatial Randomness

The simplest statistical model for a spatial point process is the homogeneous Poisson process (see ►Poisson Processes). One of several possible definitions of this process is that:

1. The number of points in any planar region A follows a Poisson distribution with mean $\lambda|A|$, where $|\cdot|$ denotes area and the parameter $\lambda > 0$ is the *intensity*, or mean number of points per unit area.
2. The numbers of events in any two disjoint areas are independent.

Properties (1) and (2) imply that, conditionally on the number of points in A, their locations form an independent random sample from the uniform distribution on A.

Models

The Poisson process provides a standard of complete spatial randomness, but is inadequate as a model for most naturally occurring phenomena. As would be the case in our hypothetical forestry example, we need models to describe a response to an inhomoegeneous environment, or a tendency for points either to cluster together or to inhibit the occurrence of mutually close sets of points.

To model a response to an inhomogeous environment, a first possibility is to replace the constant intensity λ by a function $\lambda(x)$. In practice, this is only useful if we can model $\lambda(x)$ as a function of spatially referenced explanatory variables, for example height above sea-level. In the absence of such information, we can treat $\lambda(x)$ as a realisation of an unobserved stochastic process, so defining the class of Cox processes (Cox 1955).

The first, and still widely used, model for clustering of points is the Neyman–Scott process (Neyman and Scott 1958), in which *parents* form a homogeneous Poisson process and each parent generates a family of *offspring* that are spatially dispersed around their parent. Bartlett (1964) showed that in some cases the resulting process is indistinguishable from a Cox process; specifically, a process in which family sizes are independent Poisson variates and the positions of offspring relative to their parents are an independent random sample from a bivariate distribution with density $f(\cdot)$ is also a Cox proicess with stochastic intensity proportional to $\sum_{i=1}^{\infty} f(x - X_i)$, where the X_i are the points of a homogeneous Poisson process.

The most widely used model for an inhibitory process is a Markov point process (Ripley and Kelly 1977). A Markov point process can be defined by its likelihood ratio with respect to a Poisson process with intensity $\lambda = 1$. A useful sub-class of such processes is the *pairwise interaction process*, in which the likelihood ratio for a realization $\mathcal{X} = \{x_i : i = 1, \ldots, n\}$ is

$$\ell(\mathcal{X}) = \beta^n \prod_{j \neq i} h(||x_i - x_j||),$$

where $||\cdot||$ denotes distance, $h(\cdot)$ is an *interaction function* and $\beta > 0$ determines the intensity of the process. A sufficient condition for validity of the model is that $h(\cdot)$ is inhibitory, meaning that $0 \leq h(u) \leq 1$ for all u. The case $h(u) = 1$ yields a homogeneous Poisson process.

Inference

Until relatively recently, likelihood-based inference was considered intractable for most spatial point process models. Instead, sensible ad hoc methods based on functional summary statistics were used. These included so-called nearest neighbor methods and moment-based methods (Ripley 1977). Recent developments in Monte Carlo methods of inference have made likelihood-based inference a feasible, albeit computationally intensive, alternative (Møller and Waagepetersen 2004).

General accounts of statistical models and methods for spatial point pattern data include Diggle (2003) and Ilian et al. (2008).

About the Author

Peter Diggle is Distinguished University Professor of Statistics and Associate Dean for Research in the School of Health and Medicine, Lancaster University, Adjunct Professor in the Department of Biostatistics, Johns Hopkins University School of Public Health and Adjunct Senior Researcher in the International Research Insitute for Climate and Society, Columbia University. Between 1974 and 1983 he was a Lecturer, then Reader, in Statistics at the University of Newcastle upon Tyne. Between 1984 and 1988 he was Senior, then Principal, then Chief Research Scientist and Chief of the Division of Mathematics and Statistics at CSIRO, Australia. Peter's research interests are in the development of statistical methods for spatial and longitudinal data analysis, motivated by applications in the biomedical, health and environmental sciences. He has published 8 books and around 180 articles on these topics in the open literature. He was awarded the Royal Statistical Society's Guy Medal in Silver in 1997, is a former editor of the Society's Journal, Series B and is a Fellow of the American Statistical Association. Peter was founding co-editor, with his close friend and Johns Hopkins colleague Scott Zeger, of the journal *Biostatistics* between 1999 and 2009. He is a Trustee for Biometrika, and a member of the UK Medical Research Council's Population and Systems Medicine Research Board. Away from work, Peter

plays mixed-doubles badminton with his family (partner Amanda, children Jono and Hannah). He also enjoys music, playing guitar and recorder, and listening to jazz.

Cross References

►Analysis of Areal and Spatial Interaction Data
►Point Processes
►Poisson Distribution and Its Application in Statistics
►Poisson Processes
►Spatial Statistics

References and Further Reading

Bartlett MS (1964) The spectral analysis of two-dimensional point processes. Biometrika 51:299–311
Cox DR (1955) Some statistical methods related with series of events (with discussion). J R Stat Soc B 17:129–57
Diggle PJ (2003) Statistical analysis of spatial point patterns, 2nd edn. Arnold, London
Ilian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistial analysis and modelling of spatial point patterns. Wiley, Chichester
Møller J, Waagepetersen RP (2004) Statistical inference and simulation for spatial point processes. Chapman & Hall, London
Neyman J, Scott EL (1958) Statistical approach to problems of cosmology. J R Stat Soc Ser B 20:1–43
Ripley BD (1977) Modelling spatial patterns (with discussion). J R Stat Soc B 39:172–212
Ripley BD, Kelly FP (1977) Markov point processes. J Lond Math Soc 15:188–92

Spatial Statistics

Jürgen Pilz
Professor, Head of the Institute of Statistics
University of Klagenfurt, Klagenfurt, Austria

Introduction

Spatial statistics is concerned with modeling and analysis of spatial data. By spatial data we mean data where, in addition to the (primary) phenomenon of interest the relative spatial locations of observations are recorded, too, because these may be important for the interpretation of data. This is of primary importance in earth-related sciences such as geography, geology, hydrology, ecology and environmental sciences, but also in other scientific disciplines concerned with spatial variations and patterns such as astrophysics, economics, agriculture, forestry, epidemiology and, at a microscopic scale, medical and health research.

In contrast to non-spatial data analysis, which is concerned with statistical modelling and analysis of data which just happen to phenomena in space and time, spatial statistics focuses on methods and techniques which consider explicitly the importance of the locations, or the spatial arrangement of the objects being analysed. The basic difference from classical statistics is that in spatial statistics we are concerned with non-independence of observations.

In spatial problems, observations come from a spatial random process $\mathcal{Z} = \{Z(s) : s \in S\}$, indexed by a spatial/spatiotemporal set $S \subset \mathbb{R}^d$, with $Z(s)$ taking values in some state space. The positions of observation sites $s \in S$ are either fixed in advance or random. Typically, $S \subset \mathbb{R}^2$, the study of spatial dynamics adds a temporal dimension, i.e., $S \subset \mathbb{R}^2 \times (0, \infty)$. However, S could also be one-dimensional (e.g., field trials along transect lines) or a subset of $\mathbb{R}^3$ (oil and mineral prospection, 3D imaging). In some fields such as Bayesian data analysis and simulation one even requires spaces S of dimension $d \geq 3$, this pertains, in particular, to the design and analysis of computer experiments with a moderate to large number of input variables. Comprehensive treatments of the whole field of spatial statistics are given in Ripley (1988), Cressie (1993) and Gaetan and Guyon (2010). Statistical Methods for spatio-temporal systems are given in Finkenstädt et al. (2007).

Basically, there are four classes of problems which spatial statistics is concerned with: point pattern analysis, geostatistical data analysis, areal/lattice data analysis and spatial interaction analysis. These subproblems are treated separately in a number of papers in this volume: Mase (2010), Kazianka and Pilz (2010), Vere-Jones (2010), Diggle (2010) and Spöck and Pilz (2010). Therefore, in this paper we limit ourselves to a brief overview over the areas comprising spatial statistics.

For a good overview on software for different problem areas of spatial data analysis we recommend the book by Bivand et al. (2008), for the important issue of simulation of spatial models we refer to Lantuéjoul (2002) and Gaetan and Guyon (2010).

Geostatistics

Here, S is a *continuous* subspace of $\mathbb{R}^d$ and the random field is observed at n fixed sites $\{s_1, \ldots, s_n\} \subset S$. Typical examples include rainfall data, data on soil, characteristics (porosity, humidity etc.), oil and mineral exploration data, airquality and groundwater data a.s.o. For $d \geq 2$ the random process $\mathcal{Z} = \{Z(s) : s \in S\}$ is usually termed a *random field*. The mathematical structure and the most important properties of random fields are described in Moklyachuk (2010).

The concept of stationarity is key in the analysis of spatial and/or temporal variation: roughly spoken, stationarity means that the statistical properties. (e.g., mean and

variance) of the variable of interest do not change over the considered area. However, testing for stationarity is not possible. For spatial prediction the performance of a stationary and a nonstationary model could be compared through assessment of the accuracy of predictions.

The random field is characterised by its finite dimensional distributions $P(Z(s_1) \le z_1, \ldots, Z(s_n) \le z_n)$ for all $n \in \mathbb{N}$ and $s_j \in S; j = 1, \ldots, n$. If all these distributions are Gaussian then $\mathcal{Z}$ is called a *Gaussian random field* (GRF). A GRF is completely determined by its expectation (trend function) $m(s) = E(Z(s))$ and covariance function $C(s_1, s_2) = Cov(Z(s_1), Z(s_2))$. Contrary to traditional statistics, in a geostatistical setting we usually observe only one realization of Z at a finite number of locations $s_1, \ldots, s_n$. Therefore, the distribution underlying the random field cannot be inferred without imposing further assumptions. The most simple assumption is that of (strict) stationarity, which means that the finite dimensional distributions do not change when all positions are translated by the same (lag) vector h, i.e., $(Z(s_1), \ldots, Z(s_n))$ and $(Z(s_1 + h), \ldots, Z(s_n + h))$ are identically distributed for all $n \in \mathbb{N}$ and locations $s_j \in S; j = 1, \ldots, n$. For a GRF this implies that $m(s) = \text{const}$ for all $s \in S$, and $C(s_1, s_2) = C(s_1 - s_2)$ for all $s_1, s_2 \in S$. For arbitrary RF's, the invariance of the first two moments is denoted as the property of *weak stationarity*. In geostatistics it is common to use the so-called semi-variogram $\gamma(s_1, s_2) = 0.5 * Var(Z(s + h) - Z(s))$ instead of the covariance function and to assume *intrinsic* stationarity: $m(s) = \text{const}$ and $\gamma(s, s + h) = \gamma(h)$ for all $s, h \in S$. If $Z(\cdot)$ is weakly stationary then $\gamma(h) = C(0) - C(h)$. Weak stationarity implies intrinsic stationarity, the converse is not true.

For $d = 1$, however, intrinsic stationarity is equivalent to weak stationarity of the first order differences of the underlying random process, a well-known fact from time series analysis. For an intrinsically stationary RF the semi-variogram has the important property of *conditional negative definiteness*, i.e.,

$$Var(a_1 Z(s_1) + \ldots + a_n Z(s_n)) = -\sum_{i=1}^{n} \sum_{j \neq i} a_i a_j \gamma(s_i - s_j) \ge 0$$

for all $n \in \mathbb{N}$ and real numbers $a_1, \ldots, a_n$ such that $\sum a_i = 0$. This is the reason why one usually employs parametric models (e.g., spherical, exponential, Gaussian or Matérn models) for fitting variogram functions to the data. Moreover, fitting is often done under the additional assumption of isotropy: $\gamma(h) = \gamma(|h|), |h| = \text{length of } h \in S$. For "classical" estimation methods for variogram parameters see Mase (2010), for Bayesian approaches we refer to Banerjee et al. (2004) and Kazianka and Pilz (2010). For non-stationary variogram modeling we refer to the review provided by Sampson et al. (2001) and Schabenberger and Gotway (2005).

Now, let us step to predicting Z at an unobserved location $s_0 \in S$, based on the observations $\mathbf{Z} := (Z(s_1), \ldots, Z(s_n))^T$, such that the mean squared error of prediction (MSEP) $E[Z(s_0) - \hat{Z}(s_0)]^2$ is minimized. For a GRF, the optimal predictor is known to be the mean of the conditional distribution of $Z(s_0)$ given the data:

$$\hat{Z}(s_0) = E(Z(s_0)|\mathbf{Z}) = E(Z(s_0)) + c_0^T K^{-1}(\mathbf{Z} - E(\mathbf{Z})) \quad (1)$$

where the vector c_0 has elements $C(s_0 - s_i); i = 1, \ldots, n$; and K is the covariance matrix of the observations. For non-Gaussian RF's, the predictor (1) is the best linear unbiased predictor (BLUP). Inserting the optimal estimators for $EZ(s_0)$ and $E(\mathbf{Z})$ into 1 we get various forms of Kriging predictors: assuming $EZ(s) = m$ to be constant we get $\widehat{EZ(s_0)} = \hat{m} = (\mathbf{1}^T K^{-1} \mathbf{Z})/(\mathbf{1} K^{-1} \mathbf{Z})$ and $E(\mathbf{Z}) = \hat{m}\mathbf{1}$, where $\mathbf{1}$ denotes the n-vector of one's, and this is known as the *ordinary Kriging* predictor. For non-constant m, assuming a linear regression setup for $m(s)$, one arrives at the *universal Kriging* predictor. Clearly, for non-Gaussian data, the best predictor w.r.t. MSEP is no longer linear in the observations. Comprehensive accounts of "classical" linear and nonlinear geostatistics are given in Chilés and Delfiner (1999) and Webster and Oliver (2007).

In a Bayesian setting, assuming a prior distribution for the covariance parameters, one has to determine the predictive density of $Z(s_0)|\mathbf{Z}$ via the posterior distribution of the covariance parameters given $\mathbf{Z}$, from which an optimal predictor and the associated uncertainty can be derived. For non-Gaussian data, the framework of generalized linear models or the copula framework can be used to arrive at optimal predictors (see Banerjee et al. (2004), Diggle and Ribeiro (2007) and Kazianka and Pilz (2010)). This extension of the classical geostatistical methodology has become known under the heading of *model-based geostatistics*. Concerning software for geostatistical analysis, we recommend the freely available R-packages "gstat," "geoR," "geoRglm" and the functions contained in the R-library "intamap." For spatio-temporal analysis and prediction of environmental processes we refer to Le and Zidek (2006) where also software is being described. For geostatistical space-time models particular care is needed for combining spatial and temporal variables (separability versus non-separability), a thorough treatment of this issue is given in Gneiting et al. (2007). A very exciting new development has been opened by Rue et al. (2009) who consider approximate Bayesian inference in latent Gaussian models, using an integrated nested Laplace approximation (INLA). This

approach offers computational advantages, the approximations are accurate and orders of magnitude faster than MCMC algorithms, and its generality also allows the computation of various predictive measures for doing model comparisons.

Point Process Analysis and Random Sets

By a (spatial) point process (PP) or point pattern we mean a random, locally finite collection $\mathcal{Z} = \{s_1, s_2, \ldots\}$ of points $s_i \in S \subset \mathbb{R}^d$ such that $s_i \neq s_j$ for $i \neq j$. Here, locally finite means that the number of points is finite in each bounded subset of S. The process is said to be *marked* if at each site s_i we additionally record a (random) value, for example the length of the material cracks, height or diameter of plants, intensity of earthquakes a.s.o. For statistical analysis, the process is observed in a window $W \subset S$ leading to a realization $z = \{s_1, \ldots, s_n\}$ with a random number $n = n(z)$ of points $s_i \in S$. Thus, contrary to geostatistical data analysis, in point pattern analysis the set of observation sites $\{s_1, \ldots, s_n\}$ is random, along with the number of sites n.

►Point processes are important in a variety of applications, in ecology and forestry (spatial, spatiotemporal distribution of plant/animal species), epidemiology (location of sick individuals, spatiotemporal spread of diseases), seismology (earthquake epicenters), materials science (locations of cracks and porosities), biology and medicine (centers of cells/tumours in histological sections), crime scene analysis (locations and intensities of burglaries) etc.

The probabilistic theory of PP's is quite technical and requires a good knowledge of measure theory, for a good introductory account we refer to the review articles by Møller and Waagepetersen (2007), Vere-Jones (2010) and Diggle (2010).

The PP $\mathcal{Z}$ is characterized through the finite-dimensional distributions $(N(B_1), \ldots, N(B_k))$ for all $k \in \mathbb{N}$ and bounded subsets $B_1, \ldots, B_k$ in $\mathbb{R}^d$, where the random variable $N(B_i)$ counts the number of points in B_i. The point pattern is called *stationary*, iff its finite-dimensional distributions are invariant under translations, and *isotropic* iff all these distributions are invariant under rotations.

One of the major problems is to find out whether a given point pattern can be considered as completely random, or if there is a tendency to clustering or to some "regularity." As the reference model for "no interaction between points" or "complete spatial randomness (CSR)" the *Poisson* Process (see ►Poisson Processes) is chosen (cf. Diggle 2010).

In general the mean structure of the count variables is modelled by a non-negative intensity function $\lambda(\cdot)$ such that $\mu(B) := \int_B \lambda(s)ds$ for all B in $\mathbb{R}^d$. Here the interpretation is that $\lambda(s)ds$ is the probability that there is precisely one point in the ball with center at s and area/volume ds. Likewise, the second order moment measure $\mu_2(A \times B) := E\{N(A)N(B)\}$ is modelled by a second order product density λ_2 such that $\mu_2(A \times B) = \int_A \int_B I_{A\times B}(u,v)\lambda_2(u,v)dudv$. For a Poisson PP one then has: $\mu_2(a \times B) = \mu(A)\mu(B), \lambda_2(u,v) = \lambda(u)\lambda(v)$.

The tendency of attraction or repulsion between points can be characterized by the so-called *pair correlation* function $g(u,v) := \lambda_2(u,v)/[\lambda(u)\lambda(v)]$. If points appear independently from each other then we have $\lambda_2(u,v) = \lambda(u)\lambda(v)$ and thus $g(u,v) = 1$. Thus, there is attraction between points of $\mathcal{Z}$ at locations u and v iff $g(u,v) > 1$ and repulsion iff $g(u,v) < 1$.

The characterization of point patterns becomes relatively easy in case of stationarity and additional isotropy. Then $\lambda(u) = \lambda = const$, $\lambda_2(u,v) = \lambda_2(|u-v|)$, $g(u,v) = g(|u-v|)$ and it suffices to work with the so-called K-function $K(r) = (1/\lambda)E$ {number of extra points within distance r of a randomly chosen point}. This takes the form

$$K(r) = (v_d/\lambda^2) \int_0^r u^{d-1}\lambda_2(u)du$$

where v_d stands for the surface area of the unit sphere in $\mathbb{R}^d$. For the Poisson PP in $\mathbb{R}^2$, for example, we have $K(r) = \pi r^2$. We remark, however, that second order moments and the related K function describe the dependence in point patterns only partly, i.e., the visual appearance of two point patterns may be different even if their first and second order moments are the same. Therefore, other features are considered as well, in particular the *empty space function* F_s and the *nearest neighbour function* G_s. The former is defined as $F_s(r) = P(N(b(s,r)) > 0)$, where $b(s,r)$ is the ball with radius $r > 0$ and centered at a fixed location $s \in \mathbb{R}^d$ (not necessarily $s \in \mathcal{Z}$). For a stationary PP the function F_s does not depend on s. The function G_s is the distribution function of the distance of a given point $s \in \mathcal{Z}$ to its nearest neighbour in $\mathcal{Z}$, i.e., $G_s(r) = P(N(b(s,r) > 1 | s \in \mathcal{Z})$. For the sake of comparison, the functions F and G are compared to those of a homogeneous Poisson (constant intensity) PP, for which $F(r) = 1 - exp(-\lambda|b(0,r)|) = G(r), r > 0$. Popular models of processes with dependence between points include the *Cox* PPs (less regular than Poisson PPs) and the *Gibbs* PPs (more regular than Poisson PPs). The Cox-process is defined by a two-stage model $Z|\zeta$ with random intensity $\mu(B) = \int \zeta(s)ds$ where ζ is a latent (non-observable) non-negative random field. For example, $\mathcal{Z}$ describes the (random) locations of the plants and ζ models the random environmental conditions at these

locations. Therefore, a Cox process is often termed a "doubly stochastic" Poisson PP (Poisson PP with random intensity). Assuming $\log \zeta(\cdot)$ to be a Gaussian RF leads to the widely used *log-Gaussian Cox process*: $\log \zeta(s) = g(s)^T \beta + \varepsilon(s)$, $g(s)$ includes the covariates, β is a parameter vector modeling (random) effects and $\varepsilon(s)$ is a centered Gaussian RF. Choosing $\zeta(s) = \lambda \sum_i k(s - s_i)$, where $\{s_1, s_2, \ldots\}$ form a stationary Poisson PP and $k(\cdot)$ is a density on S centered at $s_i \in \mathbb{R}^d$, we arrive at a so-called *Neyman–Scott process*. This way clustering tendencies can be modelled interpreting the points s_i as cluster centers (positions of parents) around which clusters with random numbers of descendants (children) are formed. Various special cases arise with particular choices of the density function $k(\cdot)$, choosing e.g., a Gaussian density results in a *Thomas* PP. The class of Cox models allows for many generalizations of Thomas and Neyman–Scott processes: different spatial configuration of the parents PP, interdependence (competition) and nonidentical distribution for children (variable fertility of parents) etc., all leading to aggregated PPs which are less regular than the Poisson PP.

One way to "regularize" a spatial point pattern is to disallow close points. This is appropriate for modeling situations such as tree distributions in forests and cell distributions in cellular tissues. These models are special cases of Gibbs models which are conditionally specified through the probabilities that there is a point at location s given the pattern on $\mathbb{R}^d \setminus \{s\}$: $\lambda(s|z)ds := P(N(b(s, ds)) = 1|\mathcal{Z} \cap (\mathbb{R}^d \setminus \{s\}) = z)$. The conditional intensity $\lambda(s|z)$ is usually modelled through some energy functional $U(s, z)$: $\lambda(s|z) = \exp(-U(s, z))$. For example, *Strauss* PP's correspond to the choice $U(s, z) = \exp\big(-a - b \sum_i I(\|s - s_i\| \le r)\big)$ including only the energy of the singletons and pair potentials. For $b > 0$ we have repulsion and, conversely, $b < 0$ implies attraction. We remark that the Strauss PPs are examples of *Markov* PPs since the conditional density $\lambda(s, z)$ depends only on neighboring points of s belonging to the pattern z.

For testing the CSR hypothesis, the parameters and functions introduced before $(\lambda, \lambda_2, K, F, G)$ have to be estimated on the basis of an observation window $W \subset \mathbb{R}^d$ (usually a (hyper-) rectangle). For testing this hypothesis, estimates of the following two summary statistics are in common use: $L(r) = \{K(r)/b_d\}^{1/d}$ and $J(r) = (1 - G(r))/(1 - F(r))$, b_d denotes the volume of the unit sphere in $\mathbb{R}^d$. For a stationary PP, $J > 1, J = 1$ and $J > 1$ indicate respectively that the PP is more, equally or less regular than a Poisson PP. For estimation of G the well-known ►Kaplan–Meier-estimator can be used, for a comprehensive discussion of estimators and its properties we refer to Illian et al. (2008). Baddeley et al. (2006) present a number of interesting case studies in spatial point process modeling, in areas as diverse as human and animal epidemiology, materials sciences, social sciences, biology and seismology. For practical estimation and testing we recommend the freely available R-package "spatstat."

Random Sets

These are generalizations of point patterns in such a way that $\mathcal{Z}$ defines an arbitrary random closed subset (RACS) of $\mathbb{R}^d$. Again, stationarity means that the distributions of $\mathcal{Z}$ are invariant w.r.t. translations. In this case, random closed sets can be characterized by some simple numbers and functions, resp., e.g., by (a) the covariance function $C(h) = P(\{s \in \mathcal{Z}\} \cap \{s + h \in \mathcal{Z}\})$ and (b) the contact distribution $H_B(r) = 1 - P(\mathcal{Z} \cap rB = \emptyset)/(1 - P(s \in \mathcal{Z}))$ for some (test) set $B \subset \mathbb{R}^d$, e.g., a ball or polygon.

The most simple models for RACS are Boolean models, $\mathcal{Z} = \bigcup_{i=1}^{\infty} \{Z_i + s_i\}$, where $\{s_1, s_2, \ldots\}$ is a Poisson PP with constant intensity and $Z_1, Z_2, \ldots$ a sequence of i.i.d. RACS which are independent of the PP. For instance, Z_i can be assumed to be spheres with random radii, or segments of random length and direction. In applications, the random sets are not of that simple type. However, more realistic models can be built on the basis of Boolean models using the opening and closure operations of mathematical morphology, see e.g., Serra (1988) and Lantuéjoul (2002); for interesting applications in the materials sciences we refer to Ohser and Mücklich (2000).

Lattice Data Analysis

In areal/lattice data analysis we observe the random field $\mathcal{Z} = \{Z(s) : s \in S\}$ at the points of a fixed, discrete and non-random set $S \subset \mathbb{R}^d$, which is then often called a *lattice*. Then it is sufficient to describe the joint probability function or density on S. Typical examples of such type of data are population characteristics and infections disease numbers at district or country level, remote sensing imagery and image texture data from materials sciences. The lattice may be regularly or irregularly spaced. In areal data analysis, the measurements are aggregated over spatial zones (administrative units, land parcel sections) and the points s_i are geographical regions (areas) represented as a network with a given adjacency graph. In image analysis, the lattice S is a regularly spaced set of pixels or voxels. Goals of the analysis for these types of data include the quantification of spatial correlations, prediction, classification and synthesis of textures and image smoothing and reconstruction.

For areal data analysis usually autoregressive models are employed, the spatial correlation structure is induced by the particular model chosen, e.g., SAR or CAR models. For a detailed account of this type of analysis we refer to

Lloyd (2007) and Anselin and Rey (2010), for an overview and further references see Spöck and Pilz (2010). A particular area of lattice data analysis is image analysis where $d = 2$ (or 3), $S = \{1,\dots,N\}^d$ and $N = 2^k$ for some integer $k > 1$. For modelling, Markov random fields are widely used. We call $\mathcal{Z} = \{Z(s) : s \in S\}$ a *Markov random field* if the conditional density of $Z(s)$ given $Z(y), y \neq s$, only depends on realizations of $Z(y)$ for which y belongs to some neighbourhood $\mathcal{N}(s)$ of s. As a simple example, consider a Gaussian Markov random field (GMRF). The neighborhood of s is usually defined via a symmetric neighborhood relation $s \sim y$ which is non-reflexive, i.e., $s \not\sim s$. Then the joint density on S can be written as $p(\boldsymbol{z}) \propto \exp(-0.5(\boldsymbol{z}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{\mu}))$ and the conditional density of $Z(s)$ given $Z(y), y \neq s$, is easily seen to be normal with expectation

$$E(Z(s)|Z(y) = z_y, y \in S \setminus \{s\}) = \mu_s - \frac{1}{a_{ss}} \sum_{y \neq s} a_{sy}(z_y - \mu_y)$$

and variance $1/a_{ss}$, where $\mu_y = E(Z(y))$ and a_{sy} denotes the element of the inverse of $\Sigma = (Cov(Z(s), Z(y)))_{s,y \in S}$. Therefore, a Gaussian RF is Markovian iff $a_{sy} \neq 0 \rightarrow y \in \mathcal{N}(s)$, i.e., iff Σ^{-1} is sparse. For a detailed account of GMRF we refer to Rue and Held (2005). According to the Hammersley–Clifford theorem (see Besag (1974)), MRF can be characterized as *Gibbs* RFs with local interaction potentials. The state space of a Gibbs random field can be rather general: $\mathbb{N}$ for count variables, e.g., in epidemiology, $\mathbb{R}^+$ for a positive-valued RF, e.g., a Gamma RF, a finite set of labels for categorical RFs, as e.g., in texture analysis, $\{0,1\}$ for binary RFs labeling presence or absence or alternative configurations as in *Ising models*, $\mathbb{R}^d$ for GRF, or mixtures of qualitative and quantitative states. Gibbs RFs are associated with families of conditional distributions p_Φ defined w.r.t. interaction potentials $\Phi = \{\phi_A, A \in \mathcal{S}\}$ where $\mathcal{S}$ is a family of finite subsets of S. In Bayesian image restoration, with $k > 2$ qualitative states (e.g., colours, textures or features) and finite set $S = \{0,1,\dots,255\}^2$ one often uses models of the form $p_\Phi(z) \propto \exp(-U(z))$ where U stands for the energy associated with Φ. In the simplest case one has only one interaction parameter β and $U(z) = \beta \cdot n(z)$, where $n(z)$ is the number of points of neighbouring sites with the same state. Here β plays the role of a regularization parameter: decreasing β leads to more regularity. The central goal in (Bayesian) image and signal processing is then to reconstruct an object z based on a noisy observation y from the posterior $p_\Phi(\cdot|y)$ of $\mathcal{Z}$ given y, e.g., on the basis of the MAP = maximum (mode) of the a posteriori distribution.

A good summary of the theory and applications of image data analysis based on the theory of random fields is given in Li (1995) and Winkler (2003); for description, classification and simulation of 3D-image data we refer to Ohser and Schladitz (2009).

About the Author

For biography *see* the entry ►Statistical Design of Experiments.

Cross References

►Agriculture, Statistics in
►Analysis of Areal and Spatial Interaction Data
►Environmental Monitoring, Statistics Role in
►Geostatistics and Kriging Predictors
►Model-Based Geostatistics
►Point Processes
►Poisson Processes
►Random Field
►Spatial Point Pattern

References and Further Reading

Anselin L, Rey SJ (eds) (2010) Perspectives on spatial data analysis. Springer, Berlin

Baddeley A, Gregori P, Mateu J, Stoica R, Stoyan D (eds) (2006) Case studies in spatial point process modeling. Lecture notes in statistics, Springer, New York

Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC Press, Boca Raton

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc B 36:192–236

Bivand RS, Pebesma EJ, Gomez-Rubio V (2008) Applied spatial data analysis with R. Springer, Berlin

Chilés J-P, Delfiner P (1999) Geostatistics. Modeling spatial uncertainty. Wiley, New York

Chilés J-P, Lantuéjoul Ch (2005) Prediction by conditional simulation: models and algorithms. In: Bilodeau M, Meyer F, Schmitt M (eds) Space structure and randomness. Lecture notes in statistics, vol 183. Springer, Berlin, pp 39–68

Cressie NAC (1993) Statistics for spatial data. Wiley, New York

Diggle PJ (2003) Statistical analysis of spatial point patterns, 2nd edn. Arnold, London

Diggle P (2010) Spatial pattern, this volume

Diggle PJ, Ribeiro PJ (2007) Model-based Geostatistics. Springer, New York

Finkenstädt B, Held L, Isham V (eds) (2007) Statistical methods for spatio-temporal systems. Chapman & Hall/CRC, Boca Raton

Gaetan C, Guyon H (2010) Spatial statistics and modeling. Springer, New York

Gneiting T, Genton MG, Guttorp P (2007) Geostatistical space-time models, stationarity, separability and full symmetry. In: Finkenstäedt B et al (eds) Statistical methods for spatio-temporal systems. Chapman & Hall/CRC, Boca Raton

Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modelling of spatial point patterns. Wiley, New York

Kazianka H, Pilz J (2010) Model-based geostatistics. this volume

Lantuéjoul Ch (2002) Geostatistical simulation. Models and algorithms. Springer, Berlin

Le ND, Zidek JV (2006) Statistical analysis of environmental space-time processes. Springer, New York
Li SZ (1995) Markov random field modeling in computer vision. Springer, Tokyo
Lloyd ChD (2007) Local models for spatial analyis. CRC Press, Boca Raton
Mase S (2010) Geostatistics and kriging predictors. this volume
Moklyachuk MP (2010) Random field. this volume
Møller J, Waagepetersen RP (2004) Statistical inference and simulation for spatial point processes. Chapman & Hall/CRC, Boca Raton
Møller J, Waagepetersen RP (2007) Modern statistics for spatial point processes. Scand J Stat 34:643–684
Ohser J, Mücklich F (2000) Statistical analysis of microstructures in materials science. Statistics in practice. Wiley, Chichester
Ohser J, Schladitz K (2009) 3D Images of materials structures. Wiley, Weinheim
Ripley BD (1981) Spatial statistics. Wiley, New York
Ripley BD (1988) Statistical inference for spatial processes. Cambridge University Press, Cambridge
Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman & Hall/CRC Press, Boca Raton
Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested laplace approximations. J R Stat Soc B 71:1–35
Sampson PD, Damien D, Guttorp P (2001) Advances in modelling and inference for environmental processes with nonstationary spatial covariance. In: Monestiez P, Allard D, Froidevaux R (eds) GeoENV III: Geostatistics for environmental applications. Kluwer, Dordrecht, pp 17–32
Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis. Chapman & Hall/CRC Press, Boca Raton
Serra J (ed) (1988) Image analysis and mathematical morphology. Theoretical advances. Academic, London
Spöck G, Pilz J (2010) Analysis of areal and spatial interaction data. this volume
Stein M (1999) Interpolation of spatial data. Springer, New York
Vere-Jones D (2010) Point processes, this volume
Webster R, Oliver MA (2007) Geostatistics for environmental scientists, 2nd edn. Wiley, Chichester
Winkler G (2003) Image analysis, random fields and Markov chain Monte Carlo methods: a mathematical introduction, 2nd edn. Springer, New York

Spectral Analysis

Peter Naeve
Professor Emeritus
University of Bielefeld, Bielefeld, Germany

Introduction

The term *spectral analysis* surely for most of us is connected with the experiment where a beam of sunlight is sent through a prism and split into many components of different colors, the spectrum. What looks nice is the starting point of a deeper understanding of nature, too.

The idea of splitting into components was copied by statisticians when working on time series. At first they proceeded like Kepler, who found his rules by fitting a model to data gathered by Tycho de Brahe. Deterministic modeling is a standard procedure in time series analysis. Given an economic time series x_t, one tries to fit $x_t = G_t + Z_t + S_t + R_t$ where G stands for trend, Z is a cyclic component, S a seasonal component, and R stands for the rest, the so-called noise. Regression is the important tool to study these models. The book by Davis still is a good starter. Unfortunately, this approach is not always as successful as with Kepler, "too many suns," Hotelling once complained.

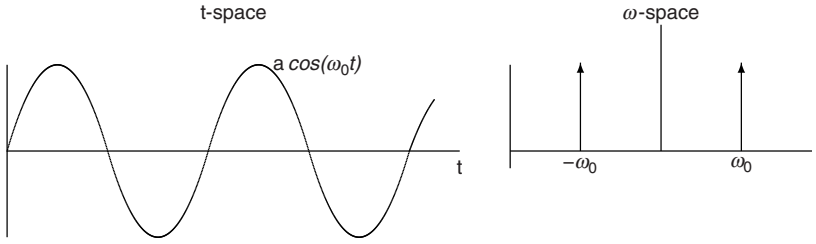
Quite another approach is to interpret a time series $\{x_t\}_{t\epsilon T}$ as a realization of a stochastic process $\{X(t)\}_{t\epsilon T}$. From now on we assume T to be a countable set. Then we might go in the direction of ARIMA-models – see, for instance, the book by Box and Jenkins – or choose spectral analysis as we will do here. So we are looking for a prism to work with.

A stochastic process is based on a system $F_n(u_1, \ldots, u_n; t_1, \ldots, t_n)$ of distribution functions. For these functions certain rules are valid, i.e., symmetric conditions $F_2(u_1, u_2; t_1, t_2) = F_2(u_2, u_1 : t_2, t_1)$, or consistency conditions such as $F_1(u_1; t_1) = F_2(u_1, \infty; t_1, t_2)$. Let E stand for the expectation operator. Then the mean function of the process is defined as $M(t) = E[X(t)]$ and the (auto-)covariance function as $C(t_1, t_2) = E[X(t_1)X(t_2)]$. A process is stationary if $M(t) = m$ and $C(t, s) = C(t-s) = C(\tau)$ for all $t, s\epsilon T$.

For such stationary processes the autocovariance function can be represented as $C(\tau) = \int e^{i\tau\omega} dF(\omega)$. The function $F(\omega)$ is called *spectral distribution*. When we have $dF(\omega) = f(\omega)d\omega$ the function $f(\omega)$ is called *spectral density*. The integration borders are $-\infty, \infty$ for continuous index set T and π, π for countable T. As can be seen by $C(0) = \int dF(\omega)$, the spectral distribution splits the variance into components. $dF(\omega)$ is the contribution to the variance of the frequencies in the interval between ω and $\omega + d\omega$. Such a stationary process can be written as $X(t) = \int e^{it\omega} dZ(\omega)$. For $\omega_1 \neq \omega_j$ $dZ(\omega_i)$, $dZ(\omega_j)$ are orthogonal random variables with $E[dZ(\omega)\overline{dZ(\omega)}] = dF(\omega)$. So the process $\{X(t)\}_{t\epsilon T}$ is split into orthogonal components $e^{it\omega} dZ(\omega)$.

What can be gained by spectral analysis may be seen by two simple examples.

Example 1 Firstly, take the process $\{X(t)\} = \{\xi \cos \omega_0 t + \eta \sin \omega_0 t\}$ where ξ and η are random variables with $E[\xi] = E[\eta] = 0$, $E[\xi^2] = E[\eta^2] = c$, and $E[\xi\eta] = 0$. The object is to get information about ω_0. The covariance function of this process is $C(\tau) = c \cos \omega_0 \tau$. In Fig. 1 the function C and the corresponding spectral density, $c\pi\{\delta(\omega - \omega_0) +$



Spectral Analysis. Fig. 1 Covariance function (left) Spectral density (right)

$\delta(\omega+\omega_0)\}$, demonstrate how the latter provides a much clearer picture of the structure of the process.

Example 2 Next let us take a stationary process $\{X(t)\}_{t\epsilon T}$ with autocovariance function $C_X(\tau)$ and spectral density $f_X(\omega)$. $Y(t)_{t\epsilon T}$ is a linear time invariant transformation of $\{X(t)\}_{t\epsilon T}$. If $w(t)$ is the impulse function of the transformation, we have $Y(t) = \int_{-\infty}^{\infty} w(\tau)X(t-\tau)d\tau$. Doing some mathematics, we get for the autocovariance function $C_Y(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(\tau_1)w(\tau_2)C_X(\tau-\tau_1-\tau_2)d\tau_1 d\tau_2$. Turning to the spectral densities of the processes, we get $f_Y(\omega) = |\phi(\omega)|^2 f_X(\omega)$, with $\phi(\omega) = \int_{-\infty}^{\infty} w(\tau)e^{i\tau\omega}d\tau$, a nice, simple multiplication of a spectral density with the square of a Fourier transform.

From now on we assume that we deal with discrete stationary processes. For these the covariance function $C(\tau) = \int_{-\pi}^{\pi} e^{i\tau\omega}f(\omega)d\omega$ and the spectral density $f(\omega) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} e^{-i\tau\omega}C(\tau)$ are a pair of Fourier transforms that are the base for further steps.

Estimation of the Spectral Density

In applications we usually don't have the full ensemble but only one member – a piece of a member – of the sample space. To go on, we have to assume that the process $\{X(t)\}_{t\epsilon T}$ is ergodic. That is, $\lim_{T_0\to 0}\frac{1}{T_0}\sum_{t=1}^{T_0}X(t) = E(X(t)$ (mean ergodic) and $\lim_{T_0\to 0}\frac{1}{T_0}\sum_{t=1}^{T_0}X(t+\tau)X(t) = E(X(t+\tau)X(t)$ (covariance ergodic). In both cases, the convergence is in quadratic mean. A simple sufficient condition for mean ergodic is $|C(\tau)| < \epsilon$, i.e., events far away are not correlated – might be true in many applications. For covariance ergodic the same must be true for the process $Z(t) = X(t+\tau)X(t)$.

To get an estimate for the spectral density there are two approaches. Either one starts with an estimate of the covariance function and take its Fourier transform as an estimate for the spectral density. Or one starts from the representation $X(t) = \int e^{it\omega}dZ(\omega)$ and $E[dZ(\omega)\overline{dZ(\omega)} = dF(\omega)$. The so-called periodogram $P_n(\omega) = \frac{1}{2\pi n}|\sum_{t=1}^{n}x(t)e^{it\omega}|^2$ combines these features. This approach is backed by the fast Fourier transform (FFT). Cooley and Tukey found this famous algorithm.

In each case, applying spectral analysis to time series of finite length leads to a lot of problems. So we only have estimates $C(\tau)$ for $|\tau| \le \tau_0$. Theory calls for an estimator for all τ. A function $L(\tau)$ with $L(0) = 1$, $L(\tau) = L(-\tau)$ for $|\tau| \le \tau_0$, and $L(\tau) = 0$ elsewhere may be a solution. $\tilde{C}(\tau) = L(\tau)C(\tau)$ is defined for all τ. Further problems emerge immediately. How does one choose τ_0? Is this estimator unbiased, consistent? What is a good $L(\tau)$? And so on. Theoretically, these questions are hard to solve. Simulation is an aid in studying these problems. The book by Jenkins and Watts may be a good introduction to this approach.

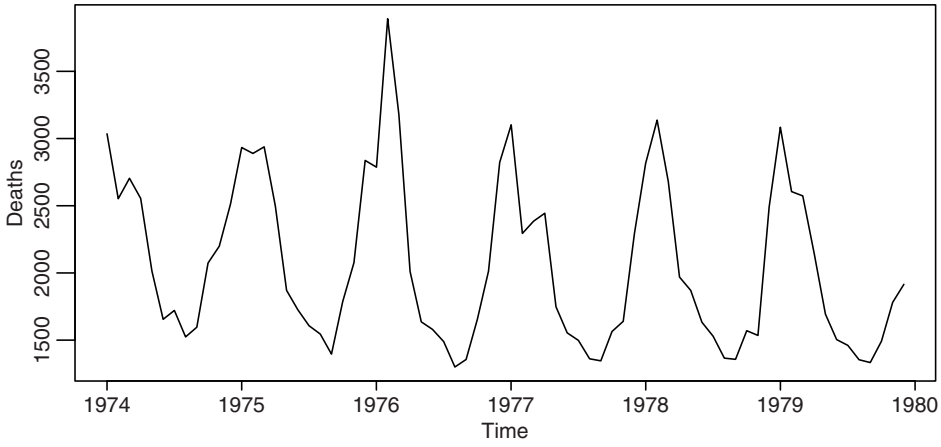
Multivariate Spectral Analysis

The simplest cases of multiple spectral analysis are two stochastic processes, $\{X(t)\}_{t\epsilon T}$ and $\{Y(t)\}_{t\epsilon T}$. The base of our analysis is the cross-variance function $C_{XY}(t_1,t_2) = E[X(t_1)Y(t_2)] = C_{XY}(t_1-t_2)$. For this function we have the representation $C_{xy}(\tau) = \int e^{i\tau\omega}dF_{XY}(\omega)$. From $C_{xy}(\tau) = \int e^{i\tau\omega}dF_{XY}(\omega)$ we get the complex cross-spectral density $f_{XY}(\omega) = k(\omega) + iq(\omega)$ $k(\omega)$ is called co-spectrum and $q(\omega)$ quadrature spectrum. A number of functions are based on these two spectra, e.g., the amplitude $A(\omega) = \sqrt{\{k(\omega)\}^2 + \{q(\omega)\}^2}$, the phase $\phi(\omega) = \arctan(q(\omega)/k(\omega))$, and the coherence $C(\omega) = \frac{A(\omega)}{f_X(\omega)f_Y(\omega)}$. Plots of these functions are nice tools to study the relation between $\{X(t)\}_{t\epsilon T}$ and $\{Y(t)\}_{t\epsilon T}$.

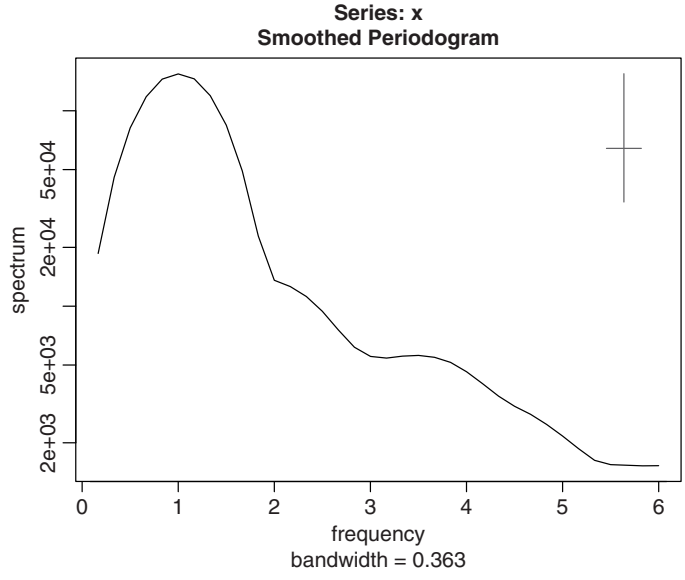
An Application

Finally we will deal with an application of spectral methods. This example is a very short version taken from the book by Venables and Ripley p. 355 f. The details are shown in Figs. 2 and 3. Figure 2 depicts the time series of monthly deaths from lung diseases in the UK 1974–1979. Figure 3 shows one estimate of the spectrum. All calculation were done with R. The function spectrum is based on FFT and smoothing by running means.

The interpretation of spectral functions and graphs calculated in applications is not an easy task. The book by Granger – the late Nobel Prize winner – might be a good starting place.



Spectral Analysis. Fig. 2 Time series



Spectral Analysis. Fig. 3 Spectrum

About the Author

Peter Naeve is Professor Emeritus at the University of Bielefeld since 2002. From 1979 on he held the Chair in Statistics and Computer Science at the Department of Economics. He is member of the International Statistical Institute, American Statistical Association, Royal Statistical Society, and International Association for Statistical Computing (IASC). His main field of interest is Computational Statistics. From the very beginning he was involved in the essential activities in this field, i.e., establishing an organized community (IASC), implementing a series of meetings (COMPSTAT Symposium on Computational Statistics), and providing a journal for this field. Among other positions he served as Co-Editor for the journal *Computational Statistics and Data Analysis* (1991–2000).

Cross References

►Box–Jenkins Time Series Models
►Stochastic Processes
►Stochastic Processes: Classification
►Time Series
►Time Series Regression

References and Further Reading

It is hard to sample a list of references from hundreds of books and an almost uncountable set of articles and discussion papers.

Being a good starting place was the (personally biased) criterion. When I entered the field, the books by Granger, Hatanka and Blackman, and Tukey were my first guides.
Blackman RB, Tukey JW (1959) The measurement of power spectra. Dover, New York
Box GEP, Jenkins GM (1970) Time series analysis forecasting and control. Holden-Day, San Francisco
Cooley JW, Tukey JW (1965) An algorithm for the machine calculation of complex Fourier series. Math Comp 19:297–301
Davis HT (1963) The analysis of economic time series. Principia, San Antonio
Granger CWJ, Hatanaka M (1964) Spectral analysis of economic time series. Princeton University Press, Princeton
Jenkins GM, Watts DG (1968) Spectral analysis and Its applications. Holden-Day, San Francisco
Venables WN, Ripley BD (1994) Modern applied statistics with S-plus. Springer, New York
To assist those who like to Google, here are the names of some other pioneers: Bartlett, Parzen, Hannan, Priestley, Brillinger, Rosenblatt, Bingham

Sport, Statistics in

Stephen R. Clarke[1], John M. Norman[2]
[1]Professor
Swinburne University, Melbourne, VIC, Australia
[2]Emeritus Professor
Sheffield University, Sheffield, UK

Fans love statistics about sport – sets of numbers that describe and summarise what is happening on the field. With developments in computer technology, global positioning systems and the internet, the range and availability of sports statistics is growing at a rapid rate. In tennis majors, for example, an on-court statistician enters the result of every rally, whether the final shot was a forehand or backhand drive or volley, a winner or forced or unforced error, and whether either or both players were at the net. Cumulative results are immediately available to spectators, the media, and the general population through the internet. Only a few years ago, the number of kicks marks and handballs each player obtained in an Australian Rules football match was provided in printed tables two days after the match. Now over 80 statistics are collected in real time and immediately available to coaches and the general public. The science of statistics can be used to add value, to make sense, to discern patterns, to separate random variation from underlying trends in these sports data.

We are discussing here not just the collection and accumulation of statistics, but statistical modeling. Collection of raw statistics is one thing (how long is it since a batsman made over 400 in an international match? how old was Stanley Matthews when he played his last soccer game?) and statistical modeling (how can statistics be used) by analysts is another. If we are interested in the chance a male player might break 60 in a golf tournament next year, past statistics might tell us the percentage of all tournament rounds in which this has occurred. But if we want to estimate the chance Tiger Woods will break 60 in the US masters next year, this is of little use. We need to do some modeling. For example we might use past statistics to obtain Tiger's scores on each hole in previous masters, and by sampling from these use simulation to get a useful estimate.

Cricket has the distinction of being the first sport used for the illustration of statistics. In *Primer in Statistics*, (Elderton and Elderton 1909) used individual scores of batsmen to illustrate frequency distributions and elementary statistics. Some previous work on correlation and consistency resulted in (Wood 1945) and (Elderton 1945) reading separate papers at the same meeting of the Royal Statistical Society. These papers investigated the distribution of individual and pairs of batsmen scores, and have some claim as the first full quantitative papers applying statistics to sport.

The literature now contains hundreds of papers detailing applications of statistical modeling in virtually every sport. Researchers in the area are not confined to Statisticians. Other disciplines include Mathematics, Operational research, Engineering, Economics and Sports Science. Learned societies such as the American Statistical Association, the Australian Mathematical Society and the Institute of Mathematics and its Applications have sections of their membership or conferences devoted to this area. The range of journals which publish articles on sport often makes it difficult to search for previous work in a particular topic.

Much early work in the area is covered in the two texts (Machol et al. 1976) and (Ladany and Machol 1977). More recently (Bennett 1998) gives an excellent overview, with chapters on particular sports: American football, baseball, basketball, cricket, soccer, golf, ice hockey, tennis, track and field; and theme chapters on design of tournaments, statistical data graphics, predicting outcomes and hierarchical models. Later collections of papers include (Butenko et al. 2004) and (Albert and Koning 2008). These provide good examples of the issues currently being investigated by researchers. We discuss here some of these issues.

As mentioned above, fitting known distributions to sporting data was amongst the earliest work performed in this area. If the performance data follow a known distribution, that tells you something about the underlying behavior of the sportsman. If a batsman's cricket scores follow an exponential (or geometric) distribution, then he has a constant hazard, or probability of dismissal, throughout

his innings. If the number of successful shots a basketball player makes in a given number of tries can be modeled by the ▶Binomial distribution, then he has a constant probability of success, and is not affected by previous success or failure. If goals scored each match by a soccer team are Poisson distributed, this implies their form is not variable throughout the season, and they are not affected by early success or failure in a match. Departures from known distributions can be used to investigate the existence of the "hot hand" in basketball or baseball, or "momentum" in tennis or soccer.

Predicting the outcomes of sporting contests is of great interest to modelers and fans alike. Statistical modelers are usually interested in not only predicting the winner, but in estimating the chance of each participant winning and likely scores or margins. These predictions have become increasingly important with the introduction of sports betting. The estimated chances developed from the statistical model can be compared with the bookmaker's odds, and inefficiencies of betting markets investigated (or exploited). If the probabilities of head to head encounters can be estimated, then the chances of various outcomes of whole tournaments or competitions can be estimated via simulation.

A usual by-product of prediction is the rating of individuals or teams. For example a simple model might predict the winning margin between two teams as the difference in their ratings plus a home advantage. ▶Least squares, maximum likelihood or other methods are then used to obtain the ratings and home advantage that give the best fit to previous results. Chess has a rating system based on exponential smoothing that is applicable to past and present players from beginners to world champions. In golf, much effort has gone into developing ratings of players (handicaps) that are fair to players of all standards from all courses.

Home advantage, the degree to which a team performs better at home than away, is present in most sports. (Stefani and Clarke 1992) show that in balanced competitions the home side wins anywhere from 54% (baseball) to 70% (international soccer) of the matches. In scoring terms 1 goal in 3 in international soccer can be attributed to home advantage, while in baseball the home advantage contributes 1 run in 34. While home advantage can be quantified it is more difficult to isolate its causes. Many papers have looked at the effects of travel, crowd, ground familiarity and referee bias without much consensus. Other research has shown that models assuming a different home advantage for different teams or groups of teams provide a better fit to the data than ones with a common home advantage.

There are many different scoring systems in sport, (for example in racquet sports), and researchers are interested in their operating characteristics. To what extent do the scoring systems affect the probabilities of each player winning, and the distribution of the number of rallies in the match? What is the chance of winning from any scoreline? Generally the longer the match the more chance for the better player. For example, a player who wins 52% of the points at tennis, will win 55% of the games, 64% of the sets and 75% of 5 set matches. But the few breaks of serve in men's tennis makes the scoring system relatively inefficient. The better player may win a higher percentage of his serves than his opponent, but the set score still reaches 6 all. Researchers have suggested alternative scoring systems, such as 4-3 tennis, where the server still has to win 4 points to win the game, but the receiver only has to win 3 points. They have also looked at the importance of points – the change in a player's chance of winning the game (or match) resulting by winning or losing the point. (In tennis the most important point in a game is the service break point). The assertion that better players win the important points can then be tested.

What often makes sport interesting is the choice of alternative strategies. Should a baseball player try and steal a base or not? Should a footballer try for a field goal or a touchdown? Should a tennis player use a fast or slow serve? Should an orienteer choose a short steep route or a longer flatter one? When should the coach pull the goalie in ice-hockey? Operational Researchers find this a fertile field for study (Wright 2009), with techniques such as Dynamic Programming and simulation used to determine optimal strategies. (Norman 1995) gives one example of the use of Dynamic Programming in each of 12 sports.

Sport is an important area for the application of statistical modeling. Sport is big business, and occupies an important role in today's society. By the use of a range of modeling and analysis techniques Statisticians can assist players, coaches, administrators and fans to better understand and improve their performance and enjoyment.

About the Authors

Dr. Stephen Clarke is a Professor of Statistics in the faculty of Life and Social Sciences at Swinburne University, Melbourne, Australia. He has authored and co-authored more than 130 papers. He received the (U.K.) Operational Research Society president's medal in 1989 for his paper on one-day cricket.

John M. Norman is an emeritus professor at Sheffield University Management School, UK. He has written two books and fifty papers, several in collaboration with Stephen Clarke.

Cross References

►Binomial Distribution
►Poisson Distribution and Its Application in Statistics
►Record Statistics
►Testing Exponentiality of Distribution

References and Further Reading

Albert J, Koning RH (eds) (2008) Statistical thinking in sports. Chapman & Hall, Boca Raton
Bennett J (ed) (1998) Statistics in sport. Arnold, London
Butenko S, Gil-Lafuente J et al (eds) (2004) Economics, management and optimization in sports. Springer-Verlag, Berlin
Elderton WE (1945) Cricket scores and some skew correlation distributions. J Roy Stat Soc (Ser A) 108:1–11
Elderton WP, Elderton EM (1909) Primer of statistics. Black, London
Ladany SP, Machol RE (1977) Optimal strategies in sports. North Holland, Amsterdam
Machol RE, Ladany SP et al (1976) Management science in sports. North Holland, New York
Norman JM (1995) Dynamic programming in sport: a survey of applications. IMA J Math Appl Bus Ind 6(December):171–176
Stefani RT, Clarke SR (1992) Predictions and home advantage for Australian rules football. J Appl Stat 19(2):251–261
Wood GH (1945) Cricket scores and geometrical progression. J Roy Stat Soc (Ser A) 108:12–22
Wright MB (2009) 50 years of OR in sport. J Oper Res Soc 60(S1):S161–S168

Spreadsheets in Statistics

Rade Stankic, Jasna Soldic-Aleksic
Professors, Faculty of Economics
Belgrade University, Belgrade, Serbia

Spreadsheet is a computer program that manipulates tables consisting of rows and columns of cells. It transforms a computer screen into a ledger sheet or grid of coded rows and columns simulating a paper worksheet. The program environment consists of one or more huge electronic worksheets (each worksheet can contain up to one million rows by a few thousands columns) organized in the form of an electronic workbook.

The general features of such programs are powerful computing and graphical capabilities, flexibility, excellent report generating feature, easy-to-use capability, and compatibility with many other data analytical software tools. These features are responsible for the substantial popularity and wide practical usage of the program. Thus, spreadsheet software is being used in academic, government, and business organizations for tasks that require summarizing, reporting, data analysis, and business modeling.

The spreadsheet concept became widely known in the late 1970s and early 1980s due to the Dan Bricklin's implementation of VisiCalc which is considered to be the first electronic spreadsheet. It was the first spreadsheet program that combined all essential features of modern spreadsheet applications, such as: WYSIWYG (*What You See Is What You Get*), interactive user interface, automatic recalculation, existence of status and formula lines, copy of cell range with relative and absolute references, and formula building by selecting referenced cells. Lotus 1–2–3 was the leading spreadsheet program in the period when DOS (Disk Operating System) prevailed as an operating system. Later on, Microsoft Excel took the lead and became the dominant spreadsheet program in the commercial electronic spreadsheet market.

The basic building blocks of a spreadsheet program are cells that represent the intersections of the rows and columns in a table. Each individual cell in the spreadsheet has a unique column and row identifier that takes specific forms in different spreadsheet programs. Thus, the top left-hand cell in the worksheet may be designated with symbols A1, 11, or 1A. The content of the cell may be a value (numerical or textual data) or a formula. When the formula is entered in a particular cell, it defines how the content of that cell is calculated and updated depending on the content of another cell (or combination of cells) that is/are referenced to in the formula. References can be relative (e.g., `A1`, or `C1:C3`), absolute (e.g., `$B$1`, or `$C$1:$C$3`), mixed row-wise or column-wise absolute/relative (e.g., `$B1` is column-wise absolute and `B$1` is row-wise absolute), three-dimensional (e.g., Sheet1!A1), or external (e.g., [Book1]Sheet1!A1). This well-defined structure of cell addresses enables a smooth data flow regardless whether data are stored in just one or several worksheets or workbooks. In most implementations, a cell (or range of cells) can be "named" enabling the user to refer to that cell (or cell range) by its name rather than by grid reference. Names must be unique within a spreadsheet, but when using multiple sheets in a spreadsheet file, an identically named cell range on each sheet can be used if it is distinguished by adding the sheet name. Name usage is primarily justified by the need for creating and running macros that repeat a command across many sheets.

What makes the spreadsheet program a powerful data analytical tool is the wide range of integrated data processing functions. Functions are organized into logically distinct groups, such as: *Arithmetic functions, Statistical functions*, *Logical functions*, *Financial functions*, *Date and Time functions*, *Text functions*, *Information*, *Mathematical function*, etc. In general, each function is determined by its name (written in uppercase by convention) and

appropriate argument(s) which is/are placed in parenthesis. The arguments are a set of values, separated by semicolons, to which the function applies. Thus, a function called *FUNCTION* would be written as follows: *FUNCTION* (argument1; argument2; etc.).

Spreadsheet software integrates a large number of built-in statistical functionalities, but some caveats about its statistical computations have been observed. A few authors have criticized the use of spreadsheets for statistical data processing and have presented some program shortcomings, such as: no log file or audit trail, inconsistent behavior of computational dialogs, poor handling of missing values, low-level of accuracy of built-in spreadsheet statistical calculations, and no sophisticated data coding techniques for specific statistical calculations. In response to such criticism directed against the statistical "incorrectness" and limitations of spreadsheet programs, many efforts have been made (both in the academic and commercial community) to compensate for them. Thus, many statistics add-ins have appeared, granting robust statistical power to the spreadsheet program environment. These add-ins are usually seamlessly integrated into a spreadsheet program and cover the range of most commonly used statistical procedures, such as: descriptive statistics, ►normality tests, group comparisons, correlation, regression analysis, forecast, etc. Some leading statistical software vendors have provided statistical modules and functionalities for spreadsheet users. For example, the statistical software package PASW Statistics 17.0 offered the following additional techniques and features for Excel spreadsheet program (*SPSS Advantage for Excel 2007*): Recency, Frequency, and Monetary value (RFM) analysis for direct marketing research (where most profitable customers are identified), classification tree analysis for group identification, unusual data detection, procedure for data preparation and transformation, and the option to save spreadsheet data as a statistical software data file.

One of the crucial spreadsheet package features is its capability to carry out "What-if" data analysis. "What-if" analysis is the process of observing and learning how the changes in some cells (as an input) affect the outcome of formulas (as an output) in the other cells in the worksheet. For example, Microsoft Excel provides the following "what-if" analytical tools: scenario manager, data tables, and Goal Seek. Scenario manager and data tables operate in a very simple way: they take sets of input values and determine possible results. While a data table works only with one or two variables, accepting many different values for those variables, a scenario manager can handle multiple variables, but has a limitation of accommodating only up to 32 values. These tools are appropriate for running the *sensitivity analysis*, which determines how a spreadsheet's output varies in response to changes to the input values. Contrary to the functioning of scenario manager and data tables, Goal Seek allows the user to compute a value for a spreadsheet input that makes the value of a given formula match a specified goal.

In the era of the Internet, networked computing, and web applications, online spreadsheet programs also came about. An online spreadsheet is a spreadsheet document edited through a web-based application that allows multiple users to have access, to edit and to share it online (multiple users can work with a spreadsheet, view changes in real time, and discuss changes). Equipped with a rich Internet application user interface, the best web-based online spreadsheets have many of the features seen in desktop spreadsheet applications and some of them have strong multiuser collaboration features. Also, there are spreadsheet programs that offer real time updates from remote sources. This feature allows updating of a cell's content when its value is derived from an external source - such as a cell in another "remote" spreadsheet. For shared, web-based spreadsheets, this results in the "immediate" updating of the content of cells that have been altered by another user and, also, in the updating of all dependent cells.

Cross References

►Statistical Software: An Overview

References and Further Reading

Albright SC, Winston WL, Zappe CJ (2009) Data analysis and decision making with Microsoft Excel, 3rd edn. South-Western Cengage Learning, USA

Monk EF, Davidson SW, Brady JA (2010) Problem-solving cases in Microsoft Access and Excel. Course technology. Cengage Learning, USA

Nash JC (2006) Spreadsheets in statistical practice – another look. Am Stat 60(3):287–289

Turban E, Ledner D, McLean E, Wetherbe J (2007a) Information technology for management: transforming organizations in the digital economy, 6th edn. Wiley, New York

Walkenbach J (2007) Excel *2007 Bible*. Wiley, New York

Winston WL (2004) Microsoft Excel – data analysis and business modeling. Microsoft, Washington

Spurious Correlation

Simon J. Sheather
Professor and Head of Department of Statistics
Texas A&M University, College Station, TX, USA

A well-known weakness of regression modeling based on observational data is that the observed association between

two variables may be because both are related to a third variable that has been omitted from the regression model. This phenomenon is commonly referred to as "spurious correlation." The term spurious correlation dates back to at least Pearson (1897).

Neyman (1952, pp. 143–154) provides an example based on fictitious data which dramatically illustrates spurious correlation. According to Kronmal (1993, p. 379), a fictitious friend of Neyman was interested in empirically examining the theory that storks bring babies and collected data on the number of women, babies born and storks in each of 50 counties. This fictitious data set was reported in Kronmal (1993, p. 383) and it can be found on the web page associated with Sheather (2009), namely, http://www.stat.tamu.edu/~sheather/book.

Figure 1 shows scatter plots of all three variables from the stork data set along with the least squares fits. Ignoring the data on the number of women and fitting the following straight-line regression model produces the output shown below.

$$\text{Babies} = \beta_0 + \beta_1 \text{Storks} + e \tag{1}$$

The regression output for model (1) shows that there is very strong evidence of a positive linear association between the number of storks and the number of babies born (p-value < 0.0001). However, to date we have ignored the data available on the other potential predictor variable, namely, the number of women.

Regression output for model (1)				
	Coefficients			
	Estimate	Std. Error	*t* value	Pr(> |*t*|)
(Intercept)	4.3293	2.3225	1.864	0.068
Storks	3.6585	0.3475	10.528	1.71e-14 ***
Residual standard error: 5.451 on 52 degrees of freedom				
Multiple *R*-Squared: 0.6807, Adjusted *R*-squared: 0.6745				

Next we consider the other potential predictor variable, namely, the number of women. Thus, we consider the following regression model:

$$\text{Babies} = \beta_0 + \beta_1 \text{Storks} + \beta_2 \text{Women} + e \tag{2}$$

Given below is the output from *R* for a regression model (2). Notice that the estimated regression coefficient for the number of storks is zero to many decimal places. Thus, correlation between the number of babies and the number of storks calculated from (1) is said to be spurious as it is due to both variables being associated with the number of women. In other words, a predictor (the number of women) exists which is related to both the other predictor (the number of storks) and the outcome variable (the number of babies), and which accounts for all of the observed association between the latter two variables. The number of women predictor variable is commonly called either an omitted variable or a confounding covariate.

Regression output for model (2)				
	Coefficients			
	Estimate	Std. Error	*t* value	Pr(> |*t*|)
(Intercept)	1.000*e*+01	2.021*e*+00	4.948	8.56*e*−06***
Women	5.000*e*+00	8.272*e*−01	6.045	1.74*e*−07***
Storks	−6.203*e*−16	6.619*e*−01	−9.37*e*−16	1
Residual standard error: 4.201 on 51 degrees of freedom				
Multiple *R*-Squared: 0.814, Adjusted *R*-squared: 0.8067				

We next briefly present some mathematics wish quantifies the effect of spurious correlation due to omitted variables. We shall consider the situation in which an important predictor is omitted from a regression model. We shall denote the omitted predictor variable by v and the predictor variable included in the one-predictor regression model by x. In the fictitious stork data x corresponds to the number of storks and v corresponds to the number of women.

To make things as straightforward as possible we shall consider the situation in which Y is related to two predictors x and v as follows:

$$Y = \beta_0 + \beta_1 x + \beta_2 v + e_{Y \cdot x,v} \tag{3}$$

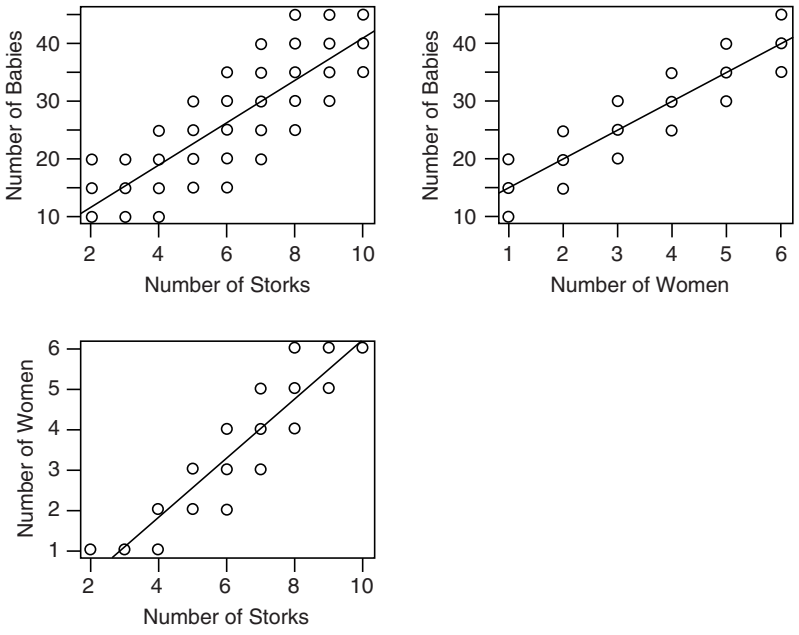
Similarly, suppose that v is related to x as follows:

$$v = \alpha_0 + \alpha_1 x + e_{v \cdot x} \tag{4}$$

Substituting (4) into (3) we will be able to discover what happens if omit v from the regression model. The result is as follows:

$$Y = (\beta_0 + \beta_2 \alpha_0) + (\beta_1 + \beta_2 \alpha_1) x + (e_{Y \cdot x,v} + \beta_2 e_{v \cdot x}) \tag{5}$$

Notice that the regression coefficient of x in (5) is the sum of two terms, namely, $\beta_1 + \beta_2 \alpha_1$.



Spurious Correlation. Fig. 1 A plot of the variables from the fictitious data on storks

We next consider two distinct cases:

1. $\alpha_1 = 0$ and/or $\beta_2 = 0$: Then the omitted variable has no effect on the regression model, which includes just x as a predictor.
2. $\alpha_1 \neq 0$ and $\beta_2 \neq 0$: Then the omitted variable has an effect on the regression model, which includes just x as a predictor. For example, Y and x can be strongly linearly associated (i.e., highly correlated) even when $\beta_1 = 0$. (This is exactly the situation in the fictitious stork data.) Alternatively, Y and x can be strongly negatively associated even when $\beta_1 > 0$.

Spurious correlation due to omitted variables is most problematic in observational studies. We next look at a real example, which exemplifies the issues. The example is based on a series of papers (Cochrane et al. 1978; Hinds 1974; Jayachandran and Jarvis 1986) that model the relationship between the prevalence of doctors and the infant mortality rate. The controversy was the subject of a 1978 Lancet editorial entitled "The anomaly that wouldn't go away." In the words of one of the authors of the original paper, Selwyn St. Leger (2001):

> When Archie Cochrane, Fred Moore and I conceived of trying to relate mortality in developed countries to measures of health service provision little did we imagine that it would set a hare running 20 years into the future... The hare was not that a statistical association between health service provision and mortality was absent. Rather it was the marked positive correlation between the prevalence of doctors and infant mortality. Whatever way we looked at our data we could not make that association disappear. Moreover, we could identify no plausible mechanism that would give rise to this association.

Kronmal (1993, p. 624) reports that Sankrithi et al. (1991) found a significant negative association ($p < 0.001$) between infant mortality rate and the prevalence of doctors after adjusting for population size. Thus, this spurious correlation was due to an omitted variable. In summary, the possibility of spurious correlation due to omitted variables should be considered when the temptation arises to over interpret the results of any regression analysis based on observational data. Stigler (2005) advises that we "discipline this predisposition (to accept the results of observational studies) by a heavy dose of skepticism."

About the Author

Professor Sheather is Head of the Department of Statistics, Texas A&M University. Prior to that he was the Head of the Statistics and Operations Group and Associate Dean of Research, Australian Graduate School of Management, at the University of New South Wales in Sydney, Australia. He is an Elected Fellow of the American Statistical Association (2001). Professor Sheather is currently listed on

ISIHighlyCited.com among the top one-half of one percent of all mathematical scientists, in terms of citations of his published work.

Cross References

►Causation and Causal Inference
►Confounding and Confounder Control
►Correlation Coefficient
►Data Quality (Poor Quality Data: The Fly in the Data Analytics Ointment)
►Role of Statistics in Advancing Quantitative Education

References and Further Reading

Cochrane AL, St. Leger AS, Moore F (1978) Health service "input" and mortality "output" in developed countries. J Epidemiol Community Health 32:200–205
Hinds MW (1974) Fewer doctors and infant survival. New Engl J Med 291:741
Jayachandran J, Jarvis GK (1986) Socioeconomic development, medical care and nutrition as determinants of infant mortality in less developed countries. Social Biol 33:301–315
Kronmal RA (1993) Spurious correlation and the fallacy of the ratio standard revisited. J R Stat Soc A 156:379–392
Neyman J (1952) Lectures and conferences on mathematical statistics and probability, 2nd edn. US Department of Agriculture, Washington DC, pp 143–154
Pearson K (1897) Mathematical contributions to the theory of evolution: on a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond 60:489–498
Sankrithi U, Emanuel I, Van Belle G (1991) Comparison of linear and exponential multivariate models for explaining national infant and child mortality. Int J Epidemiol 2:565–570
Sheather SJ (2009) A modern approach to regression with R. Springer, New York
St. Leger S (2001) The anomaly that finally went away? J Epidemiol Community Health 55:79
Stigler S (2005) Correlation and causation: a comment. Persp Biol Med 48(1 Suppl.):588–594

St. Petersburg Paradox

James M. Joyce
Chair and Professor of Philosophy and of Statistics
University of Michigan, Ann Arbor, MI, USA

The St. Petersburg "Paradox" concerns a betting situation in which a gambler's fortune will be increased by $\$2^n$ if the first tail appears on the *nth* toss a fair coin. Nicholas Bernoulli introduced this problem in 1713 as a challenge to the then prevailing view that the fair price of a wager (the price at which one should be equally happy to buy or sell it) is equal to its expected monetary payoff. While Bernoulli's wager has an infinite expected payoff, any reasonable person will sell it for $20. By 1727 Gabriel Cramer had recognized that the prevailing view goes wrong because it assumes that people value money linearly. As he wrote, "mathematicians evaluate money in proportion to its quantity while, in practice, people with common sense evaluate money in proportion to the (practical value) they can obtain from it" (Bernoulli 1954, p. 33). Since an extra increment of money buys less happiness for a prince than a pauper, Cramer observed, the St. Petersburg wager can have a finite "practical value" provided that the worth of an extra dollar falls off rapidly enough as a person's fortune grows. In modern terms, Cramer had understood that money has declining marginal utility and that the St. Petersburg wager can have a finite expected utility if the marginal decrease in utility is sufficiently steep. He noted, for example, that a utility function of the form $u(\$x) = x^{1/2}$ produces an expected utility of $\sum_n \left(\frac{1}{2}\right)^n 2^{n/2} \approx 2.41421$ for Bernoulli's wager, which is equivalent to a fair price of $5.83.

Cramer never published, and it was left to Daniel Bernoulli to report Cramer's contributions and to write the definitive treatment (1954) of his cousin Nicholas's problem in the *St. Petersburg Academy Proceedings* of 1738, from which the Paradox derives its name. Daniel, who hit upon the declining utility of money independently of Cramer, went further by advocating the general principle that rational agents should value wagers according to their expected utility. He also argued that a person's marginal utility for an extra sum of money should be both inversely proportional to the person's fortune and directly proportional to the size of the sum. This means that the utility of $x is a function of the form $u(\$x) = k \cdot \ln(x)$. When evaluated using such a utility function, the St. Petersburg wager has a finite expected utility of $k \cdot \ln(4)$.

Bernoulli was also explicit that, as a general matter, the value of any gamble is its expected utility, and not its expected payoff. Specifically, he maintained that if the utility function $u(x)$ measures the "practical value" of having fortune $x, then the value of any wager X is $E(u(X)) = \int_0^1 P(X = x) \cdot u(x)dx$ and its fair price is that sum $\$f$ such that $u(f) = E(u(X))$. Though this was perhaps Bernoulli's deepest insight, its implications were not fully appreciated until the early 1950s when the work of Savage (1954) and von Neumann and Morgenstern (1953) moved the hypothesis of expected utility maximization to the very center of both microeconomics and ►Bayesian statistics.

Until that time, Bernoulli was better known among economists and statisticians for postulating that money has declining marginal utility and for solving the St. Petersburg

Paradox. The thesis that money has declining marginal utility has been immensely influential since it serves as the basis for the standard theory of risk aversion, which explains a wide variety of economic phenomena. In economic parlance, a *risk averse* agent prefers a straight payment of a gamble's expected payoff to the gamble itself. Economists seek to explain risk aversion by postulating *concave* utility functions for money, with greater concavity signaling more aversion. If $u(x)$ is concave for $a \leq x \leq b$, and if a wager X's payouts are confined to $[a, b]$, then it is automatic that $E(u(X)) \geq u(E(X))$. Moreover, if v is a concave transformation of u, the absolute risk aversion associated with v exceeds that associated with u, where absolute risk aversion is measured by the Arrow (1965)–Pratt (1964) coefficient $v''(x)/v'(x)$. Agents with Bernoulli's logarithmic utility are everywhere risk averse, and their absolute level of risk aversion decreases with increases in x since $u''(x)/u'(x) = 1/x$.

Interestingly, the Cramer/Bernoulli solution to the St. Petersburg Paradox failed the test of time. As Karl Menger (1934) first recognized (Basset 1987), if money has *unbounded* utility then one can always construct a "Super St. Petersburg Paradox." For example, using $u(\$x) = \ln(x)$, a wager that pays $e^2, e^4, e^8, \ldots$ if a tail appears first on the 1st, 2nd, 3rd,... toss will have infinite expected utility. One can avoid this either by insisting that realistic utility functions are bounded or by restricting the allowable gambles so that events of high utility are always assigned such low probabilities that gambles with infinite expected utilities never arise. On either view, the St. Petersburg Paradox ceases to be a problem since there is no chance that anyone will ever face it. Most standard treatments, e.g., (Ingersoll 1978), endorse bounded utility functions on the grounds that arbitrarily large payoffs are impossible in a finite economy. Others, who want to leave open the theoretical possibility of unbounded utility, require all realizable wagers to be limits of wagers with uniformly bounded support, where limits are taken in the weak topology. For a well-developed approach of this sort see (Kreps 1988, pp. 63–68).

About the Author

James M. Joyce is Professor of Philosophy and Statistics at the University of Michigan, Ann Arbor. He is the author of *The Foundations of Causal Decision Theory* (Cambridge Studies in Probability, Induction and Decision Theory, Cambridge University Press, 1999), as well as a number of articles on decision theory and Bayesian approaches to epistemology and the philosophy of science.

Cross References

►Statistics and Gambling

References and Further Reading

Arrow KJ (1965) Aspects of the theory of risk-bearing. Markham, Chicago

Basset GW (1987) The St. Petersburg paradox and bounded utility. Hist Polit Econ 19:517–523

Bernoulli D (1738) Specimen theoriae de mensura sortis. Commentarii academiae scientiarum imperialis petropolitanae. In: Proceedings of the royal academy of science, St. Petersburg. English translation (1954) by Louise Sommer with notes by Karl Menger. Exposition of a new theory on the measurement of risk. Econometrica 22:23–36

Ingersoll J (1978) Theory of financial decision making. Rowman and Littlefield, Oxford

Kreps D (1988) Notes on the theory of choice. Westview, Boulder

Menger K (1934) Das unsicherheitsmoment in der wertlehre. Zeitschrift für Nationalöcokonomie 51:459–485

Pratt JW (1964) Risk aversion in the small and in the large. Econometrica 32:122–136

Savage LJ (1954) The foundations of statistics. Wiley, New York

von Neumann J, Morgenstern O (1953) Theory of games and economic behavior, 3rd edn. Princeton University Press, Princeton

Standard Deviation

Sekander Hayat Khan M.
Professor of Statistics
Institute of Statistical Research and Training
University of Dhaka, Dhaka, Bangladesh

Introduction

Standard deviation is a measure of variability or dispersion. The term *Standard deviation* was first used in writing by Karl Pearson in 1894. This was a replacement for earlier alternative names for the same idea: for example, "mean error" (Gauss), "mean square error," and "error of mean square" (Airy) have all been used to denote standard deviation. Standard deviation is the most useful and most frequently used measure of dispersion. It is expressed in the same units as the data. Standard deviation is a number between 0 and ∞. A large standard deviation indicates that observations/data points are far from the mean and a small standard deviation indicates that they are clustered closely around the mean.

Definition

If X is a random variable with mean value $\mu = E(x)$, the standard deviation of X is defined by

$$\sigma = \sqrt{E(X-\mu)^2}. \quad (1)$$

That is, the standard deviation σ is the square root of the average value of $(X-\mu)^2$. The standard deviation of a continuous real-valued random variable X with probability density function $f(x)$ is

$$\sigma = \sqrt{\int (x-\mu)^2 f(x)dx}, \tag{2}$$

where $\mu = \int xf(x)dx$, and the integrals are the definite integrals taken over the range of X. If the variable X is discrete with probability function $f(x)$, the integral signs are replaced by summation signs.

In the case where X takes random values from a finite data set $x_1, x_2, \cdots, x_N$, the standard deviation is given by

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}, \tag{3}$$

where μ is the mean of X.

Estimation

For estimating the standard deviation from sample observations, μ in Eq. 3 is to be replaced by the sample mean $\bar{x}$ given by $\bar{x} = \sum_{i=1}^{n} x_i/n$, and then it is denoted by s_n. This s_n is the maximum likelihood estimate of σ when the population is normally distributed.

For estimating the standard deviation from a small sample, the sample standard deviation, denoted by s, can be computed by

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}, \tag{4}$$

where $\{x_1, x_2, \cdots, x_n\}$ is the sample, and $\bar{x}$ is the sample mean. This correction (use of $n-1$ instead of n), known as Bessel's correction, makes s^2 an unbiased estimator for the variance σ^2.

It can be shown that $\hat{\sigma}$ = IQR/1.35, where IQR is the interquartile range of the sample, is a consistent estimate of σ. The asymptotic relative efficiency of this estimator with respect to sample standard deviation is 0.37. It is, therefore, better to use sample standard deviation for normal data, while $\hat{\sigma}$ can be more efficient when the distribution of data is with thicker tail[3]. Standard deviation is independent of change of origin but not of scale.

Interpretation and Application

Standard deviation is the most useful and frequently used measure of dispersion. Standard deviation is used both as a separate entity and as a part of other analyses, such as computing confidence intervals and in hypotheses testing. Standard deviation is zero if all the elements of a population or data set are identical. It becomes larger if the data tend to spread over a larger range of values.

In science, researchers use standard deviation of experimental data for testing statistical significance. σ and $\hat{\sigma}$ are used in making certain tests of statistical significance. Standard deviation of a group of repeated measurements gives the precision of those measurements. In finance, it is used as a measure of risk on an investment. Standard deviation can be used to examine if a professional is consistent in his work. Similarly, standard deviation of scores (runs) made by a cricket player in a season tells about the consistency in his performance.

Standard deviation of an estimate, called the *Standard error*, is used to have an idea of the precision of that estimate.

►Chebyshev's inequality, (which enables to find probability without knowing probability function of a random variable), throws light on the connection between standard deviation and dispersion. For all distributions for which standard distribution is defined, it states that at least $\left(1-\frac{1}{k^2}\right)100\%$ of the values are within k standard deviation from the mean.

About the Author

Professor Khan is Editor of the *Journal of Statistical Research* (JSR), official publication of the Institute of Statistical Research and Training, University of Dhaka, Bangladesh.

Cross References

►Chebyshev's Inequality
►Coefficient of Variation
►Portfolio Theory
►Variance

References and Further Reading

Pearson Karl (1894) On the dissection of asymmetrical curves. Philos Tr R Soc S-A 185:719–810

Miller J. Earliest known uses of some of the words of mathematics. http://jeff560.tripod.com/mathword.html

Das Gupta A, Haff L (2006) Asymptotic expansions for correlations between measures of spread. J Stat Plan Infer 136: 2197–2213

Yule GU, Kendall MG (1958) An introduction to the theory of statistics, 14th edn. 3rd Impression. Charles Griffin & Company, London

Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences

David S. Jones
Professor of Biomaterial Science, Chair of Biomaterial Science
School of Pharmacy, Queens University of Belfast, Belfast, UK

Introduction

Essential to the efficacy of performance of drug delivery systems is the ability of the drug to diffuse from the said delivery systems and dissolve within the biological medium. Following this, the drug may diffuse through the biological media and subsequently diffuse across the attendant biological membranes, thereby gaining entry into the systemic circulation. In certain systems, the rate at which the drug dissolves within the biological fluid is the slowest and hence the rate-limiting step whereas in other scenarios the diffusion of the drug across the biological membrane may present the greatest challenge. In light of the importance of drug release, it is essential to ensure that the statistical analysis of the data from such experiments is successfully performed to enable rational conclusions to be drawn.

The conductance and design of drug release experiments is relatively straightforward and is defined within the scientific literature and within Pharmacopoeial monographs, e.g., the British Pharmacopoeia, the United States Pharmacopoeia. However, there is a relative paucity of information concerning methods that may be used to statistically quantify the outcomes of these experiments. Experimentally the analysis of drug release is typically performed by immersion of the dosage form within a defined volume of fluid designed to mimic a particular biological matrix, e.g., simulated gastric fluid, simulated intestinal fluid. The volume of fluid is chosen to ensure that the subsequent dissolution is typically not affected by the concentration of dissolved drug within the fluid. Thereafter, at defined time intervals, a sample of the surrounding fluid is removed and the mass of drug quantified using an appropriate analytical method, e.g., ultraviolet spectroscopy, fluorescence spectroscopy. After this analysis, there are two major challenges to the pharmaceutical scientist to ensure that the interpretation of the data is satisfactorily performed, namely:

(1) Selection of the appropriate mathematical model to define release.

(2) Use of statistical methods to examine formulation effects or release fluid effects on drug release.

The intention of this paper is to define appropriate statistical methods to address the above issues and thereby to define a protocol for the analysis of data that has been derived from drug release experiments.

Drug Release from Pharmaceutical Systems

Since the first publication of papers on the modelling of drug release for drug delivery systems (see Baker 1987, Chien 1992) there have been several papers that have applied mathematical concepts to understand the mechanism of drug release from such systems. For the purpose of this article, these methods may be summarised into three categories defined according to the mechanism of drug release, as follows:

(a) *Controlled (Fickian) release from monolithic devices*
In this method the release of a homogeneously dispersed drug from the delivery system is controlled by conventional diffusion (as initially described by Adolf Fick). Mathematically, Fickian diffusion of a drug from a slab geometry may be defined as follows:

$$\frac{M_t}{M_\infty} = 1 - \sum_{n=0}^{\infty} \frac{8\exp\left[-D(2n+1)^2\pi^2 t/l^2\right]}{(2n+1)^2\pi^2}. \qquad (1)$$

At early time approximations $\left(0 \le \frac{M_t}{M_\infty} \le 0.6\right)$, the following approximation may be made:

$$\frac{M_t}{M_\infty} = 4\left(\frac{Dt}{\pi l^2}\right)^{0.5}, \qquad (2)$$

where: D is the diffusion coefficient of the drug
t is time
l is the thickness of the slab geometry
M is the mass of drug released.
Accordingly it may be observed that the fraction of drug release is proportional to the square root of time.

(b) *Reservoir devices*
In these systems, drug diffusion from the device is controlled by the presence of a membrane. Mathematically, drug diffusion from the core of the device is defined by the following equations:

$$\frac{dM_t}{dt} = \frac{DAKC_s}{l} \quad \text{for a slab geometry} \qquad (3)$$

$$\frac{dM_t}{dt} = \frac{2\pi hDKC_s}{\ln\left(\frac{r_0}{r_1}\right)} \quad \text{for a cylinder geometry} \qquad (4)$$

$$\frac{dM_t}{dt} = \frac{4\pi hDKC_s r_0 r_1}{r_0 - r_1} \quad \text{for a sphere geometry} \qquad (5)$$

where: D is the diffusion coefficient
l is the thickness of the slab geometry
M_t is the mass of drug released at time t
h is the length of the cylinder
r_0 and r_1 are the outside and inside radii of the cylinder/sphere
A is the area of the device
K is the partition coefficient of the drug between the core and membrane

Under the above circumstances it may be observed that the mass of drug released is directly proportional to time.

More recently, Peppas (1985) described the use of a generic equation to model and characterise drug release from pharmaceutical platforms, as follows:

$$\frac{M_t}{M_\infty} = kt^n \tag{6}$$

where:
k is the release constant
$\frac{M_t}{M_\infty}$ is the fractional drug release
n is the release exponent.

In this approach, the equation encompasses the previous mathematical model, the value of the release exponent being used to define whether the mechanism of drug release from slab systems is:

(a) Fickian ($n = 0.5$)
(b) Reservoir controlled ($n = 1$)
(c) Anomalous ($0.5 < n < 1$)

Defining the Statistical Problem

Whilst the mathematical approaches described above seem quite straightforward, there is an ongoing issue with the application of these models within a statistical framework. There are several issues, which may be defined as follows:

(1) *Use of the incorrect mathematical model*

The choice of the correct mathematical model should be performed following consideration of the design of the dosage form and also the experimental conditions. In many situations, the limitations of the models are overlooked to render the mathematical analysis more straightforward. For example, in Fickian diffusion controlled systems, the mathematical model may only be used whenever there is no swelling of the pharmaceutical device. Furthermore, as highlighted in one of the examples above, the geometry of the device will affect the choice of equation. However, whilst the above concerns may seem obvious to those experienced in the pharmaceutical sciences, one common concern regards the modelling process. Typically the Peppas model is used to model release data however, in the early stages the model may yield an exponent of unity which may not be a true reflection of the release kinetics of the system as both diffusion controlled release and anomalous release will also yield similar exponents over this period of testing.

(2) *Choice of Statistical Tests*

Having acquired drug diffusion/dissolution data, the next challenge to the pharmaceutical scientist concerns the choice of the correct statistical method. One test that is recommended by the FDA is the f_2 test, which is used to compare the dissolution of two products, typically a test product (e.g., a generic product) and a reference product. The f_2 value is calculated using the following equation (Bolton and Bon 2004):

$$f_2 = 50 \log\left(\left[1 + \frac{1}{N}\right]\sum(R_t - T_t)^2 \times 100\right), \tag{7}$$

where: R_t and T_t are the % dissolution of the reference and test product at time t.

In this test an f_2 value >50 illustrates similarity of dissolution profiles. However, it should be noted that this test has several limitations; most notably individual differences at early time points may render the dissolution of two formulations different whenever the overall profiles are similar. The f_2 test has been principally used in the pharmaceutical industry to compare the dissolution of two dosage forms however; it is not commonly used within pharmaceutical research due to its relative inflexibility. The question may then be asked, "How are the drug release profiles of two, or more than two dosage forms compared?" Examples of the strategies that may be used are provided below.

(a) *Comparison of the release rates of the different formulations*

Mathematically the release of a drug from a dosage form is frequently described using the release rate, i.e., the slope of the plot of cumulative drug release against time^n. To use this method it must initially be *correctly* proven that the mechanisms of drug release from the different formulations are similar, a point often overlooked within the scientific literature. In light of the potential similarities of the kinetics of drug release for diffusion controlled, anomalous and zero order systems at early time points, it is essential to statistically establish similarity. Therefore, drug release should be allowed to progress to ensure that up to 60% release has occurred. To establish similarity of release mechanisms, it is appropriate to model drug release using the Peppas model and to then compare the release exponent values. For this purpose the Peppas model is transformed logarithmically, the release exponent (n) being the

resultant slope of the line following linear regression.

$$\ln \frac{M_t}{M_\infty} = \ln k + n \ln t. \quad (8)$$

The underlying prerequisite of this approach is the requirement for linearity. Typically linearity should be proven using both an ►Analysis of Variance and reference to Pearson's correlation coefficient (this should be greater than 0.99 [Jones 2002]). To facilitate meaningful statistical analysis of the data, it is suggested that approximately six replicate measurements should be performed as this increase the likelihood of the use of parametric tests for subsequent comparisons of the release exponents. Following the acquisition of this information the following points should be considered:

- To establish the release mechanism of the drugs from the pharmaceutical systems, the calculated release exponent should be statistically compared to 0.5 and also to 1.0. This is typically performed using a one sample *t* test. Retaining of the null hypothesis in these tests confirms that the release is either zero-order or diffusion controlled. Rejection of the null hypothesis verifies that the release mechanism is anomalous, i.e., $0.5 < n < 1.0$. The reader should note that the values of *n* representative of diffusion controlled and zero-order release are dependent on the geometry of the system. For a cylindrical system the release exponents are 0.45 and 0.89 for Fickian controlled and zero-order systems, respectively whereas for spherical systems these values become 0.43 and 0.85.
- Assuming that the release mechanism of all formulations under examination is similar, it is therefore appropriate to statistically compare the drug release kinetics from the various formulations. Therefore, for reservoir systems (in which the mechanism of release is zero-order), the plot of cumulative drug release against time is linear whereas in Fickian diffusion, the plot of cumulative drug release against $\sqrt{}$time is linear. Using linear regression analysis (and remembering not to include the point 0,0 in the analysis), the slope of the plot may be statistically determined for each individual replicate, which for diffusion controlled release and reservoir (zero-order) controlled release have the units of (concentration)(time)$^{-0.5}$ and (concentration)(time)$^{-1}$. Replication of these analyses (e.g., $n = 6$) enables calculation of the mean ± standard deviation or the median and ranges of the rates of release. Finally comparison of the rates of release may be easily performed using either the Analysis of Variance or the Kruskal-Wallis test if more than two samples/formulations require to be compared or, alternatively, the unpaired *t* test or the Mann Whitney *U* test, if the number of formulations under comparison is two. The choice of parametric or non-parametric tests to analyse the data is performed according to conventional statistical theory, the former tests being used if the populations from which the data were sampled were normally distributed (commonly tested using, e.g., the ►Kolmogorov-Smirnov test or the Shapiro-Wilk test) and if the variances of the populations from which the data were samples were statistically similar (commonly tested using e.g., Leveine's test or ►Bartlett's test). It should be noted that this approach is employed if the release mechanisms of different formulations are statistically similar, independent of the mechanism of drug release. Accordingly, the release exponent of different formulations may be identical within the range of $0.5 < n < 1.0$.

(b) *Comparing drug release from pharmaceutical systems that exhibit different release mechanisms*

In the above scenarios, the release rate of the drug from the pharmaceutical platform was obtained from linear regression of the associated cumulative drug release plot, i.e., cumulative drug release against time for the zero-order system and cumulative drug release against the square root of time for diffusion control systems. The above approach is predicated on the identical mechanisms of drug release; however, this requirement does raise a statistical dilemma. Consequently if the release mechanisms (and hence measured units) are different, therefore it is impossible to generate a single parameter that may be used as the basis for comparisons of the various formulations.

Under these conditions there are two approaches that may be employed to generate meaningful comparisons of drug release from different formulations.

(1) *Analysis of the data sets using a repeated measures Analysis of Variance*

This approach uses a repeated measures experimental design to compare drug release from different formulations. In this the repeated measure is time (which should be identical for each formulation) and the factor is formulation type. Individual differences between the various formulations may then be identified using an appropriate post hoc test. It is essential to ensure that the experimental design does not become overly complicated and that the demands of the ANOVA (with respect to homogeneity of population variances and the use of normally-distributed populations) hold.

(2) *Analysis of data at single time points*

The main requirements for the use of the repeated measures Analysis of Variance are, firstly that the requirements for the use of this test are met and secondly, that the times at which the data were collected (sampled) are identical for each formulation. In practice these problems are straightforward to overcome at the experimental design stage however, there may be issues concerning the ability to perform the required number of replicates (typically ≥5) to allow a parametric test is suitable to use for the data analysis. For example, experiments in which the release is relatively rapid (<48 h) may be easier to perform with many replicates whereas the converse is true for experiments in which the release is protracted. In such circumstances (e.g., whenever there are few replicates, typically $n \leq 3$), one method that may be employed to compare the drug release profiles of different formulation involves the comparison of the formulations at each sampling point using a multiple hypothesis test, e.g., the Kruskal-Wallis test. In a similar fashion, individual differences between formulations may be identified by the application of an appropriate post hoc test, e.g., Dunn's test, Nemenyi's test.

In an alternative approach, typically encountered whenever the sampling periods differ, comparison of the drug release kinetics of candidate formulations may be performed by ascertaining the time required for a defined fraction of the initial drug loading to be released. A regression of the release profile (using the Peppas model) is performed and, using the output from this model, the times required for each formulation to release a defined fraction is obtained and statistically compared using the appropriate statistical test (Jones et al. 1999; Jones et al. 2000). The choice if test to perform the analysis is important and the reader should be reminded that the use of parametric statistical tests (the unpaired *t* test and the ANOVA) should be validated.

Conclusions

Analysing release data is an essential component in the development and assessment of the performance of pharmaceutical systems. In spite of this, suitable methods to analyse release data are not clearly defined. In this monograph strategies for the statistical comparisons of release data are defined.

About the Author

David Jones is Professor of Biomaterial Science at the Queen's University of Belfast. Professor Jones is a Chartered Engineer, Chartered Chemist and holds Fellowships of the Royal Statistical Society and the Institute of Materials, Minerals and Mining and is a Member of the Royal Society of Chemistry, the Institute of Engineers in Ireland and the Pharmaceutical Society of Northern Ireland. He is the Editor of the *Journal of Pharmacy and Pharmacology* and has been the Statistical Advisor to the International Journal of Pharmacy Practice. Professor Jones is a former winner of the Eli Lilly Award and the British Pharmaceutical Conference Science Medal.

Cross References

►Analysis of Variance
►Biopharmaceutical Research, Statistics in
►Medical Research, Statistics in
►Parametric Versus Nonparametric Tests
►Pharmaceutical Statistics: Bioequivalence
►Repeated Measures
►Student's t-Tests
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Baker RW (1987) Controlled release of biologically active agents. Wiley-Interscience, New York

Bolton S, Bon C (2004) Pharmaceutical statistics: practical and clinical applications, vol 135. Marcel Dekker, New York, p 755

Chien YW (1992) Novel drug delivery systems, 2nd edn. vol 50. Marcel Dekker, New York

Jones DS (2002) Pharmaceutical statistics. Pharmaceutical Press, London, p 608

Jones DS, Irwin CR, Woolfson AD, Djokic J, Adams V (1999) Physicochemical characterization and preliminary in vivo efficacy of bioadhesive, semisolid formulations containing flurbiprofen for the treatment of gingivitis. J Pharm Sci 88(6):592–598

Jones DS, Woolfson AD, Brown AF, Coulter WA, McClelland C, Irwin CR (2000) Design, characterisation and preliminary clinical evaluation of a novel mucoadhesive topical formulation containing tetracycline for the treatment of periodontal disease. J Cont Rel 67(2–3):357–368

Peppas NA (1985) Analysis of Fickian and Non-Fickian drug release from polymers. Pharm Acta Helvet 60(4):110–111

Statistical Analysis of Longitudinal and Correlated Data

David Todem
Michigan State University, East Lansing, MI, USA

Introduction

Correlated data are typically generated from studies where the outcomes under investigation are collected on clustered units. Specific examples include; (1) longitudinal data where outcomes are collected on the same experimental unit (for instance, the same person) at two or more different points in time; and (2) studies where outcomes

are recorded at one single point in time on clustered units. Such studies have one major attraction, the ability to control for unobserved variables in making inferences. Sampled units serve as controls for other units in the same cluster. As an example, in a longitudinal study, each subject serves as his or her own control in the study of change across time. Therefore, these studies allow the researcher to eliminate a number of competing explanations for observed effects. The determination of causal ordering in making solid inferences contitutes another attraction for longitudinal studies.

Despite these advantages, statistical analysis of correlated data raises a number of challenging issues. It is well known, for example, that the multiplicity of outcomes recorded over time on the same unit necessitates the use of methods for correlated data. This entry reviews some of the common statistical techniques to analyze such data. A focus is on longitudinal data as statistical models for clustered data are typically simple versions of techniques for longitudinal data. In longitudinal data analysis, the response $y(t)$ is a time-varying variable and the covariate can be a baseline vector x, a time-varying covariate vector $x(t)$, or a combination of both. A key issue for such data is to relate the longitudinal mean responses to covariates and draw related inferences while accounting for the within-subject association. In essence, two classes of models exist for modeling the mean outcomes and covariates relationship; (1) the parametric models and; (2) the semiparametric and nonparametric models. This entry examines each of these models in some detail, with an eye to discerning their relative advantages and disadvantages. A discussion on emerging issues in analyzing longitudinal data is also given but touched on briefly.

Parametric Models

Parametric models are the predominant approaches for longitudinal data. They make parametric assumptions about the relationship between the mean of a longitudinal response to covariates. They are known as growth curve models and include the popular mixed-effects models (Laird and Ware 1982) and generalized estimating equations models (Liang and Zeger 1986). Verbeke and Molenberghs (2000) and Diggle et al. (2002) provide an extensive review of this literature.

Mixed-Effects Models

Mixed-effects models are a useful tool to analyze repeated measurements recorded on the same subject. They were primarily developed for continuous outcomes in time (Laird and Ware 1982) and were later extended to categorical and discrete data (Breslow and Clayton 1993). For continuous outcomes with an identity link, they are known as linear mixed-effects models. Generalized linear mixed-effects models constitute the broader class of mixed-effects models for correlated continuous, binary, multinomial, ordinal and count data (Breslow and Clayton 1993). They are likelihood-based and often are formulated as hierarchical models. At the first stage, a conditional distribution of the responses given random effects is specified, usually assumed to be a member of the exponential family. At the second stage, a prior distribution is imposed on the random effects. The conditional expectations (given random effects) are made of two components, a fixed-effects and a random-effects term. The fixed-effects term represents covariate effects that do not change with the subject. Random effects represent a deviation of a subject's profile from the average profile. Most importantly, they account for the within-subject correlation across time under the conditional independence assumption. For continuous outcomes with an identity link function, these models have an appealing feature in that the fixed-effects parameters have a subject-specific as well as a population-averaged interpretation (Verbeke and Molenberghs 2000). For non continuous data and nonlinear relationships, this elegant property is lost. The fixed-effects parameters, with the exception of few link functions, only have a subject-specific interpretation, conditional on random effects. This interpretation is only meaningful for covariates that change within a subject such as time-varying covariates. These effects capture the change occurring within an individual profile. To assess changes for time-independent covariates, the modeler is then required to integrate out the random effects from the quantities of interest.

Mixed-effects models are likelihood-based and therefore can be highly sensitive to any distribution misspecification. But they are known to be robust against less restrictive missing data mechanisms. There exist other likelihood-based methods for analyzing correlated data. Before the advent of ▶linear mixed models, longitudinal continuous data were analyzed using techniques such as repeated measures analysis of variance (ANOVA). This approach has a number of disadvantages and has generally been superseded by linear mixed-effects models, which can easily be fit in mainstream statistical software. For example, repeated measures ANOVA models require a balanced design in that measurements should be recorded at the same time points for all subjects, a condition not required by linear mixed models.

Generalized Estimating Equations Models

Although there is a variety of standard likelihood-based models available to analyze data when the outcome is approximately normal, models for discrete outcomes (such

as binary outcomes) generally require a different methodology. Liang and Zeger (1986) have proposed the so-called Generalized Estimating Equations-GEE model, which is an extension of ▶generalized linear models to correlated data. The basic idea of this family of models is to specify a function that links the linear predictor to the mean response, and use a set of estimating functions with any working correlation model for parameter estimation. A sandwich estimator that corrects for any misspecification of the working correlation model is then used to compute the parameters' standard errors. GEE-based models are very popular as an all-round technique to analyze correlated data when the exact likelihood is difficult to specify. One of the strong points of this methodology is that the full joint distribution of the data does not need to be specified to guarantee asymptotically consistent and normal parameter estimates. Instead, a working correlation model between the clustered observations is required for estimation. GEE regression parameter estimates have a population-averaged interpretation, analogous to those obtained from a cross-sectional data analysis. This property makes GEE-based models desirable in population-based studies, where the focus is on average affects accounting for the within-subject association viewed as a nuisance term.

The GEE approach has several advantages over a likelihood-based model. It is computationally tractable in applications where the parametric approaches are computationally very demanding, if not impossible. It is also less sensitive to distribution misspecification as compared to full likelihood-based models. A major limitation of GEE-based models at least in their 1986 original formulation is that they require a more stringent missing data mechanism (missing data completely at random) to produce valid inferences. Weighted GEE-based models have been proposed to accommodate a less stringent missing data mechanism, the missing data at random process (Robins et al. 1995).

Semiparametric and Nonparametric Models

A major limitation of parametric models is that the relationship of the mean of a longitudinal response to covariates is assumed fully parametric. Although such parametric mean models enjoy simplicity and ease of interpretation, they often suffered from inflexibility in modeling complicated relationships between the response and covariates in various longitudinal studies. Specific examples include modeling of; (1) longitudinal *CD4+* counts as function of time in HIV/AIDS research; and (2) trajectories of angiogenic and antiogenic factors in maternal plasma concentrations (s-eng, sVEGFR-1 and PlGF) in perinatal research. Parametric models typically require higher degree polynomials to capture the relationship between these mean responses and covariates. This has been seen as an indication of poor fit and has motivated the development of more complex and flexible approaches to model these data. Semiparametric and nonparametric regression models, well known to be more data adaptive, have emerged as promising alternative to parametric models in these settings. Nonparametric models make no parametric assumption about the relationship between the mean response and covariates. Semiparametric models assume a parametric relationship between some covariates and the mean response while maintaining a nonparametric relationship between other covariates and the mean response. These methods are well developed for independent data, but their extensions to longitudinal data remain an active area of research. A major difficulty often cited in the literature for this extension is the inherent within-subject correlation in longitudinal studies. This correlation presents significant challenges in the development of kernel and spline smoothing methods for longitudinal data. Specifically, as reported by many researchers in the field (see for example, Lin and Carroll 2000; Lin et al. 2004), local likelihood-based kernel methods are not able to effectively account for the within-subject correlation in longitudinal data.

Discussion

This entry has reviewed some of the common techniques to model longitudinal data. A focus was on parametric models. Nonparametric and semiparametric approaches based on smoothing techniques have emerged as a flexible way to model longitudinal data. Other approaches that do not require smoothing have recently been proposed (Lin and Ying 2001). But much research, especially from a theoretical standpoint, is needed to understand these methods. Moreover, statistical software to fit these models routinely in real time is much needed. This is in contrast to parametric models which can be fit using mainstream statistical software such as SAS, Stata, R, Splus and SPSS. There are emerging areas in connection to longitudinal data analysis that need further research such as; (1) the joint modeling of longitudinal and ▶survival data, (2) missing data and (3) causal inference. These areas have enjoyed some significant developments in the past several years. But there are numerous open questions that remain unanswered and are the subject of future research.

About the Author

Dr. David Todem is a Biostatistics Associate Professor in the Division of Biostatistics of the Department of Epidemiology at Michigan State University, USA. He has authored

and co-authored more than 30 papers and 2 entries in encyclopedic publications. Dr Todem is an Editorial Board member for *The Open Statistics and Probability Journal*.

Cross References

►Data Analysis
►Exponential Family Models
►Linear Mixed Models
►Medical Statistics
►Multilevel Analysis
►Nonlinear Mixed Effects Models
►Nonparametric Regression Using Kernel and Spline Methods
►Panel Data
►Random Coefficient Models
►Repeated Measures
►Semiparametric Regression Models
►Statistical Software: An Overview

References and Further Reading

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25
Diggle PJ, Heagerty PJ, Liang K-Y, Zeger S (2002) Analysis of longitudinal data. Oxford University Press, Oxford
Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38:963–974
Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22
Lin D, Ying Z (2001) Semiparametric and nonparametric regression analysis of longitudinal data. J Am Stat Assoc 96:103–126
Lin X, Carroll RJ (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error. J Am Stat Assoc 95:520–534
Lin X, Wang N, Welsh A, Carroll RJ (2004) Equivalent kernels of smoothing splines in nonparametric regression for clustered data. Biometrika 91:177–193
Robins J, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes under the presence of missing data. J Am Stat Assoc 90:106–121
Verbeke G, Molenberghs G (2000) Linear mixed models for longitudinal data. Springer, New York

Statistical Approaches to Protecting Confidentiality in Public Use Data

Jerome P. Reiter
Associate Professor
Duke University, Durham, NC, USA

Many national statistical agencies, survey organizations, and researchers – henceforth all called agencies – collect data that they intend to share with others. Wide dissemination of data facilitates advances in science and public policy, enables students to develop skills at data analysis, and helps ordinary citizens learn about their communities. Often, however, agencies cannot release data as collected, because doing so could reveal data subjects' identities or values of sensitive attributes. Failure to protect confidentiality can have serious consequences for agencies, since they may be violating laws or institutional rules enacted to protect confidentiality. Additionally, when confidentiality is compromised, the agencies may lose the trust of the public, so that potential respondents are less willing to give accurate answers, or even to participate, in future studies (Reiter 2004).

At first glance, sharing safe data with others seems a straightforward task: simply strip unique identifiers like names, tax identification numbers, and exact addresses before releasing data. However, these actions alone may not suffice when quasi-identifiers, such as demographic variables, employment/education histories, or establishment sizes, remain on the file. These quasi-identifiers can be used to match units in the released data to other databases. For example, Sweeney (1997) showed that 97% of the records in a medical database for Cambridge, MA, could be identified using only birth date and nine-digit ZIP code by linking them to a publicly available voter registration list.

Agencies therefore further limit what they release, typically by altering the collected data (Willenborg and de Waal 2001). Common strategies include those listed below. Most public use data sets released by national statistical agencies have undergone at least one of these methods of statistical disclosure limitation.

Aggregation. Aggregation reduces disclosure risks by turning atypical records – which generally are most at risk – into typical records. For example, there may be only one person with a particular combination of demographic characteristics in a city, but many people with those characteristics in a state. Releasing data for this person with geography at the city level might have a high disclosure risk, whereas releasing the data at the state level might not. Unfortunately, aggregation makes analysis at finer levels difficult and often impossible, and it creates problems of ecological inferences.

Top coding. Agencies can report sensitive values exactly only when they are above or below certain thresholds, for example reporting all incomes above $200,000 as "$200,000 or more." Monetary variables and ages are frequently reported with top codes, and sometimes with bottom codes as well. Top or bottom coding by definition eliminates detailed inferences about the distribution

beyond the thresholds. Chopping off tails also negatively impacts estimation of whole-data quantities.

Suppression. Agencies can delete sensitive values from the released data. They might suppress entire variables or just at-risk data values. Suppression of particular data values generally creates data that are not missing at random, which are difficult to analyze properly.

Data swapping. Agencies can swap data values for selected records – for example, switch values of age, race, and sex for at-risk records with those for other records – to discourage users from matching, since matches may be based on incorrect data (Dalenius and Reiss 1982). Swapping is used extensively by government agencies. It is generally presumed that swapping fractions are low – agencies do not reveal the rates to the public – because swapping at high levels destroys relationships involving the swapped and unswapped variables.

Adding random noise. Agencies can protect numerical data by adding some randomly selected amount to the observed values, for example a random draw from a normal distribution with mean equal to zero (Fuller 1993). This can reduce the possibilities of accurate matching on the perturbed data and distort the values of sensitive variables. The degree of confidentiality protection depends on the nature of the noise distribution; for example, using a large variance provides greater protection. However, adding noise with large variance introduces measurement error that stretches marginal distributions and attenuates regression coefficients (Yancey et al. 2002).

Synthetic data. The basic idea of synthetic data is to replace original data values at high risk of disclosure with values simulated from probability distributions (Rubin 1993). These distributions are specified to reproduce as many of the relationships in the original data as possible. Synthetic data approaches come in two flavors: partial and full synthesis (Reiter and Raghunathan 2007). Partially synthetic data comprise the units originally surveyed with some subset of collected values replaced with simulated values. For example, the agency might simulate sensitive or identifying variables for units in the sample with rare combinations of demographic characteristics; or, the agency might replace all data for selected sensitive variables. Fully synthetic data comprise an entirely simulated data set; the originally sampled units are not on the file. In both types, the agency generates and releases multiple versions of the data (as in multiple imputation for missing data, see ►Multiple Imputation). Synthetic data can provide valid inferences for analyses that are in accord with the synthesis models, but they may not give good results for other analyses.

Statisticians play an important role in determining agencies' data sharing strategies. First, they measure the risks of disclosures of confidential information in the data, both before and after application of data protection methods. Assessing disclosure risks is a challenging task involving modeling of data snoopers' behavior and resources; see Reiter (2005) and Elamir and Skinner (2006) for examples. Second, they advise agencies on which protection methods to apply and with what level of intensity. Generally, increasing the amount of data alteration decreases the risks of disclosures; but, it also decreases the accuracy of inferences obtained from the released data, since these methods distort relationships among the variables. Statisticians quantify the disclosure risks and data quality of competing protection methods to select ones with acceptable properties. Third, they develop new approaches to sharing confidential data (*see* ►Data Privacy and Confidentiality). Currently, for example, there do not exist statistical approaches for safe and useful sharing of network and relational data, remote sensing data, and genomic data. As complex new data types become readily available, there will be an increased need for statisticians to develop new protection methods that facilitate data sharing.

About the Author

Jerry Reiter is currently an Associate Editor for the *Journal of the American Statistical Association*, *Survey Methodology*, the *Journal of Privacy and Confidentiality*, and the *Journal of Statistical Theory and Practice*. He is the current Chair of the Committee on Privacy and Confidentiality of the American Statistical Association (2009–2012). He is an Elected member of the International Statistical Institute. He has authored more than 60 papers, including foundational works on the use of multiple imputation for confidentiality protection. He was awarded the Alumni Distinguished Undergraduate Teaching Award at Duke University.

Cross References

►Census
►Data Analysis
►Data Privacy and Confidentiality
►Federal Statistics in the United States, Some Challenges
►Multi-Party Inference and Uncongeniality

References and Further Reading

Dalenius T, Reiss SP (1982) Data-swapping: a technique for disclosure control. J Stat Plan Infer 6:73–85

Elamir E, Skinner CJ (2006) Record level measures of disclosure risk for survey microdata. J Off Stat 22:525–539

Fuller WA (1993) Masking procedures for microdata disclosure limitation. J Off Stat 9:383–406

Reiter JP (2004) New approaches to data dissemintation: a glimpse into the future (?). Chance 17(3):12–16

Reiter JP (2005) Estimating identification risks in microdata. J Am Stat Assoc 100:1103–1113
Reiter JP, Raghunathan TE (2007) The multiple adaptations of multiple imputation. J Am Stat Assoc 102:1462–1471
Rubin DB (1993) Discussion: statistical disclosure limitation. J Off Stat 9:462–468
Sweeney L (1997) Computational disclosure control for medical microdata: the Datafly system. In: Proceedings of an international workshop and exposition, pp 442–453
Willenborg L, de Waal T (2001) Elements of statistical disclosure control. Springer, New York
Yancey WE, Winkler WE, Creecy RH (2002) Disclosure risk assessment in perturbative microdata protection. In: Domingo-Ferrer J (ed) Inference control in statistical databases. Springer, Berlin, pp 135–152

Statistical Aspects of Hurricane Modeling and Forecasting

Mark E. Johnson[1], Charles C. Watson[2]
[1]Professor
University of Central Florida, Orlando, FL, USA
[2]Watson Technical Consulting, Savannah, GA, USA

Hurricanes are complex, natural phenomena that can cause property damage on a catastrophic scale. The human toll depends on the preparedness of the population – historical events with thousands of casualties are rare but do occur (e.g., the 1900 Galveston storm – Larson 1999). Depending on where hurricanes form and traverse, they have other names such as typhoons (western Pacific) and cyclones (Indian Ocean and Australia). Officially, a hurricane is defined as a closed circulation, warm core, and convective weather system with maximum 10-min average winds of 33 m/s or higher, measured at 10 m above ground level (WMO 2007). This precise and technical definition is important since insurance payouts for losses often depend on the declaration of a hurricane event. The definition also provides a threshold for establishing the event frequency at specific locations, a criterion especially important for climate change studies. For planning purposes, the return period of hurricanes of various intensities is needed – i.e., what is the probability that 100 mph winds will strike a specific location this season or what wind speed corresponds to the 100 year worst event? Fortunately, hurricanes are relatively rare events (as compared to thunderstorms or tornadoes) and thus, extreme value methods are used to assess their frequencies (Embrechts et al. 1997). An excellent introduction to hurricanes is given by Emanuel (2005) while a more technical treatise is available by Anthes (1982).

Iman et al. (2006) reviewed many aspects of statistical forecasting and planning in the premier Interdisciplinary Section of The American Statistician. The invitation to prepare this article was motivated in part by the hyperactive 2004 and 2005 Atlantic hurricane seasons which stunned the American public following relatively minor hurricane activity in the United States since Hurricane Andrew in 1992. Various researchers took these two seasons as the onset of sustained, increased activity, only to witness the four subsequent years of little hurricane activity impacting Florida (O'Hagan et al. 2008). This perspective illustrates a United States-centric perspective regarding hurricane activity. The 2007 season endured two very strong events (Hurricanes Dean and Erin) which pummeled the Mexican Yucatan and the Gulf of Campeche, causing massive havoc with their oil and gas industry. Similarly, in 2009, the Philippines experienced multiple typhoons left nearly 1,000 dead, thousands homeless, and widespread agricultural devastation, yet received little media attention.

Forecasting hurricane track and intensity are key problems that must be addressed in real time for actual events under a harsh public and media spotlight as hurricane watches and warnings go into effect. The "obvious" forecast is to extrapolate the current track with a linear trend in intensity. A more sophisticated version of this forecast is to draw upon the historical record to develop a regression model using comparable information on the movement of storms getting to the current position of the storm (CLIPER and CLIPER5 in use by the National Hurricane Center). More advanced models take into account current and forecast upper level winds ("steering currents"), while the most advanced include fluid dynamics calculations of mesoscale storm structure. In addition to the many individual forecast models, ensemble models are also in use (for a technical summary, see www.nhc.noaa.gov/modelsummary.shtml). The increase in skill (accuracy of prediction) of the more sophisticated models is offset by data input needs and computational run times. Forecasts must be timely – a 6 h forecast that takes 5 h to produce may be inferior to a much simpler forecast that can be formulated in a matter of minutes. For a further discussion of the many pitfalls associated with forecasts, especially the problems encountered with Hurricane Charley in 2004, see the aforementioned article by Iman et al. (2006).

In determining hurricane impacts for insurance purposes, a more leisurely time frame for computation is available. The computational burden is severe in that a probabilistic assessment of hurricane losses is necessary.

Most approaches have proceeded by choosing specific, individual models of hurricane frequency, wind field, track, friction impacts, wind field decay, damage, and actuarial summaries. Given the approximately 150 year Atlantic storm history, less in other regions, practitioners have tended to fit probability distributions to key characteristics and then proceed to simulate 50–300,000 years of future hurricane seasons, accumulating losses for each generated event. To assess the uncertainty and sensitivity of the parameter specifications for these models, the Florida Commission on Hurricane Loss Methodology has prescribed the use of Latin hypercube sampling (McKay et al. 1979). One specific implementation pertinent to hurricane modeling is described by Iman et al. (2005a, b). The latest research focuses on the use of climate models to provide track and intensity guidance (Watson and Johnson 2008).

A basic issue with evaluating hurricane modeling efforts is that every hurricane is somewhat different and any model that "fine tunes" its modeling approach to a specific event will ultimately suffer for it (not all future events are just like the particular event. For some historical events, a very simple hurricane windfield model can do extremely well with respect to matching modeled to actual losses. An approach used by the Florida Commission to address this difficulty follows the contextual analysis developed by Watson and Johnson (2004) and expounded from an actuarial perspective by Watson, Johnson and Simons. In brief, a factorial combination of model components are considered (nine wind fields, four friction models, nine damage functions and three frequency approaches) and the loss costs for specific models are placed in the context of 972 model combination results. ►Outliers with respect to the range of the factorial models generate relevant probing questions of specific models.

Nelder (2010) noted the importance of learning another jargon for statisticians doing interdisciplinary research. The effort is well-rewarded for statisticians dealing with the topic of hurricanes which will likely entail collaborations with meteorologists, atmospheric scientists, geophysicists, and wind engineers.

About the Authors

Dr. Mark E. Johnson is professor, Department of Statistics, University of Central Florida. He was Department Chair (1990–1996). Dr Johnson is a Fellow of the American Statistical Association (1988), Elected Member, International Statistical Institutue (1994), Chartered Statistician, Royal Statistical Society (1993). He has (co-)authored more than 60 refereed papers and is author of *Multivariate Statistical Simulation* (Wiley 1987). Professor Johnson was awarded the Jack Youden Prize (1981), ASQC Shewell Award (1985), Thomas L. Saaty Prize (1984, 1989 and 1997), and ASQC Brumbaugh Award (1991), He was Associate Editor of *Technometrics* (1984–1991), *Journal of Quality Technology*, *American Journal of Management and Mathematical Sciences*, and *Journal of Statistical Computation and Simulation*.

Mr. Charles C. Watson Jr. is the founder and Director of Research and Development of Kinetic Analysis Corporation. He has authored or co-authored more than 70 papers and book contributions in the fields of satellite remote sensing, geophysics, and meteorology, in such diverse publications such as *Bulletin of the American Meteorological Society*, *Photogrammetric Engineering & Remote Sensing*, the *Journal of Insurance Regulation*, and *The American Statistician*. Mr. Watson has served or is active as a scientific consultant on hazard planning and remote sensing to numerous national and international projects and agencies such as the Caribbean Catastrophe Risk Insurance Facility, the Intergovernmental Panel on Climate Change, UN Agencies such as the World Meteorological Organization, UN Environment Program, and World Food Program, as well as US agencies such as National Aeronautics and Space Administration.

Cross References

►Actuarial Methods
►Forecasting: An Overview
►Statistics and Climate Change
►Statistics of Extremes
►Stochastic Difference Equations and Applications
►Time Series

References and Further Reading

Anthes RA (1982) Tropical cyclones, their evolution, structure, and effects, American Meteorological Society meteorological monographs, vol 19(41). AMS, Boston

Emanuel K (2005) Divine wind: the history of science and hurricanes. Oxford University Press, New York

Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events. Springer, Berlin

Iman RL, Johnson ME, Watson C Jr (2005a) Sensitivity analysis for computer model projections of hurricane losses. Risk Anal 25:1277–1297

Iman RL, Johnson ME, Watson C Jr (2005b) Uncertainty analysis for computer model projections of hurricane losses. Risk Anal 25:1299–1312

Iman RL, Johnson ME, Watson C Jr (2006) Statistical aspects of forecasting and planning for hurricanes. Am Stat 60(2):105–121

Larson E (1999) Isaac's Storm. A man, a time, and the deadliest hurricane in history. Crown Publishers, New York

McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245

Nelder J (2010) "Statistics: Nelder's view," International encyclopedia of statistical science. Springer, New York
O'Hagan T, Ward B, Coughlin K (2008) How many Katrinas? Predicting the number of hurricanes striking the USA. Significance 5(4):163–167
Watson C Jr, Johnson ME (2008) Integrating hurricane loss models with climate models. In: Murnane R, Diaz H (eds) Climate extremes and society. Cambridge University Press, Cambridge, pp 209–224
Watson C Jr, Johnson ME (2004) Hurricane loss estimation models: opportunities for improving the state of the art. B Am Meteorol Soc 84:1713–1726
Watson C Jr, Johnson ME, Simons M (2004) Insurance rate filings and hurricane loss estimation models. J Insur Regul 22(3):39–64
World Meteorological Organization (WMO) (2007) Global guide to tropical cyclone forecasting, WMO/TC-No. 560, Report No. TCO-31, World Meteorological Organization, Geneva, Switzerland

Statistical Consulting

Rolf Sundberg
Professor of Mathematical Statistics
Stockholm University, Stockholm, Sweden

What Is Statistical Consulting?

Here is a sketch of a normal consultation in the consulting unit of my department, in a faculty of sciences. One or a couple of researchers/Ph.D-students from a biology/geology/... department contact us asking for help with the analysis of data from a study they are carrying out. At the meeting the client first describes the background, the set-up, and (some of) the data of the study. The aims of the study are often in a general, vague form that needs specification and statistical reformulation in quantifiable units. What is the client's problem, really, and what kind of questions can possibly be answered from that kind of data? Often the clients will be forced to think about their problems in fresh ways. The consultant will also ask a lot of questions in order to make clear how the data were collected. What populations do the data represent? Was there ▶randomization, stratification, censoring, etc? On what parts of the data should the focus be? Explore the data! What is the structure of these data? This can lead up to a tentative statistical model, and later to parameter estimation procedures and hypothesis tests, etc.

The first meeting hopefully ends at a stage where the client and the statistician have agreed about what questions should be addressed statistically, and how this might be attempted on the data. Either this appears so simple and clear that the clients want and can do this themselves, or else a time plan and a work plan for the contribution by the statistician is agreed on. After a week or two, with some e-mail correspondence in between, client and consultant meet again to discuss the results so far and what kind of report from the statistician that the clients might want. Often also the answer to one question triggers new questions.

Another statistical consultation type of work could be more of a collaborative/partnership character, where the statistician is a member of a team, and the aims are more far-reaching. The statistician then invests a lot of time and effort, to become knowledgeable in the subject matter area and expert in the applications of statistical methods in that area, but can therefore also expect more influence and credit, and is a natural coauthor of the project publications.

Also a consultation where the client is seen only once or twice is rewarding for the statistician, but in a more indirect way. Hopefully it will be an intellectually stimulating challenge that together with other such experiences can have a profound influence on our personal development as statisticians. And it might still lead to a joint publication.

Consultation work is typically done under time pressure from one or both parties. Too often the client has unrealistic expectations in this respect. On the other hand, the clients usually do not need or want a perfect model for data (remember the George Box phrase: "All models are wrong, but some are useful") or the most sophisticated method of analysis. A solution that is approximately right is much better than one that is precisely wrong. The consultant should think of the acronym KISS, here read out as "Keep It Simple, but Scientific," or rephrased as another quotation: "as simple as possible, but no simpler." "Errors of the third kind" (testing the wrong hypothesis) are most dangerous, Common sense and a critical mind are important. As statistical consultants we must beware of falling in the traps of being a More Data Yeller or a Nit Picker, or any other of the consultant stereotypes coined by Hyams (1971).

Desirable Qualities for a Statistical Consultant

Among the desirable qualities to be possessed by an ideal consultant are:

- Interest in the statistical problems of others (Derr: "Regard each client as a potential collaborator"), and a general interest in science, technology, nature, society.

- Sound basis in theoretical and applied statistics. As a start it should certainly include linear and loglinear models (►generalized linear models), some experimental design (and sampling), and some multivariate analysis, but also experience from a few courses in methods for particular fields of application, and experience from applying such methods to data.
- Eagerness to extend and improve one's statistical knowledge.
- Computer skills in at least one (preferably more) statistical packages.
- Good ability to communicate with clients (includes understanding and adjusting to the client's statistical level).
- Skills in report writing (using a word processor).
- Efficiency under time restrictions and time pressure.
- Awareness of ethical dilemmas that can appear, and an ability to deal with problematic clients.

Teaching Statistical Consulting

Nowadays a large number of universities provide education in statistical consulting, in one form or the other. At my department, as an example, this is a master level course for mathematical statistics students, involving real clients, and real problems in real time. Much of the training in the course is orientated towards three aspects:

- The first meeting with a client (in particular asking questions to find out about the problem)
- Statistical thinking
- Structuring problems and seeing the structure in data

The students are also provided some extended knowledge of statistical methods and models, and they are in a concrete way involved in one consulting project, ending with the writing of a project report.

Some Suggested Reading

The entry by Stinnet et al. (2009) in Encyclopedia of Biostatistics describes the roles of biostatisticians in a variety of medical/biological environments (medical school, pharmaceutical industry, governmental agency, etc.), and discusses some of the special challenges in consulting with physicians, as well as the training of consultants in biostatistics. Joiner's (1982) older entry in Encyclopedia of Statistical Sciences also exemplifies what consulting statisticians might do, before it sets up and discusses a list of desirable skills. The discussion of computers and literature is a bit out-of-date, for natural reasons.

Mallows (1998) discusses "statistical thinking" and the question "how do the data relate to the problem?", in an attempt to formulate a "theory of applied statistics." Cox (2007) provides a review of applied statistics in his typical style, while Chatfield's (1995) nicely written book provides more concrete advice.

Efficient communication is a key element in statistical consultation, and it is the topic of Derr's (2000) book, with an accompanying CD-ROM showing illustrative short movies of positive and negative examples. Communication is the main topic also of Boen and Zahn (1982), who provide much discussion of how to deal with clients, not least with difficult clients, cf. Hyams (1971).

Cabrera and McDougall (2002) is written as a textbook on the whole topic. The first half is on consulting, communication, and statistical methods. I do not agree fully with the statistical methods chapter, but who would expect two statisticians to agree fully? The second half consists of case studies. Such a mix also characterizes Chatfield's (1995) book, and the older book by Cox and Snell (1981), that can be recommended in this context for a section on strategy and for its many case studies. More case studies are found in Hand and Everitt (1987) and in Tweedie et al. (1998). Greenfield's contribution to the former is an entertaining chapter on the encounters he has had with some difficult client characters (cf. Hyams 1971, again).

To finish, here is a quote from one of Terry Speed's columns in the IMS Bulletin (2005), entitled "How to do Statistical Research." Former IMS President Speed explains his research strategy to be that of doing

- *Consulting*: a very large amount
- *Collaboration*: quite a bit
- *Research*: some

"Why? A very large amount of consulting means meeting many people and many problems, learning a lot, including finding out where we are ignorant. Then we might spot some low-hanging fruit."

About the Author

For biography *see* the entry ►Chemometrics.

Cross References

►Careers in Statistics
►Data Analysis
►Generalized Linear Models
►Model Selection
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Research Designs
►Sample Survey Methods
►Statistical Design of Experiments (DOE)

►Statistical Literacy, Reasoning, and Thinking
►Statistical Software: An Overview
►Statistics: Nelder's view

References and Further Reading

ASA Section on Statistical Consultation (2003) When you consult a statistician ... what to expect? Downloadable from www.amstat.org/sections/cnsl/brochures/SCSbrochure.pdf
Boen JR, Zahn DA (1982) The human side of statistical consulting. Wadsworth, Belmont, CA
Cabrera J, McDougall A (2002) Statistical consulting. Springer, New York
Chatfield C (1995) Problem solving. A statistician's guide, 2nd edn. Chapman & Hall, London
Cox DR (2007) Applied statistics: a review. Ann Appl Stat 1:1–16
Cox DR, Snell EJ (1981) Applied statistics. Principles and examples. Chapman & Hall, London
Derr J (2000) Statistical consulting. A guide to effective communication. Duxbury Press, Pacific Grove, CA
Hand DJ, Everitt BS (1987) The statistical consultant in action. Cambridge University Press, Cambridge
Hyams L (1971) The practical psychology of biostatistical consultation. Biometrics 27:201–211
Joiner BL (1982) Consulting, statistical. In: Encyclopaedia of statistical sciences. Wiley, New York, pp 147–155
Mallows C (1998) The zeroth problem. Am Stat 52:1–9
Speed TP (2005) Terence's stuff: How to do statistical research. IMS Bull 1:6. http://bulletin.imstat.org/archive/34/1
Stinnet SS, Derr JA, Gehan EA (2009) Statistical consulting. In: Encyclopedia of biostatistics, 2nd edn. Wiley, Chichester, UK
Tweedie R et al (1998) Consulting: real problems, real interactions, real outcomes. Stat Sci 13:1–29

Statistical Design of Experiments (DOE)

Jürgen Pilz
Professor, Head of the Institute of Statistics of the University of Klagenfurt
University of Klagenfurt, Klagenfurt, Austria

Model and Denotations

As in regression analysis, DoE is concerned with modelling the dependence of a random target variable Y in dependence of a number of controllable deterministic variables $x_1, \ldots, x_k$ (called *factors*). The major goal of DoE is to find configurations for $\boldsymbol{x} = (x_1, \ldots, x_k)$ out of a given region $V \subset R^k$ which lead to "optimal" results for the target variable under consideration. The different configurations $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}$ for the factors are summarized in a statistical design $d_n = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)}) \in V^n$ of *size* n. The optimality criterion is usually defined through some objective function, e.g., the information or ►entropy associated with an experiment, the variance of some predictor $\hat{Y}(\boldsymbol{x}^*)$ for an unobserved configuration $\boldsymbol{x}^* = (x_1^*, \ldots, x_k^*)$ etc. The main areas of concern in DoE are:

(a) statistical design in regression analysis and analysis of variance
(b) factorial designs
(c) identification and elimination of disturbing influences (blocking)

This often includes, as a first step, the design of the size of the experiment; i.e., the number of observations n to be taken in order to achieve a predefined goal, see e.g., Rasch et al. (2010). The mean function of $Y = Y(\boldsymbol{x})$ given $\boldsymbol{x} = (x_1, \ldots, x_k) \in V$ is called the *response surface*, usually denoted by $\eta(\boldsymbol{x}) = EY(\boldsymbol{x})$, and the model becomes

$$Y(\boldsymbol{x}) = \eta(\boldsymbol{x}) + \varepsilon, \ \boldsymbol{x} \in V \tag{1}$$

where the random error term is assumed to be independent of $\boldsymbol{x}$ and such that $E(\varepsilon) = Var(\varepsilon) = \sigma^2$. Interpreting $\boldsymbol{x}$ as realisation of a random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$, the response function is simply the regression function of Y w.r.t. $\boldsymbol{X}$. The unknown response surface is often modelled through a linear setup

$$\eta(\boldsymbol{x}) = \beta_0 + \beta_1 f_1(\boldsymbol{x}) + \ldots + \beta_r f_r(\boldsymbol{x}) \tag{2}$$

with given functions $f_1, \ldots, f_r$. For example, $\eta(\boldsymbol{x})$ could be a second order polynomial setup

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i \le j=1}^{k} \beta_{ij} x_i x_j \tag{3}$$

arising from a second order Taylor expansion of η. Here, the first sum contains all *main effects* $x_1, \ldots, x_k$ and the second sum contains the (second order) interactions $x_i x_j$.

Optimal Designs

For any given concrete design $d_n = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})$ of size n; where $\boldsymbol{x}_{(i)} = (x_{i1}, x_{i2}, \ldots, x_{ik})$; $i = 1 \ldots n$ are not necessarily distinct from each other, it is well-known that the estimated response surface yields the best linear unbiased estimate (BLUE)

$$\hat{\eta}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 f_1(\boldsymbol{x}) + \ldots + \hat{\beta}_r f_r(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^T \hat{\boldsymbol{\beta}}$$

where $\boldsymbol{f}(\boldsymbol{x}) = (1, f_1(\boldsymbol{x}), \ldots, f_r(\boldsymbol{x}))^T$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_r)^T$ provided the parameters are estimated by the method of ►least squares (LS); i.e., $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$.
Here $\boldsymbol{Y} = (Y(\boldsymbol{x}_{(1)}), \ldots, Y(\boldsymbol{x}_{(n)}))$ stands for the vector of observations taken at the design points and X for the so-called *design matrix*

$$X = (f_j(\boldsymbol{x}_{(i)})) = \begin{pmatrix} 1 & f_1(\boldsymbol{x}_{(1)}) & \ldots & f_r(\boldsymbol{x}_{(1)}) \\ \vdots & \vdots & & \vdots \\ 1 & f_1(\boldsymbol{x}_{(n)}) & & f_r(\boldsymbol{x}_{(n)}) \end{pmatrix} \tag{4}$$

which is of type $n \times (r+1)$. For a first order regression setup $\eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ we have $r = k$ and the design matrix has the simple form

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ & \vdots & & \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{x}_{(1)}^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_{(n)}^T \end{pmatrix} \tag{5}$$

Criteria for the optimal choice of a design, as e.g., minimum prediction variance, are based on the covariance matrix

$$Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

of the LSE $\hat{\beta}$. For i.i.d. normally distributed observations this matrix is proportional to the Fisher information matrix, therefore

$$M(d_n) = \frac{1}{n} X^T X \tag{6}$$

is called the *information matrix* of the design $d_n = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(n)})$. Thus it makes sense to base optimality criteria for designs on functionals of (the inverse of) this matrix.

Definition The design d_n^* is called

(a) *L-optimal* w.r.t. some positive definite matrix U if

$$tr\left(UM\left(d_n^*\right)^{-1}\right) = \min_{d_n} tr(UM(d_n)^{-1})$$

(b) *G-optimal* if it minimizes the maximum variance of $\hat{\eta}(x) = \boldsymbol{f}(\boldsymbol{x})^T \hat{\boldsymbol{\beta}}$ over some region $H \subset R^k$, i.e., $\max_{x \in H} f(x)^T M\left(d_n^*\right)^{-1} f(x) = \min_{d_n} \max_{x \in H} f(x)^T M(d_n)^{-1} f(x)$

(c) *D-optimal* if it minimizes the determinant:

$$\det\left(M\left(d_n^*\right)^{-1}\right) = \min_{d_n} \det\left(M(d_n)^{-1}\right)$$

Important special cases of L-optimality include A-optimality and $\boldsymbol{c}$-optimality, where $U = I_{r+1}$ and $U = \boldsymbol{c}\boldsymbol{c}^T$ for a given vector $\boldsymbol{c} \in R^{r+1}$, respectively. An A-optimal design minimizes the sum of the variances $Var(\hat{\beta}_0) + \ldots + Var(\hat{\beta}_r)$ and thus the average variance of the regression coefficients, and a **c**-optimal design minimizes the variance of the linear combination $Var(\boldsymbol{c}^T \hat{\boldsymbol{\beta}}) = Var(c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \ldots + c_r\hat{\beta}_r)$. A D-optimal design minimizes the volume of the dispersion (confidence) ellipsoid for $\hat{\boldsymbol{\beta}}$.

Further criteria and numerical procedures for the construction of optimal designs may be found in Pukelsheim (1993), Atkinson et al. (2001), and Fedorov and Hackl (1997) on the basis of fundamental results by *Kiefer* and Wolfowitz in the late 1950s and early 1960s. Bayesian extensions of this theory are given in Pilz (1991) and Chaloner and Verdinelli (1995). An extensive theory of optimal designs for correlated errors in a spatial setting can be found in Müller (2007), Pilz and Spöck (2008) and Spöck and Pilz (2010) develop a theory of optimal spatial design for the construction of environmental monitoring networks using spectral theory for random fields. Optimal designs for higher-dimensional random fields are considered in Santner et al. (2003), with applications in the area of the design of computer experiments, see also Fang et al. (2005). Here, *Kriging* approximation models are constructed and then used as surrogates for the computer model. The design problem then refers to the optimal choice of the inputs at which to evaluate the computer model. Several software toolboxes are available for constructing optimal designs, see, e.g., Santner et al. (2003), DACE (http://www.2.imm.dtu.dk/hbn/dace) and the R-toolbox DoE (see Rasch et al. 2010).

Factorial Designs

Contrary to the mathematically well-defined optimality criteria considered in the last section, it is also customary to consider heuristically motivated and "practically useful" criteria for the construction of designs. Briefly, the first branch is called the "Kiefer design theory" and the latter branch is referred to as "Box design theory," in honour of their pioneers.

We assume that the response surface can be sufficiently well described by a polynomial of degree $g \geq 1$ in $k \geq 2$ factors $x_1, \ldots, x_k$. In order to guarantee the non-singularity of the information matrix it is necessary that each factor can take at least $g + 1$ different values, the latter are called the *levels* of the factors. A *factorial design* then means a design which defines a subset of all possible combinations of the levels of the k factors. It is said to be a *full factorial design* if it contains all of the $(g + 1)^k$ combinations of the levels, otherwise it is said to be a *fractional factorial* design. In most applications the response surface is investigated in a sequential manner. In a first step, a screening of the essential factors has to be made, using tools from regression analysis or from multivariate analysis (e.g., ▶principal component analysis). Hereafter, a first order polynomial in the remaining (essential) factors is formed to study the response surface and quantities of interest (e.g., extrema). If this setup is insufficient then a second or third order polynomial setup is chosen and the factor levels are updated until no further significant improvements are obtained. A formal way for proceeding in this manner had already been developed by Box and Wilson in 1951, with the aim of finding factor configurations leading to optimum experimental results.

Full Factorial Designs of the Type 2^k

Usually, one starts with a full factorial design, where all factors are controlled at two levels, "high" and "low," say. Such a design contains 2^k configurations (design points). By an appropriate scaling the design region can be transformed to the k-dimensional cube $V = \{\boldsymbol{x} = (x_1, \ldots, x_k) : -1 \le x_i \le +1,\ i = 1, \ldots, k\}$ and the design points are just the vertices of the cube. The full factorial design of size $n = 2^k$, $d_n = FF(2^k)$ for short, allows the estimation of all 2^k parameters of the model

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{\substack{i,j=1 \\ i<j}} \beta_{ij} x_i x_{j+\ldots+} \beta_{12\ldots k} x_1 x_2 \ldots x_k \tag{7}$$

As an example, consider a full factorial 2^3 design with factors x_1, x_2, and x_3 which can be adjusted at two levels -1 ("low") and $+1$ ("high"), respectively. The design has $n = 8$ points and allows the estimation of all parameters of the model $\eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3$. The basic structure of this design is displayed in the following table:

Trial no.	Coding	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
1	(1)	–	–	–	+	+	+	–
2	a	+	–	–	–	–	+	+
3	b	–	+	–	–	+	–	+
4	c	–	–	+	+	–	–	+
5	ab	+	+	–	+	–	–	–
6	ac	+	–	+	–	+	–	–
7	bc	–	+	+	–	–	+	–
8	abc	+	+	+	+	+	+	+

The coding follows the usual standard in the literature; the letters $a, b, c, \ldots$ represent the factors $x_1, x_2, x_3, \ldots$ and are used to indicate that the corresponding factor is adjusted at the level $+1$.

It is easily seen that $M(d_n) = \frac{1}{n} X^T X = I_n$ for a full factorial $d_n = FF(2^k)$ and the estimated regression coefficients are uncorrelated, in case of normally distributed observations they are even independent, and have a simple structure: $\hat{\beta} = \frac{1}{n} X^T \boldsymbol{Y}$, $Cov(\hat{\beta}) = \frac{\sigma^2}{n} I_n$.

Such designs are called *orthogonal*, they can easily be constructed using Hadamard matrices. When restricting attention to first order polynomials $\eta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ then an $FF(2^k)$ design leads to minimum variance estimates with $Var(\hat{\beta}_i) = \sigma^2/n$, moreover these full factorial designs turn out to be A-, D- and G-optimal. Finally, the estimated response surface has variance $Var(\hat{\eta}(\boldsymbol{x})) = \frac{\sigma^2}{n}(1 + \boldsymbol{x}^T \boldsymbol{x})$ which only depends on the distance of $\boldsymbol{x} = (x_1, \ldots, x_k)$ from the center point $\boldsymbol{0} = (0, \ldots, 0)^T$ of the design region V. Such designs are called *rotatable*, i.e., for first order polynomial setups full factorial designs of the type 2^k are rotatable.

Fractional Factorial Designs of the Type 2^{k-p}

If the number of factors is getting large, then one is interested in having less than 2^k observations to reduce the experimental efforts. On the other hand, such a reduction is justified if it is clear that there are no higher-order interactions between all or some of the factors. In practical applications it is very common that only the main effects and second-order interaction effects matter. To illustrate this: a full factorial 2^6 design requires $n = 64$ observations, but only 6 degrees of freedom are needed to estimate the main effects and another 15 are needed for the extimation of the two-factorial interchanging effects. Thus, only one third of the 64 observations would be needed for parameter estimation if third- and higher-order interactions were negligible. Therefore, fractional (incomplete) factorial designs are widely used in practice. They had first been introduced by Finney in 1945.

We call a design d_n of size $n = 2^{k-p}$, $1 \le p < k$, a *fractional factorial design* of the type 2^{k-p} if it forms the 2^{-p}-th part of a full factorial design of type 2^k. Such designs are constructed algorithmically by means of p defining relations. To illustrate the ideas, let $k = 4$ and $p = 1$, i.e., we construct a half replication of the $FF(2^k)$ using the defining relation $x_4 = x_1 x_2 x_3$ or, equivalently, multiplying by $x_4, 1 = x_1 x_2 x_3 x_4$.

Using the coding of the previous full factorial $FF(2^3)$ for the new $FF(2^4)$ and observing the defining relation $1 = x_1 x_2 x_3 x_4$ we arrive at the coding for the required fractional factorial 2^{4-1} design: $(1), ab, ac, ad, bc, bd, cd, abcd$. Finally, using the alternative defining relation $1 = -x_1 x_2 x_3 x_4$ we arrive at the alternative 2^{4-1} design: $a, b, c, d, abc, abd, acd, bcd$. The union of both half replicates results in the full factorial $FF(2^4)$ design.

The reduction of the number of observations achieved with fractional factorial designs, however, comes at the price of confounded parameter estimates. In our example, multiplying the defining relation $1 = x_1 x_2 x_3 x_4$ by x_1, x_2, x_3, and x_4, respectively, we obtain $x_1 = x_2 x_3 x_4$, $x_2 = x_1 x_3 x_4$, $x_3 = x_1 x_2 x_4$, $x_4 = x_1 x_2 x_3$, which implies that the main effects parameters $\beta_1, \beta_2, \beta_3$, and β_4 are confounded with

the third-order interaction parameters $\beta_{234}, \beta_{134}, \beta_{124}$, and β_{123}, respectively. From the defining relation $1 = x_1x_2x_3x_4$ itself follows that the intercept term β_0 is confounded with the fourth-order interaction parameter β_{1234}. However, there is no confounding of main effects with low-order interaction (second order interaction) parameters $\beta_{12}, \ldots, \beta_{34}$. Designs d_n for which $n = 2^s$ for some integer $s \geq 2$ are called *regular*, designs for which $n = r + 1$ (= number of unknown regressin parameters) are called *saturated*. Clearly, full factorial as well as fractional factorial designs are regular; full factorial designs $FF(2^k)$ are saturated for the linear regression setup (7) including all possible interactions between the main factors. The construction of saturated orthogonal designs for the hypercube region $V = \{\boldsymbol{x} = (x_1, \ldots, x_k) : -1 \leq x_i \leq +1, i = 1, \ldots, k\}$ is only possible for sizes n which are multiples of 4, such designs had already been constructed by Plackett and Burman in 1946.

Blocking in Factorial Designs

Random disturbances in the experimental conditions lead to an increased variance of the experimental error. In order to reduce this variance it is necessary to randomize the sequence of level combinations of a given design. If the number of factors k is getting larger (which usually implies an increased duration of experimentation in time) then systematic changes in the experimental conditions can occur (e.g., changing weather conditions in agricultural experiments). In this case, reductions in the variance of the experimental error can be achieved by *blocking*. *Blocks* are subsets of an experimental design which are constructed such that they guarantee the homogeneity of experimental conditions within the corresponding subsets. Such blocks can be formed, e.g., from subsets of full or fractional factorial designs, the sequence of trials within the blocks again chosen at random. For example, having k factors $x_1, \ldots, x_k$ and assuming that only the main effects and two-factorial interaction effects are significant, then the response surface takes the form

$$\eta(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{\substack{i,j=1 \\ i<j}}^{k} \beta_{ij} x_i x_j$$

For an unconfounded estimation of the effects a full factorial $FF(2^k)$ may be chosen, or, for $k \geq 6$, some fractional factorial 2^{k-p} with small $p \geq 1$. In order to take account of the block effect a block factor x_B is introduced, adjusted to the levels of the product $x_1x_2 \ldots x_k$ (or some other generator when starting with a fractional factorial). The block factor x_B can then be interpreted as an indicator variable taking values +1 and −1, and the resulting design can be interpreted as a fractional factorial design of type $2^{(k+1)-1}$ wtih the defining relation $1 = x_1x_2 \ldots x_kx_B$. Assuming the interaction effects $\beta_{1B}, \ldots, \beta_{kB}, \beta_{12B}, \beta_{13B}, \ldots, \beta_{12\ldots kB}$ to be negligible, the main effects and two-factorial interaction effects can be estimated without confounding. Moreover, since the design is orthogonal, blocking has no influence on these estimates.

For further results on fractional factorial designs, blocking, multilevel designs and other topics relevant in the vast field of statistical (optimum) experimental design we refer to the extensive monograph by Wu and Hamada (2009).

About the Author

Dr. Jürgen Pilz is a Professor and Head, Department of Statistics, University of Klagenfurt, Austria. He is the Head of the Department of Statistics since 2007. He is also Director of the Ph.D study program in Mathematics and Statistics at the University of Klagenfurt. He is an Elected member of the International Statistical Institute (1996). He has authored and co-authored more than 100 papers and 6 books, including *Bayesian Estimation and Experimental Design in Linear Regression Models* (Wiley 1991) and *Interfacing Geostatistics and GIS* (Springer, 2009). Professor Pilz was Associate Editor of the following international journals: *Journal of Statistical Planning and Inference* (1992–1999) and Metrika (2004–2009). Currently, he is an Associate editor of *Stochastic Environmental Research and Risk Assessment*. He has supervised more than 25 Ph.D dissertations.

Cross References

►Clinical Trials: An Overview
►Design of Experiments: A Pattern of Progress
►Factorial Experiments
►Optimum Experimental Design
►Randomization
►Uniform Experimental Design

References and Further Reading

Atkinson AC, Bogacka B, Zhigljavsky A (eds) (2001) Optimum design – 2000. Kluwer, Dordrecht, The Netherlands

Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. Stat Sci 10:237–304

Fang KT, Fang K, Runze L (2005) Design and modeling for computer experiments. Chapman & Hall/CRC Press, Boca Raton, FL

Fedorov VV, Hackl P (1997) Model-oriented design of experiments. Lecture notes in statistics 125. Springer, Berlin

Müller WG (2007) Collecting spatial data: optimum design of experiments for random fields. Springer, Berlin

Pilz J (1991) Bayesian estimation and experimental design in linear regression models. Wiley, Chichester, UK

Pilz J, Spöck G (2008) Bayesian spatial sampling design. In: Ortiz JM, Emery X (eds) Proceedings of 8th international geostatistics congress Gecamin Ltd., Santiago de Chile, pp 21–30
Pukelsheim F (1993) Optimal design of experiments. Wiley, New York
Rasch D, Pilz J, Verdooren R, Gebhardt A (2011) Optimal Experimental Design with R. Chapman & Hall/CRC Press, Boca Raton, FL
Santner Th, Williams BJ, Notz W (2003) The design and analysis of computer experiments. Springer, Berlin
Spöck G, Pilz J (2010) Spatial sampling design and covariance-robust minimax prediction based on convex design ideas. Stoch Environ Res Risk Assess 24(3):463–482
Wu CFJ, Hamada M (2009) Experiments: planning, analysis and parameter design optimization, 2nd edn. Wiley, New York

Statistical Distributions: An Overview

Kalimuthu Krishnamoorthy
Philip and Jean Piccione Professor of Statistics
University of Louisiana at Lafayette, Lafayette, LA, USA

Introduction

Statistical distributions are used to model sample data that were collected from a population or to model the outcomes of a *random* experiment. The statistical distribution is simply the probability distribution of a random variable. These probability models are commonly used in many applied areas such as economics, education, engineering, social, health, and biological sciences. The distributions of discrete random variables (whose possible values are countable) are referred to as the discrete distribution while those of continuous random variables are called continuous distribution. To begin with an example, let X denote the number of heads that can be observed by flipping a fair coin three times. The sample space of X includes eight outcomes, namely, HHH, HTH, THH, TTH, HHT, HTT, THT, TTT, where H denotes the head and T denotes the tail. The probability that X equals one is the probability of observing any one of the mutually exclusive outcomes TTH, HTT and THT. As all eight outcomes are equally likely, $P(X = 1) = \frac{3}{8}$. Proceeding this way, we obtain the probability distribution of X as

x	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

The above distribution is a member of the family of binomial distributions indexed by n and p, where n is the number of independent Bernoulli trials (each trial results into either "success" or "failure") and p is the probability of observing a success in each trial. The function that gives the probability that a discrete random variable takes a specified value is referred to as the probability mass function (pmf). For example, the pmf of a binomial random variable is given by

$$P(X = x|n,p) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0,1,2,\ldots,n.$$

For a continuous random variable X, $P(X = x) = 0$ for any fixed x, and so we consider only $P(X \in A)$ for any given interval $A \in R$, and this probability can be evaluated as $P(X \in A) = \int_A f(x;\theta)dx$, where $f(x;\theta)$ is called the probability density function (pdf), and θ is a parameter vector. The pdf $f(x)$ should satisfy two conditions: $f(x) \geq 0$ for all x, and $\int_{-\infty}^{\infty} f(x;\theta)dx = 1$.

In the following we shall list some commonly used discrete and continuous distributions, their physical significance, relations among them and some measures that describe features of a distribution.

Discrete Distributions

Most commonly used discrete distributions are the binomial, Poisson, hyper geometric, negative binomial and logarithmic series distributions. The first four distributions are closely related. The ▶binomial distribution is used to estimate the proportion of individual with an attribute of interest in a population. In particular, the number of individuals with an attribute of interest in a random sample from a large population (e.g., proportion of defective items in a large shipment) is a binomial random variable with the sample size as the value of n, and the true proportion (usually unknown) in the sampled population is the parameter p. On the other hand, if the sample is drawn (without replacement) from a finite population, then the number of units in the sample with the characteristic of interest is a hypergeometric random variable with the size of the population N (usually known) as the "lot size," the true number of units M (usually unknown) with the attribute in the population as the parameter, and the sample size n as another (known) parameter. The pmf of a hypergeometric random variable is given by $P(X = x|n,M,N) = \binom{M}{x}\binom{N-M}{n-x}/\binom{N}{n}$, $L \leq x \leq U$, where $L = \max\{0, M-N+n\}$ and $U = \min\{n, M\}$. If the population is reasonably large, then one could use the binomial model instead of the hypergeometric.

The Poisson distribution (see ▶Poisson Distribution and Its Application in Statistics) is postulated to model the probability distribution of rare events. Specifically, if

Statistical Distributions: An Overview. Table 1 Some discrete distributions

Distribution	Probability mass function	Description
Uniform	$f(x;N) = \frac{1}{N}, \quad k = 1, \ldots, N.$	Positive integer N
Binomial	$f(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n$	n = No. of trials p = Success probability
Hypergeometric	$f(x;n,M,N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$, $\max\{0, M-N+n\} \le x \le U = \min\{n, M\}$	n = Sample size; M = No. of defects N = Lot size
Poisson	$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$	λ = Mean
Geometric	$f(x;p) = (1-p)^x p$, $x = 0, 1, 2, \ldots$	p = Success probability x = No. of failures until the first success
Negative binomial	$f(x;r,p) = \binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, 2, \ldots$	p = Success probability x = Number of failures until the rth success
Logarithmic series	$f(x;\theta) = -\frac{\theta^x}{x \ln(1-\theta)}$	$0 < \theta < 1$

an event is almost unlikely to occur in a moment of time, but the number of occurrences over a long period of time could be very large, then a Poisson model is appropriate to describe the frequency distribution of the event. This description implies that the binomial distribution with large n and small p can be approximated by a Poisson distribution with mean $\lambda = np$. More specifically, for a binomial(n,p) random variable with large n and small p, $P(X \le x|n,p) \simeq \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}$, $x = 0, 1, \ldots, n$, where $\lambda = np$ and $e^{-\lambda}\lambda^x/x!$ is the pmf of a Poisson random variable with mean λ.

The geometric distribution arises as the probability distribution of number of trials in a sequence of independent Bernoulli trials needed to get the first success. The negative-binomial distribution is a generalization of the geometric distribution where we consider the number of trials required to get r successes. Note that in the binomial distribution, the number of successes in a fixed number of independent Bernoulli trials is a random variable where as in the case of negative-binomial the number of trials is a random variable. The number of failures K in a sequence of independent Bernoulli trials that can be observed before observing exactly r successes is also referred to as the negative-binomial random variable. In the former case, n takes on values $r, r+1, r+2, \ldots$. whereas in the latter case K takes on values 0, 1, 2,…. Both binomial and negative-binomial distributions are related to the beta distribution: If X is a binomial(n, p) random variable then, for $x \neq 0$, $P(X \ge x|n,p) = P(Y \le p)$, where Y is a beta$(x, n-x+1)$ random variable. Also, for $x \neq n$ $P(X \le x|n,p) = P(W \ge p)$, where W is a beta$(x+1, n-x)$ random variable. If X is the number of failures before the rth success (in a sequence of independent Bernoulli trials), then $P(X \le x|r,p) = P(W \le p)$, where W is a beta$(r, x+1)$ random variable. Similar relation exists between the Poisson and the chi-square distributions. Specifically, $P\left(\chi^2_n > x\right) = P(Y \le n/2 - 1)$, where Y is a Poisson random variable with mean $x/2$.

The probability mass function of a logarithmic series distribution with parameter θ is given by $P(X = k) = \frac{a\theta^k}{k}$, $0 < \theta < 1$, $k = 1, 2, \ldots$, where $a = -1/[\ln(1-\theta)]$. The logarithmic series distribution is useful to describe a variety of biological and ecological data. It is often used to model the number of individuals per species. This distribution is also used to fit the number of products requested per order from a retailer.

Some popular discrete distributions are listed in Table 1. For detailed descriptions, properties and applications of various discrete distributions, see the books by Johnson et al. (1992), Evans et al. (2000), and Krishnamoorthy (2006).

Continuous Distributions

Continuous distributions are grouped into a few families based on the form of pdfs: location family, scale family, location-scale family and exponential family, etc. In the following we shall describe some of these families.

Location-Scale Family: The pdf of a location-scale distribution can be expressed as $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where μ is the location parameter, $\sigma > 0$ is the scale parameter and f is any

pdf that does not depend on any parameter. As an example, the pdf of a normal distribution can be expressed as

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right), \text{ with}$$
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The two-parameter exponential distribution, normal, Cauchy, double exponential (Laplace), extreme-value and logistic are popular location-scale distributions. The cumulative distribution function (cdf) of a location-scale random variable can be computed using its standard form as $P(X \le x) = P\left(Z \le \frac{x-\mu}{\sigma}\right)$. For a location-scale family, $\frac{\hat{\mu}-\mu}{\sigma}$ and $\frac{\hat{\sigma}}{\sigma}$ are pivotal quantities provided $\hat{\mu}$ and $\hat{\sigma}$ are equivariant estimators. These pivotal quantities are useful to find inferential procedures for μ, σ or for any invariant function of (μ, σ).

The normal distribution is the most popular among the location-scale families. In fact there is nothing inherently normal about the normal distribution, and its common use in applications is due its simplicity. Distributions of many commonly used statistics can be approximated by the standard normal distribution via the central limit theorem (see ►Central Limit Theorems). Furthermore, the asymptotic distribution of a maximum likelihood estimator is normal with the variance determined by the Fisher information matrix.

Exponential Family: A family of distributions whose pdf or pmf can be written in the form $f(x;\theta) = h(x)c(\theta) \exp\left(\sum_{i=1}^{k} q_i(\theta) w_i(x)\right)$ is called an exponential family. As an example, the binomial family is an exponential family because the pmf $f(x;p) = h(x)c(p) \exp(q_1(p)w_1(x))$, with $h(x) = \binom{n}{x}$, $c(p) = (1-p)^n$, $q_1(p) = \ln(p/(1-p))$ and $w_1(x) = x$. The normal distribution and lognormal distribution are members of exponential families. A statistical model from an exponential family is easy to work with because exponential families have some nice mathematical properties. For instance, it is easier to find sufficient statistics for an exponential family. In fact, for a sample $X_1, \ldots, X_n$ from an exponential family, $\left(\sum_{i=1}^{n} w_1(X_i), \ldots, \sum_{i=1}^{n} w_k(X_i)\right)$ is a sufficient statistic for θ.

Some distributions are routinely used to model lifetime data, and they are referred to as lifetimes (or failure times) distributions. The ►Weibull distribution is one of the most widely used lifetime distributions in reliability and survival analysis. It is a versatile distribution that can take on the characteristics of other types of distributions, based on the value of the shape parameter. If X follows a Weibull distribution with shape parameter c and the scale parameter b, then $\ln(X)$ has the extreme-value distribution with the location parameter $\mu = \ln(b)$ and the scale parameter $\sigma = 1/c$. This one–one relation allows us to transform the results based on a Weibull model to an extreme-value distribution (see ►Weibull distribution). Other lifetime distributions include exponential, two-parameter exponential, lognormal, and gamma distributions. Some popular continuous distributions are listed in Table 2.

Relations Among Distributions: Many of the continuous distributions have one–one relation with others. For example, normal and lognormal (via logarithmic transformation of lognormal random variable), two-parameter exponential and Pareto (via logarithmic transformation of Pareto random variable), two-parameter exponential and power distribution (via negative log transformation of power random variable). This one–one relation enables us to transform some invariant inferential procedures for one distribution to another. Another important distribution that has relation with the t, F, binomial and negative binomial distributions is the beta distribution. An efficient program that evaluates the beta distribution can be used to computes the cumulative distribution functions (cdfs) of other related random variables just cited. The gamma distribution with the shape parameter $\alpha = n/2$ and the scale parameter $\beta = 2$ specializes to the ►chi-square distribution with n degrees of freedom; when $\alpha = 1$, it simplifies to the exponential distribution with mean β. A diagram that describes relations among various distributions is given in Casella and Berger (2002, p. 627).

Moments and Other Measures

Moments are set of measures that are useful to judge some important properties of a probability distribution. Mean and median are commonly used measure of location or center of the distribution. Range and variance are used to quantify the variability of a random variable. We shall now overview some of these measures that describe important characteristics of a distribution.

The mean of a random variable is usually denoted by μ, which is expectation of the random variable. For a discrete random variable X, $\mu = E(X) = \sum_k kP(X = k)$, where the sum is over all possible values of X. If X is continuous, then $\mu = \int_{-\infty}^{\infty} xf(x)dx$, where $f(x)$ is the pdf of X. The expectation $E(X^k)$, $k = 1, 2, \ldots$, is referred to as the kth moment about the origin, while $E(X-\mu)^k$ is called the kth moment about the mean or the kth *central moment*. The second moment about the mean is the variance (denoted by σ^2), and its positive square root is called the standard deviation. The absolute moment $E(|X-\mu|)$ is referred to as the *mean deviation*. The mean deviation and variance

Statistical Distributions: An Overview. Table 2 Some continuous distributions

Distribution	Probability density function	Description of parameters
Uniform	$f(x;a,b) = \frac{1}{b-a}, \quad a \le x \le b$	$a < b$; known or unknown
Normal	$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$	$-\infty < \mu < \infty, \sigma > 0$ Mean μ Standard deviation σ
Chi-square	$f(x;n) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{n/2-1}, \quad x > 0$	Degrees of freedom (df) $n > 0$
F-distribution	$f(x;m,n) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{\left(\frac{m}{2}\right)^{m/2} x^{m/2-1}}{\left(\frac{n}{2}\right)^{m/2}\left[1+\frac{mx}{n}\right]^{m/2+n/2}}, x > 0$	m = Numerator df n = Denominator df
Student's-*t*	$f(x;n) = \frac{\Gamma[(n+1)/2]}{\Gamma(n/2)\sqrt{n\pi}} \frac{1}{(1+x^2/n)^{(n+1)/2}}, \quad -\infty < x < \infty$	df $n \ge 1$
Exponential	$f(x;\mu,\sigma) = \frac{1}{\sigma} \exp\left(-\frac{(x-\mu)}{\sigma}\right), \quad x > \mu$	Location μ Scale $\sigma > 0$
Gamma	$f(x;a,b) = \frac{1}{\Gamma(a)b^a} e^{-x/b} x^{a-1}, \quad x > 0$	Shape $a > 0$ Scale $b > 0$
Beta	$f(x;a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$	Shape $a > 0$ Scale $b > 0$
Noncentral Chi-square	$f(x;n,\delta) = \sum_{k=0}^{\infty} \frac{\exp\left(-\frac{\delta}{2}\right)\left(\frac{\delta}{2}\right)^k}{k!} \frac{\exp\left(-\frac{x}{2}\right)x^{\frac{n+2k}{2}-1}}{2^{\frac{n+2k}{2}}\Gamma\left(\frac{n+2k}{2}\right)}$	df $n > 0$ δ = Noncentrality parameter > 0
Noncentral *F*	$\text{cdf} = \sum_{k=0}^{\infty} \frac{\exp\left(-\frac{\delta}{2}\right)\left(\frac{\delta}{2}\right)^k}{k!} P\left(F_{m+2k,n} \le \frac{mx}{m+2k}\right)$	Numerator df $m > 0$ Denominator df $n > 0$ Noncentrality parameter $\delta > 0$
Noncentral *t*	$f(x;n,\delta) = \frac{n^{n/2}\exp(-\delta^2/2)}{\sqrt{\pi}\,\Gamma(n/2)(n+x^2)^{(n+1)/2}} \sum_{i=0}^{\infty} \frac{\Gamma[(n+i+1)/2]}{i!} \left(\frac{x\delta\sqrt{2}}{\sqrt{n+x^2}}\right)^i$	df $n \ge 1$ $-\infty < \delta < \infty$
Laplace (Double exponential)	$f(x;a,b) = \frac{1}{2b} \exp\left[-\frac{\lvert x-a\rvert}{b}\right], -\infty < x < \infty$	$-\infty < a < \infty, b > 0$ Location a, scale $b > 0$
Logistic	$f(x;a,b) = \frac{1}{b} \frac{\exp\left\{-\left(\frac{x-a}{b}\right)\right\}}{\left[1+\exp\left\{-\left(\frac{x-a}{b}\right)\right\}\right]^2}, -\infty < x < \infty$	Location a, scale $b > 0$
Lognormal	$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}x\sigma} \exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right], \quad x > 0$	$\sigma > 0, -\infty < \mu < \infty$
Pareto	$f(x;a,b) = \frac{ba^b}{x^{b+1}}, x \ge a$	$a > 0; b > 0$
Weibull	$f(x;b,c,m) = \frac{c}{b}\left(\frac{x-m}{b}\right)^{c-1} \exp\left\{-\left[\frac{x-m}{b}\right]^c\right\}, x > m$	Scale $b > 0$ Shape $c > 0$ Location m
Extreme-value	$f(x;a,b) = \frac{1}{b} \exp\left[-\frac{x-a}{b}\right] \exp\left\{-\exp\left[-\frac{x-a}{b}\right]\right\}$	Location a Scale $b > 0$
Cauchy	$f(x;a,b) = \frac{1}{\pi\, b[1+((x-a)/b)^2]}, \quad -\infty < x < \infty$	Location a, scale $b > 0$
Inverse Gaussian	$f(x;\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left(\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right), x > 0$	$\lambda > 0, \mu > 0$

are used to judge the spread of a distribution. The measure of variability that is independent of the units of measurements is called *coefficient of variation*, and is defined as (standard deviation/mean = σ/μ).

The measures that are used to judge the shape of a distribution are the *coefficient of* ►*skewness* and the *coefficient of kurtosis* (see ►Kurtosis: An Overview). The coefficient of skewness is defined as (the third moment about the

mean)/(variance)$^{3/2}$. The skewness measures the lack of symmetry. A negative coefficient of skewness indicates that the distribution is left-skewed (larger proportion of the population is below the mean) while a positive value indicates that the distribution is right-skewed. The *coefficient of kurtosis*, defined as γ = (the fourth moment about the mean)/(variance)^2, is a measure of peakedness or flatness of the probability density curve. As an example, for the normal distribution, the coefficient of skewness is zero (symmetric about the mean), and the coefficient of kurtosis is three. For a Student t distribution with n degrees of freedom, the coefficient of skewness is zero and the coefficient of kurtosis is $3(n-2)/(n-4)$, which approaches 3 as $n \to \infty$.

The ►moment generating function for a random variable is defined as $M_X(t) = E[e^{tX}]$ provided the expectation exists for t in some neighborhood of zero. Note that the kth derivative of $M_X(t)$ evaluated at $t = 0$ is $E(X^k)$, the kth moment about the origin. The logarithm of moment generating function, $G_X(t) = \ln(M_X(t))$, is called the cumulant generating function. The kth derivative of $G_X(t)$ evaluated at $t = 0$ is the kth moment about the mean. Thus, $G'(t)|_{t=0} = \mu$, $G''(t)|_{t=0} = \sigma^2$, and so on.

Fitting a Probability Model

There are several methods available to fit a probability distribution for a given sample data. A popular simple method is quantile–quantile plot (Q-Q plot) which is the plot of the sample quantiles (percentiles) and the corresponding population quantiles. The population quantiles are usually unknown, and they are obtained using the estimates of the model parameters. If the Q–Q plot exhibits a linear pattern, then the data can be regarded as a sample from the postulated probability distribution. There are other rigorous approaches available to check if the sample is from a specific family of distributions. For instance, the Wilks–Shapiro test and the Anderson–Darling test (see ►Anderson-Darling Tests of Goodness-of-Fit) are popular tests to determine if the sample is from a normal population. Another well-known nonparametric test is the ►Kolmogorov–Smirnov test which is based on the difference between the empirical distribution of the sample and the cumulative distribution function of the hypothesized probability model.

Multivariate Distributions

The probability distribution of a random vector is called multivariate distribution. In general, it is assumed that all the components of the random vector are continuous or all of them are discrete. The most popular continuous multivariate distribution is the multivariate normal (see ►Multivariate Normal Distributions). A random vector $\boldsymbol{X}$ is multivariate normally distributed with mean vector $\boldsymbol{\mu}$ and the variance–covariance matrix $\boldsymbol{\Sigma}$ if and only if $\boldsymbol{\alpha X} \sim N(\boldsymbol{\alpha'\mu}, \boldsymbol{\alpha'\Sigma\alpha})$ for every non-zero $\boldsymbol{\alpha'} \in R^p$. Many results and properties of the univariate normal can be extended to the multivariate normal distribution (see ►Multivariate Normal Distributions) using this definition. Even though there are other multivariate distributions, such as multivariate gamma and multivariate beta, are available in literature, their practical applications are not well-known. One of the most popular books in the area of multivariate analysis is Anderson (2003) and its earlier editions.

A popular multivariate discrete distribution is the ►multinomial distribution, which is a generalization of the ►binomial distribution. This distribution is routinely used to analyze the categorical data in the form of contingency table. Another distribution to model a sample of categorical vector observations from a finite population is the multivariate hypergeometric distribution. A useful reference for multivariate discrete distributions is the book by Johnson et al. (1997).

About the Author

Dr. Kalimuthu Krishnamoorthy is Professor, Department of Mathematics, University of Louisiana at Lafayette, Louisiana, USA. He is holder of Philip and Jean Piccione Professor of statistics. He has authored and co-authored more than 75 papers and 2 books, *Handbook of Statistical Distributions with Applications* (Chapman & Hall/CRC, 2006), and *Statistical Tolerance Regions: Theory, Applications and Computation* (Wiley 2009). He is currently an Associate editor for *Communications in Statistics*.

Cross References

►Approximations to Distributions
►Beta Distribution
►Binomial Distribution
►Bivariate Distributions
►Chi-Square Distribution
►Contagious Distributions
►Distributions of Order K
►Exponential Family Models
►Extreme Value Distributions
►F Distribution
►Financial Return Distributions
►Gamma Distribution
►Generalized Extreme Value Family of Probability Distributions
►Generalized Hyperbolic Distributions
►Generalized Rayleigh Distribution
►Generalized Weibull Distributions

►Geometric and Negative Binomial Distributions
►Heavy-Tailed Distributions
►Hyperbolic Secant Distributions and Generalizations
►Hypergeometric Distribution and Its Application in Statistics
►Inverse Gaussian Distribution
►Location-Scale Distributions
►Logistic Distribution
►Logistic Normal Distribution
►Multinomial Distribution
►Multivariate Normal Distributions
►Multivariate Statistical Distributions
►Normal Distribution, Univariate
►Poisson Distribution and Its Application in Statistics
►Relationships Among Univariate Statistical Distributions
►Skew-Normal Distribution
►Skew-Symmetric Families of Distributions
►Student's t-Distribution
►Testing Exponentiality of Distribution
►Uniform Distribution in Statistics
►Univariate Discrete Distributions: An Overview
►Weibull Distribution

References and Further Reading

Anderson TW (2003) An introduction to multivariate statistical analysis. Wiley, New York
Casella G, Berger RL (2002) Statistical inference. Duxbury, Pacific Grove, CA
Evans M, Hastings N, Peacock B (2000) Statistical distributions. Wiley, New York
Johnson NL, Kotz S, Kemp AW (1992) Univariate discrete distributions. Wiley, New York
Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions. Wiley, New York
Johnson NL, Kotz S, Balakrishnan N (1997) Discrete multivariate distributions. Wiley, New York
Krishnamoorthy K (2006) Handbook of statistical distributions with applications. Chapman & Hall/CRC Press, Boca Raton, FL
Patel JK, Kapadia CH, Owen DB (1976) Handbook of statistical distributions. Marcel Dekker, New York

Statistical Ecology

David Fletcher
Associate Professor
University of Otago, Dunedin, New Zealand

Ecologists study complex systems, and often need to use non-standard methods of sampling and data analysis. The data might be collected over a long-time scale, involve little spatial replication, or be highly aggregated in space. There have been many fruitful collaborations between ecologists and statisticians, often leading to the development of new statistical methods. In this brief overview of the subject, I will focus on three areas that have been of particular interest in the management of animal populations. I will also discuss the use of statistical methods in other areas of ecology, the aim being to highlight interesting areas of development rather than a comprehensive review.

Mark-Recapture Methods

Mark-recapture methods are commonly used to estimate abundance and survival rates of animal populations (Lebreton et al. 1992; Williams et al. 2002). Typically, a number of individuals are physically captured, marked and released. The information obtained from successive capture occasions is summarized in a "capture history," which indicates whether or not an individual was captured on the different occasions. The likelihood is specified in terms of demographic parameters of interest, such as annual survival probabilities, and nuisance parameters that model the capture process. A range of goodness-of-fit diagnostics have been developed, including estimation of overdispersion (Anderson et al. 1994). Overdispersion usually arises as a consequence of heterogeneity, or lack of independence, amongst individuals in the survival and/or capture probabilities; attempts have also been made to model such heterogeneity directly (Pledger et al. 2003). ►Model selection often involves use of ►Akaike's information criterion (AIC), and model-averaging is also commonly used (Johnson and Omland 2004). Bayesian methods are becoming popular, particularly as means of fitting hierarchical models (Brooks et al. 2000). Recent developments include the use of genotyping of fecal, hair or skin samples to identify individuals (Lukacs and Burnham 2005; Wright et al. 2009), and spatially-explicit models that allow estimation of population density (Borchers and Efford 2008). A related area of recent interest has been the estimation of the occupancy rate, i.e., the proportion of a set of geographical locations that are occupied by a species (MacKenzie et al. 2006). This can be of interest in large-scale monitoring programs, for which estimation of abundance is too costly, and in understanding metapopulation dynamics. In this setting, the "individuals" are locations and the "capture history" records whether or not a species was observed at that location, on each of several occasions.

Distance Sampling

A common alternative method for estimating population abundance or density is distance sampling. This involves recording the distance of each observed individual from

a transect line or a point. The analysis then involves estimation of the probability of detection of an individual as a function of distance (Buckland et al. 2004), thereby allowing estimation of the number of individuals that have not been detected. Two important assumptions in using this method is that detection is certain for an individual on the line or point and that individuals do not move during the observation process, although modifications have been suggested for situations in which these assumptions are not met (Borchers et al. 1998; Buckland and Turnock 1992). Compared to the use of mark-recapture methods for estimating abundance, distance sampling typically provides savings in terms of field effort, and will usually be more appropriate when the population is widely dispersed. A useful discussion of the theory underlying use of distance sampling is given by Fewster and Buckland (2004), while Schwarz and Seber (1999) provide an extensive review of methods for estimating abundance.

Population Modeling

Population projection models have long been used as a tool in the process of managing animal and plant populations, most often as means of assessing the impact of management on the population growth rate or on the probability of quasi-extinction (Caswell 2001; Burgman et al. 1993). A population model will typically involve one or more demographic parameters, such as annual survival probabilities and annual reproductive rates, for individuals in different ages or stages. In the past, estimation of the parameters has been performed by separately fitting statistical models to the different sets of data; recent work in this area has focussed on regarding the population model as a statistical model that can be fitted to all the available data (Buckland et al. 2007). The benefit of this approach is that all the uncertainty can be allowed for, and that estimation of the parameters can be improved by including data that provide a direct indication of the population growth rate (Besbeas et al. 2002). This development has the potential to allow ecologists to fit a broad range of population models to their data, including ones that allow for immigration (cf., Nichols and Hines 2002; Peery et al. 2006).

Other Developments

A key aspect of studying many plant and animal populations is their aggregated spatial distribution. This distribution might be of interest in itself, or be something that needs to be allowed for in the sampling and data analysis. There is a long tradition of the analysis of spatial pattern in ecology, involving a range of statistical techniques, including distance-based methods and spatial ►point processes (Fortin and Dale 2005). Various statistical distributions have been suggested as a means of allowing for the fact that aggregation often leads to zero-inflated and/or positively skewed data. These include the negative binomial, lognormal and gamma distributions, plus zero-inflated versions of these (Dennis and Patil 1984; Martin et al. 2005; Fletcher 2008). Likewise, methods have been developed for fitting models that incorporate spatial autocorrelation (Legendre 1993; Fortin and Dale 2005).

►Adaptive sampling is a modification of classical sampling that aims to allow for spatial aggregation by adaptively increasing the sample size in those locations where the highest abundances have been found in an initial sample (Thompson and Seber 1996; Brown and Manly 1998). Information on the number and relative abundance of individual species in one or more geographical areas has been of interest to many ecologists, leading to the use of species abundance models (Hughes 1986; Hill and Hamer 1998), estimation of species richness (Chao 2005), modeling species-area relationships (Connor and McCoy 2001), and the analysis of species co-occurrence (Mackenzie et al. 2004; Navarro-Alberto and Manly 2009).

In studying ecological communities, it is often natural to consider the use of multivariate methods. There is a large literature in this area, primarily focussing on classification and ordination techniques for providing informative summaries of the data (McGarigal et al. 2000). Likewise, multivariate analysis of variance (see ►Multivariate Analysis of Variance (MANOVA)) has been used to assess the ecological impact of human disturbance on a range of species (Anderson and Ter Braak 2003).

In order to study processes operating at large spatial scales, it is useful to carry out studies at those scales. In doing so, there is a tension between satisfying the statistical requirements of replication and keeping the study at a scale that is large enough to provide meaningful results (Schindler 1998; Hewitt et al. 2007). There has been some discussion in the ecological literature regarding appropriate statistical methods for such studies (Cottenie and De Meester 2003). One approach is to consider a single large-scale study as insufficient to provide the level of evidence that is usually required of a small-scale experiment, with the hope that information from a number of studies can eventually be combined, either informally of using meta analysis (Gurevitch and Hedges 1999).

Future

It is clear that the increasing popularity of computationally-intensive Bayesian methods of analysis will lead to ecologists being able to fit statistical models that provide them

with a better understanding of the spatial and temporal processes operating in their study populations (Clark 2007). Likewise, recently-developed techniques such as ►neural networks (Lek et al. 1996) and boosted trees (Elith et al. 2008), are likely to appear more frequently in the ecological literature. In tandem with the development of new techniques, there will always be a need to balance complexity and simplicity in the analysis of ecological data (Murtaugh 2007).

About the Author

David Fletcher is regarded as one of New Zealand's top ecological statisticians. He has worked in statistical ecology since arriving in New Zealand 20 years ago, both in academia and as a private consultant. He is the author of 70 refereed papers and numerous technical reports for government agencies.

Cross References

►Adaptive Sampling
►Akaike's Information Criterion
►Analysis of Areal and Spatial Interaction Data
►Distance Sampling
►Non-probability Sampling Survey Methods
►Spatial Point Pattern
►Statistical Inference in Ecology

References and Further Reading

Anderson MJ, Ter Braak CJF (2003) Permutation tests for multi-factorial analysis of variance. J Stat Comput Sim 73:85–113

Anderson DR, Burnham KP, White GC (1994) AIC model selection in overdispersed capture-recapture data. Ecology 75:1780–1793

Besbeas P, Freeman SN, Morgan BJT, Catchpole EA (2002) Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. Biometrics 58:540–547

Borchers DL, Efford MG (2008) Spatially explicit maximum likelihood methods for capture-recapture studies. Biometrics 64:377–385

Borchers DL, Zucchini W, Fewster RM (1998) Mark-recapture models for line transect surveys. Biometrics 54:1207–1220

Brooks SP, Catchpole EA, Morgan BJT (2000) Bayesian annual survival estimation. Stat Sci 15:357–376

Brown JA, Manly BJF (1998) Restricted adaptive cluster sampling. Environ Ecol Stat 5:49–63

Buckland ST, Turnock BJ (1992) A robust line transect method. Biometrics 48:901–909

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (eds) (2004) Advanced distance sampling: estimating abundance of biological populations. Oxford University Press, Oxford

Buckland ST, Newman KB, Fernández C, Thomas L, Harwood J (2007) Embedding population dynamics models in inference. Stat Sci 22:44–58

Burgman MA, Ferson S, Akcakaya HR (1993) Risk assessment in conservation biology. Chapman and Hall, London

Caswell H (2001) Matrix population models, 2nd edn. Sinauer Associates, Massachusetts

Chao A (2005) Species richness estimation. In: Encyclopedia of statistical sciences, 2nd edn. Wiley, New York

Clark JS (2007) Models for ecological data: an introduction. Princeton University Press, Princeton, NJ

Connor EF, McCoy ED (2001) Species-area relationships. In: Encyclopedia of biodiversity, vol 5. Academic, New York, pp 397–411

Cottenie K, De Meester L (2003) Comment to Oksanen (2001): reconciling Oksanen (2001) and Hurlbert (1984). Oikos 100:394–396

Dennis B, Patil GP (1984) The gamma distribution and weighted multimodal gamma distributions as models of population abundance. Math Biosci 68:187–212

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813

Fewster RM, Buckland ST (2004) Chapter 10 of advanced distance sampling: estimating abundance of biological populations. In: Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (eds) Advanced distance sampling: estimating abundance of biological populations. Oxford University Press, Oxford

Fletcher DJ (2008) Confidence intervals for the mean of the delta-lognormal distribution. Environ Ecol Stat 15:175–189

Fortin M-J, Dale MRT (2005) Spatial analysis: a guide for ecologists. Cambridge University Press, Cambridge

Gurevitch J, Hedges LV (1999) Statistical issues in ecological meta-analyses. Ecology 80:1142–1149

Hewitt JE, Thrush SF, Dayton PK, Bonsdorff E (2007) The effect of spatial and temporal heterogeneity on the design and analysis of empirical studies of scale-dependent systems. Am Nat 169:398–408

Hill JK, Hamer KC (1998) Using species abundance models as indicators of habitat disturbance in tropical forests. J Appl Ecol 35:458–460

Hughes RG (1986) Theories and models of species abundance. Am Nat 128:879–899

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19:101–108

Lebreton J-D, Burnham KP, Clobert J, Anderson DR (1992) Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecol Monogr 62:67–118

Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? Ecology 74:1659–1673

Lek S, Delacoste M, Baran P, Dimopoulos I, Lauga J, Aulagnier S (1996) Application of neural networks to modelling nonlinear relationships in ecology. Ecol Model 90:39–52

Lukacs PM, Burnham KP (2005) Review of capture-recapture methods applicable to noninvasive genetic sampling. Mol Ecol 14:3909–3919

McArdle BH (1996) Levels of evidence in studies of competition, predation, and disease. New Zeal J Ecol 20:7–15

Mackenzie DI, Bailey LL, Nichols JD (2004) Investigating species co-occurrence patterns when species are detected imperfectly. J Anim Ecol 73:546–555

MacKenzie D, Nichols J, Royle J, Pollock K, Bailey L, Hines J (2006) Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Academic
McGarigal K, Cushman S, Stafford S (2000) Multivariate statistics for wildlife and ecology research. Springer, New York
Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett 8:1235–1246
Murtaugh PA (2007) Simplicity and complexity in ecological data analysis. Ecology 88:56–62
Navarro-Alberto JA, Manly BFJ (2009) Null model analyses of presence–absence matrices need a definition of independence. Popul Ecol 51:505–512
Nichols JD, Hines JE (2002) Approaches for the direct estimation of λ, and demographic contributions to λ, using capture-recapture data. J Appl Stat 29:539–568
Peery MZ, Becker BH, Beissinger SR (2006) Combining demographic and count-based approaches to identify source–sink dynamics of a threatened seabird. Ecol Appl 16:1516–1528
Pledger S, Pollock KH, Norris JL (2003) Open capture-recapture models with heterogeneity: I Cormack-Jolly-Seber model. Biometrics 59:786–794
Schindler DW (1998) Replication versus realism: the need for ecosystem-scale experiments. Ecosystems 1:323–334
Schwarz CJ, Seber GAF (1999) Estimating animal abundance: review III. Stat Sci 14:427–456
Taylor LR (1961) Aggregation, variance and the mean. Nature 189:732–735
Thompson SK, Seber GAF (1996) Adaptive sampling. Wiley, New York
Williams BK, Conroy MJ, Nichols JD (2002) Analysis and management of animal populations. Academic, San Diego, CA
Wright JA, Barker RJ, Schofield MR, Frantz AC, Byrom AE, Gleeson DM (2009) Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. Biometrics 65:833–840

Statistical Estimation of Actuarial Risk Measures for Heavy-Tailed Claim Amounts

Abdelhakim Necir
Professor
Mohamed Khider University of Biskra, Biskra, Algeria

Introduction

Risk measures are used to quantify insurance losses and measure financial risk assessments. Several risk measures have been proposed in actuarial science literature, namely, the value at risk, the expected shortfall or the conditional tail expectation, and the distorted risk measures (DRM). Let X be a nonnegative random variable (rv) representing losses of an insurance company with a continuous distribution function (df) F. The DRM of X is defined by

$$\Pi_g = \int_0^\infty g\left(1 - F\left(x\right)\right) dx,$$

where the distortion function g is an increasing function such that $g\left(0\right) = 0$ and $g\left(1\right) = 1$ (see, Wang 1996). In terms of the generalized inverse function $Q(s) := \inf\left\{x : F(x) \geq s\right\}$, the DRM may rewritten as

$$\Pi_g = \int_0^1 g'\left(s\right) Q\left(1 - s\right) ds,$$

provided that g is differentiable. In this entry, we consider the DRM corresponding to the distortion function $g\left(s\right) = s^{1/\rho}$, $\rho \geq 1$ called the proportional hazard transform (PHT) risk measure. In this case we write

$$\Pi_\rho = \rho^{-1} \int_0^1 s^{1/\rho - 1} Q\left(1 - s\right) ds.$$

Empirical Estimation of Π_ρ

Suppose we have independent random variables $X_1, X_2, \ldots,$ each with the cdf F, and let $X_{1:n} < \ldots < X_{n:n}$ be the ▶order statistics corresponding to $X_1, \ldots, X_n$. It is most natural to define an empirical estimator of Π_ρ as follows

$$\widehat{\Pi}_\rho := \rho^{-1} \int_0^1 s^{1/\rho - 1} Q_n\left(1 - s\right) ds, \ \rho \geq 1, \tag{1}$$

where $Q_n(s)$ is the empirical quantile function, which is equal to the ith order statistic $X_{i:n}$ when $s \in \left((i-1)/n, i/n\right]$, $i = 1, \ldots, n$. We note that $\widehat{\Pi}_\rho$ is a linear combinations order statistics, that is, $\widehat{\Pi}_\rho = \sum_{i=1}^n a_{i,n} X_{n-i+1,n}$, with $a_{i,n} := \rho^{-1} \int_{(i-1)/n}^{i/n} s^{1/\rho - 1} ds$, $i = 1, \ldots, n$, and $n \in \mathbb{N}$. A statistic having the form (1) is an L-statistic (see, for instance, Shorack and Wellner 1986, p. 260). The ▶asymptotic normality of the estimator $\widehat{\Pi}_\rho$ is discussed in Jones and Zitikis (2003).

Theorem 1 (Jones and Zitikis, 2003). *For any* $1 < \rho < 2$, *we have*

$$n^{1/2}\left(\widehat{\Pi}_\rho - \Pi_\rho\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_\rho^2\right), \ as \ n \to \infty,$$

where

$$\sigma_\rho^2 := \rho^{-2} \int_0^1 \int_0^1 \left(min(s,t) - st\right) s^{1/\rho - 1} t^{1/\rho - 1} dQ\left(1 - s\right) dQ\left(1 - t\right),$$

provided that $\mathbb{E}\left[X^\eta\right] < \infty$ *for some* $\eta > 2\rho/\left(2 - \rho\right)$.

The premium, which is greater than or equal to the mean risk, must be finite for any $\rho \geq 1$. That is, we have $1 \leq \rho < 1/\gamma$. For $\gamma > 1/2$, the second-order moment $\mathbb{E}\left[X^2\right]$ is infinite and $1 \leq \rho < 2$. In this case, we have $2\rho/\left(2 - \rho\right) > 2$ that implies that $\mathbb{E}\left[|X|^\eta\right]$ is infinite for any $\eta > 2\rho/\left(2 - \rho\right)$. Therefore, Theorem 1 does not hold for regularly varying

distributions with tail indices $-1/\gamma > -1/2$. To solve this problem, we propose an alternative estimator for Π_ρ with normal asymptotic distribution for any $-1/\gamma > -1/2$. To get into a more general setting, assume that F is heavy-tailed, which means that $\lim_{x\to\infty} e^{\lambda x}(1-F(x)) = \infty$ for every $\lambda > 0$. The class of regularly varying cdfs is a good example for heavy-tailed models: The cdf F is said to be regularly varying at infinity with index $(-1/\gamma) < 0$ if the condition

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-1/\gamma}, \tag{2}$$

is satisfied for every $x > 0$. This class includes a number of popular distributions such as Pareto, Generalized Pareto, Burr, Fréchet, Student, …, which are known to be appropriate models for fitting large insurance claims, large fluctuations of prices, log-returns, etc. (see, e.g., Beirlant et al. 2001). In the remainder of this entry, we therefore restrict ourselves to this class of distributions, and for more information on them we refer to, for example, de Haan and Ferreira (2006).

New Estimator for Π_ρ : Extreme Values Based Estimation

We have already noted that the estimator $\widehat{\Pi}_\rho$ does not yield asymptotic normality beyond the condition $\mathbf{E}[X^2] < \infty$. For this reason, Necir and Meraghni (2009) proposed an alternative of PHT estimator, which would take into account differences between moderate and high quantiles, that is

$$\widetilde{\Pi}_\rho := \sum_{i=k+1}^{n} a_{i,n} X_{n-i+1,n} + (k/n)^{1/\rho} \frac{X_{n-k,n}}{1-\rho\widehat{\gamma}_n},$$

where we assume that the tail index $\gamma \in [1/2, 1)$ and estimate it using the Hill (1975) estimator $\widehat{\gamma}_n := k^{-1}\sum_{i=1}^{k} \log X_{n-i+1:n} - \log X_{n-k:n}$. Here, let $k = k_n$ be a sequence such that $k \to \infty$, and $k/n \to 0$ as $n \to \infty$. The construction of this estimator is inspired from the work of Necir et al. (2007) and Necir and Boukhetala (2004).

Asymptotic Normality of $\widetilde{\Pi}_\rho$

The main theoretical result of this entry is Theorem 2, below, in which we establish weak approximations for $\widetilde{\Pi}_\rho$ by functional of Brownian bridges and therefore asymptotic confidence bounds for Π_ρ. To formulate it, we need to introduce an assumption that ensures the weak approximation of Hill's estimator $\widehat{\gamma}_n$. The assumption is equivalent to the following second-order condition (see Geluk et al. 1997). Namely, it said that the cdf F satisfies the generalized second-order regular variation condition with second-order parameter $\beta \le 0$ (see de Haan and Stadtmüller 1996) if there exists a function $a(s)$, which does not change its sign in a neighborhood of infinity and is such that, for every $x > 0$,

$$\lim_{s\to\infty} (a(s))^{-1} \left\{ \frac{1-F(sx)}{1-F(s)} - x^{-1/\gamma} \right\} = x^{-1/\gamma} \frac{x^{\rho/\gamma} - 1}{\rho/\gamma}, \tag{3}$$

where $\rho \le 0$ is the so-called second-order parameter; when $\rho = 0$, then the ratio on the right-hand side of Eq. (3) should be interpreted as $\log x$. In the formulation of Theorem 2, we shall use $A(z) := \gamma^2 a(\mathbb{U}(z))$ with $a(s)$ as above and $\mathbb{U}(z) := Q(1-1/z)$.

Theorem 2 (Necir and Meraghni 2009). *Let F be a df satisfying (2) with $\gamma > 1/2$ and suppose that $Q(\cdot)$ is continuously differentiable on $[0,1)$. Let $k = k_n$ be such that $k \to \infty$, $k/n \to 0$ and $k^{1/2}A(n/k) \to 0$ as $n \to \infty$. For any $1 \le \rho < 1/\gamma$, there exists a sequence of independent Brownian bridges (B_n) such that*

$$\frac{n^{1/2}\left(\widetilde{\Pi}_\rho - \Pi_\rho\right)}{(k/n)^{1/\rho-1/2} Q(1-k/n)} =_d \mathcal{L}_1(B_n, \rho, \gamma) + o_p(1),$$

where

$$\mathcal{L}_1(B_n,\rho,\gamma) := \delta(\rho,\gamma)(n/k)^{1/2} B_n(1-k/n) - \lambda_{\rho,\gamma}(n/k)^{1/2}\int_{1-k/n}^{1} \frac{B_n(s)}{1-s} ds - \frac{\rho^{-1}\int_{k/n}^{1} s^{1/\rho-1} B_n(1-s) Q'(1-s)\, ds}{(k/n)^{1/\rho-1/2} Q(1-k/n)},$$

with $\delta(\rho,\gamma) := \lambda_{\rho,\gamma}\left(\rho\gamma^2 - \gamma + 1 - \gamma\lambda_{\rho,\gamma}^{-1}\right)$, and $\lambda_{\rho,\gamma} := \frac{\rho\gamma}{(1-\rho\gamma)^2}$.

corollary 1 Under the assumptions of Theorem 2, we have

$$\frac{n^{1/2}\left(\widetilde{\Pi}_{\rho,n} - \Pi_\rho\right)}{(k/n)^{1/\rho-1/2} X_{n-k:n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \sigma_{\rho,\gamma}^2\right), \text{ as } n \to \infty,$$

where the asymptotic variance $\sigma_{\rho,\gamma}^2$ is given by the sum of the following terms

$$\kappa_1 = \frac{\left(\gamma\rho - \gamma + \gamma^2\rho\right)^2}{(1-\rho\gamma)^4}, \kappa_2 = \frac{2\rho^2\gamma^2}{(1-\rho\gamma)^4}$$

$$\kappa_3 = \frac{2\gamma^2}{(1-\rho-\rho\gamma)(2-\rho-2\rho\gamma)}, \kappa_4 = \frac{2\rho\gamma\left(\gamma-\gamma\rho-\gamma^2\rho\right)}{(1-\rho\gamma)^4}$$

and $\kappa_5 = -\frac{2\rho\gamma^3}{(1-\rho\gamma)^2}$.

Cross References

►Actuarial Methods
►Asymptotic Normality

►Estimation: An Overview
►Heavy-Tailed Distributions
►Insurance, Statistics in
►Risk Analysis

References and Further Reading

Artzner Ph, Delbaen F, Eber J-M, Heath D (1999) Coherent measures of risk. Math Financ 9:203–228
Beirlant J, Matthys G, Dierckx G (2001) Heavy-tailed distributions and rating. Astin Bull 31:37–58
de Haan L, Ferreira A (2006) Extreme value theory: an introduction. Springer, New York
de Haan L, Stadtmüller U (1996) Generalized regular variation of second order. J Aust Math Soc A 61:381–395
Geluk J, de Haan L, Resnick S, Starica C (1997) Second order regular variation, convolution and the central limit theorem. Stoch Proc Appl 69:139–135
Hill BM (1975) A simple approach to inference about the tail of a distribution. Ann Stat 3:1136–1174
Jones BL, Zitikis R (2003) Empirical estimation of risk measures and related quantities. N Am Actuarial J 7:44–54
Necir A, Boukhetala K (2004) Estimating the risk adjusted premium of the largest reinsurance covers. In: Antoch J (ed) Proceeding of computational statistics, Physica-Verlag, pp 1577–1584. http://www.springer.com/statistics/computanional+statistics/book/978-3-7908-1554-2
Necir A, Meraghni D (2009) Empirical estimation of the proportional hazard premium for heavy-tailed claim amounts. Insur Math Econ 45:49–58
Necir A, Meraghni D, Meddi F (2007) Statistical estimate of the proportional hazard premium of loss. Scand Actuarial J 3:147–161
Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley, New York
Wang SS (1996) Premium calculation by transforming the layer premium density. Astin Bull 26:71–92

Statistical Evidence

Subhash R. Lele[1], Mark L. Taper[2]
[1]Professor
University of Alberta, Edmonton, AB, Canada
[2]Research Scientist
Montana State University, Bozeman, MT, USA

Scientists want to know how nature works. Different scientists have different ideas or hypotheses about the mechanisms that underlie a phenomenon. To test the validity of these ideas about mechanisms, they need to be translated into quantitative form in a mathematical model that is capable of predicting the possible outcomes from such mechanisms. Observations of real outcomes, whether obtained by designed experiment or observational study, are used to help discriminate between different mechanisms. The classical approach of hypothesis refutation depends on showing that the data are impossible under a specific hypothesis. However, because of the intrinsic stochasticity in nature, appropriate mathematical models tend to be statistical rather than deterministic. No data are impossible under a statistical model and hence this classic approach cannot be used to falsify a statistical model. On the other hand, although not impossible, data could be more improbable under one statistical model than a competing one. Quantifying evidence for one statistical model vis-à-vis a competing one is one of the major tasks of statistics. The evidential paradigm in statistics addresses the fundamental question: How should we interpret the observed data as evidence for one hypothesis over the other? Various researchers have tried to formulate ways of quantifying evidence, most notably Barnard (1949) and Edwards (1992). The monograph by Hacking (Hacking 1965) explicitly stated the problem and its solution in terms of the law of the likelihood:

► *If hypothesis A implies that the probability that a random variable X takes the value x is $p_A(x)$, while hypothesis B implies that the probability is $p_B(x)$, then the observation $X = x$ is evidence supporting A over B if and only if $p_A(x) > p_B(x)$ and the likelihood ratio $p_A(x) > p_B(x)$, measures the strength of that evidence.*

Royall (1997) developed this simple yet powerful idea and turned it into something that is applicable in practice. He emphasized that the commonly used approaches in statistics are either decision-theoretic (Neyman-Pearson-Wald) that address the question "given these data, what should I do?" or, are belief based (Bayesian) that address the question "given these data, how do I change my beliefs about the two hypotheses?" He suggested that statisticians should first address the more fundamental question "how should we interpret the observed data as evidence for one hypothesis over the other?", and only then think about how the beliefs should be changed or decisions should be made in the light of this evidence. Royall also pointed out that evidence is a strictly comparative concept. We need two competing hypotheses before we can compare the evidence for one over the other. His critique of the commonly used evidence measures showed that the practice of using Fisherian p-value as a measure of evidence is incorrect because it is not a comparative measure, while the Bayesian posterior probability, aside from being dependent on the prior beliefs and not solely on the observed data, is also an incorrect measure of evidence because it is not invariant to the choice of the parameterization.

One of the reasons, the Neyman-Pearson ideas are prominent in science is that they accept the fact that decisions can go wrong. Hence in scientific practice, one quantifies and controls the probabilities of such wrong decisions. Royall (1997) introduced concepts of error probabilities that are similar to the Type-I and Type-II error probabilities in the Neyman-Pearson formulation, but relevant to the evidential paradigm. He realized, evidence, properly interpreted, can be misleading and asked how often would be we be misled by strong evidence (see below) if we use the law of the likelihood and how often would we be in a situation that neither hypothesis is supported to the threshold of strong evidence.

Three concepts answer those questions. Suppose we say that hypothesis A has strong evidence supporting it over hypothesis B if the likelihood ratio is greater than K, for some a priori fixed $K > 0$. Then:

(a) The probability of misleading strong evidence: $M(K) = P_A\left(x : \frac{p_B(x)}{p_A(x)} > K\right)$,
(b) The probability of weak evidence: $W(K) = P_A\left(x : \frac{1}{K} < \frac{p_B(x)}{p_A(x)} < K\right)$,
(c) The probability of strong evidence for the correct model: $S(K) = P_A\left(x : \frac{p_A(x)}{p_B(x)} > K\right)$.

A remarkable result that follows is that there exists a universal upper bound on the probability of misleading evidence under any model, namely $M(K) \leq 1/K$. Furthermore, as one increases the sample size, both $M(K)$ and $W(K)$ converge to 0 and $S(K) \to 1$. Thus, with enough observations we are sure to reach the right conclusion without any error. This is in stark contrast with the Neyman-Pearson Type-I error that remains fixed, no matter how large the sample size. In the Neyman-Pearson formulation, as sample size increases, K increases while error probability is held constant. Thus, as one increases the sample size, the criterion for rejection changes so that it is harder and harder to distinguish the hypotheses. This seems quite counter-intuitive and makes it difficult to compare tests of different sample size.

The concepts of misleading and weak evidence have implications in the sample size calculations and optimal experimental designs. For example, the experimenter should make sure the minimal sample size is such that probability of weak evidence is quite small and at the end of the experiment one can reach a conclusion. Furthermore, by controlling the probability of misleading evidence through sample size, experimental/sampling design and evidence threshold one can also make sure that the conclusions reached are likely to be correct. Besides these *a priori* uses, the probability of misleading evidence can be calculated as a post data error statistic reminiscent of a p-value, but explicitly constructed for the comparison of two hypotheses (Taper and Lele 2010).

There are, however, limitations to the evidential ideas developed by Royall and described above. One major limitation is that the law of likelihood can only quantify evidence when the hypotheses are simple, but most scientific problems involve comparing composite hypotheses. This may arise because the scientist may be interested in testing only some feature of the model without restrictions on the rest of the features. Similarly, a proper statistical model might involve infinitely many nuisance parameters in order to model the underlying mechanism realistically but the parameters of interest may be finite. Such cases arise in many practically situations, for example, the longitudinal data analysis or random effects models among others. Aside from raising the need to consider composite hypothesis, in these situations, the full likelihood function may be difficult to write down. One may want to specify only a few features of the model such as the mean or the variance, leading to the use of quasi-likelihood, estimating functions and such other modifications. The question of ►model selection where one is selecting between families of models instead of a specific element of a given family is important in scientific practice. For example, whether to use a linear regression model (see ►Linear Regression Models) or a non-linear regression model (see ►Nonlinear Regression) is critical for forecasting.

Can we generalize the law of likelihood and concepts of error probabilities to make it applicable in such situations? An initial attempt is described in Lele (2004), Taper and Lele (2004, 2010). The key observation in such a generalization is that quantifying the strength of evidence is the same as comparing distances between the truth and the competing models that are estimated from data. The likelihood ratio simply compares an estimate of the ►Kullback-Leibler divergence.

One can consider many different kinds of divergences, each leading to different desirable properties. For example, if one uses Hellinger distance to quantify strength of evidence, one gets a measure that is robust against ►outliers. If one uses Jeffrey's divergence, one needs to specify only the mean and variance function, similar to the quasi-likelihood formulation, to quantify strength of evidence. One can use profile likelihood or integrated likelihood or conditional likelihood to compare evidence about a parameter of interest in the presence of nuisance parameters. These simply correspond to different divergence measures and hence have different properties. Lele (2004) terms these as "evidence functions". They may be

compared in terms of how fast the probability of strong evidence for the correct model converges to 1. Not surprisingly, for simple versus simple hypothesis comparison, it turns out that the Kullback-Leibler divergence or the likelihood ratio is the best evidence function, provided the model is correctly specified. Other evidence functions, however, might be more robust against outliers or may need less specification; and hence may be more desirable in practice.

Error probabilities can be calculated for general evidence functions using bootstrapping (Taper and Lele 2010). When the data are independent and identically distributed one can circumvent the conceptual constraint that the true model is in one of the alternative hypotheses by using a non-parametric bootstrap. We briefly describe this in the likelihood ratio context. Notice that the likelihood ratio is simply a statistic, a function of the data. One can generate a ►simple random sample with replacement from the original data and compute the strength of evidence based on this new sample. By repeating this procedure large number of times, one obtains the bootstrap estimate of the distribution of the strength of evidence. The percentile-based confidence interval tells us the smallest level of strength of evidence one is likely to obtain if the experiment is repeated. One of the vexing questions in evidential paradigm is how to relate evidence to decision making without invoking beliefs. It may be possible to use the bootstrap distribution of the strength of evidence, in conjunction with the ►loss function, for decision-making. Because this distribution is obtained empirically from the observations, such decisions will be robust against model specifications.

The evidential paradigm is still in its adolescence, with much scope for innovation. Nevertheless the paradigm is sufficiently developed to make immediate contributions; in fact, information criterion comparisons, which are evidence functions, have already revolutionized the practice of many sciences. The references below will be useful to further widen the reader's knowledge and understanding beyond just our views.

About the Authors

Dr. Subhash Lele is a professor of statistics in the Department of Mathematical and Statistical Sciences, University of Alberta, Canada. He has published over 70 papers in statistical and scientific journals on various topics such as morphometrics, quantitative ecology, hierarchical models, estimating functions and philosophy of statistics. He has served as the President of the Biostatistics section of the Statistical Society of Canada and Secretary of the Statistical Ecology section of the Ecological Society of America. He is an Elected member of the International Statistical Institute. He has served on the editorial boards of the Journal of the American Statistical Association, Ecology and Ecological Monographs and Ecological and Environmental Statistics. He has co-authored (with Dr. J.T. Richtsmeier) a book *An invariant approach to statistical analysis of shapes* (Chapman and Hall, 2000) and co-edited a book (with Dr. M.L. Taper) on *The nature of scientific evidence: Empirical, philosophical and statistical considerations* (University of Chicago press, 2004). He has served on three U.S. National Academy of Sciences committees on climate change, public health and other issues.

Dr. Mark L. Taper (Department of Ecology, Montana State University, USA) is a statistical and quantitative ecologist who uses analytic and computational modeling to answer questions in conservation biology, spatial ecology, population dynamics, macro ecology, and evolutionary ecology. He also has a deep interest in the epistemological foundations of both statistics and science. Dr. Taper has chaired the Ecological Society of America's statistics section and is the founding Director of the Montana State University interdisciplinary program in Ecological and Environmental Statistics. Dr. Taper has published over 90 scientific, statistical, and philosophical articles and co-edited with Subhash Lele a volume titled *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations* published by The University of Chicago Press in 2004. He has served on the editorial boards of *Frontiers in Ecology and the Environment* and of *Ecological and Environmental Statistics*.

Cross References

►Bootstrap Methods
►Frequentist Hypothesis Testing: A Defense
►Kullback-Leibler Divergence
►Likelihood
►Marginal Probability: Its Use in Bayesian Statistics as Model Evidence
►Model Selection
►Most Powerful Test
►Neyman-Pearson Lemma
►Presentation of Statistical Testimony
►P-Values
►Sample Size Determination
►Significance Testing: An Overview
►Significance Tests: A Critique
►Statistical Fallacies

►Statistical Inference: An Overview
►Statistics and the Law

References and Further Reading

Research Papers

Barnard GA (1949) Statistical Inference. J Roy Stat Soc B 11:115–149

Blume JD (2002) Likelihood methods for measuring statistical evidence. Stat Med 21:2563–2599

Lele SR (2004) Evidence functions and the optimality of the law of likelihood. In: Taper ML, Lele SR (eds) The nature of scientific evidence: statistical, philosophical and empirical considerations. University of Chicago Press, Chicago

Strug LJ, Hodge SE (2006a) An alternative foundation for the planning and valuation of linkage analysis I. Decoupling "error probabilities" from "measures of evidence". Hum Hered 61:166–188

Strug LJ, Hodge SE (2006b) An alternative foundation for the planning and evaluation of linkage analysis II. Implications for multiple test adjustments. Hum Hered 61:200–209

Strug LJ, Rohde C, Corey PN (2007) An introduction to evidential sample size. Am Stat 61(3):1–5

Taper ML, Lele SR (2010) Evidence, evidence functions and error probabilities. In: Forster MR, Bandyophadhyay PS (eds) Handbook for philosophy of statistics. Elsevier

Books

Edwards AWF (1992) Likelihood, Expanded edn. Johns Hopkins University Press, Baltimore

Forster MR, Bandopadhyay PS (2010 expected) Handbook for philosophy of statistics. Elsevier

Hacking I (1965) Logic of statistical inference. Cambridge University Press, Cambridge

Mayo DG (1996) Error and the Growth of Experimental Knowledge. University of Chicago Press, Chicago

Royall R (1997) Statistical Evidence: a likelihood paradigm. Chapman & Hall, London

Taper ML, Lele SR (eds) (2004) The nature of scientific evidence: statistical, philosophical and empirical considerations. University of Chicago Press, Chicago

Taper ML, Lele SR (2010) Evidence, evidence functions and error probabilities. In: Forster MR, Bandyophadhyay PS (eds) Handbook for philosophy of statistics. Elsevier

Statistical Fallacies

Watter Krämer
Professor and Chairman
Technische Universität Dortmund, Dortmund, Germany

The range of possible fallacies in statistics is as wide as the range of statistics itself (see Cohen 1938; Good 1962, 1978; Moran 1973 for convenient overviews); there is probably no application and no theory where one does not find examples of intentional or unintentional misuse of statistical facts and theories (which of course is not unique to statistics – there is probably no science or social science whatsoever which is immune to such abuse). When collecting data, there is the well known problem of biased or self-selected samples, or ill-phrased questionnaires where answers are already imbedded in the questions. A nice example is provided by two surveys on workers' attitude towards working on Saturdays which were conducted in Germany in the same months of the same year (Krämer 2008, p. 121). The first survey produced a rejection rate of 95% whereas in the second survey, 80% of workers who were asked were happy to work on Saturdays if only they could. After inspection of the questionnaires it was clear how these results came about: The first survey was sponsored by a trade union and started with reminding the audience of the hard work it had taken to push through the five day work week, ending with the question (I exaggerate slightly): Are you really prepared to sacrifice all of what your fellow workers have fought about so hard? The second survey started with a comment on fierce competition for German industry from Asia which in the end let to the final question of whether workers were prepared to work on Saturdays if otherwise their employer went bankrupt.

Such extreme examples are of course quite rare, but it is rather easy to lead people in any direction which is convenient from the researcher's point of view.

In the area of biased and self-selected samples, the best known example is of course the historical disaster of the Literary Digest magazine back in 1936. The magazine had asked well above ten million Americans, a record sample by any standards, whom they intended to vote for in the upcoming presidential election. According to this survey, the republican candidate was going to win handsomely whereas in reality Roosevelt, the incumbent, won by a landslide. The Digest's sample was drawn from lists of automobile and telephone owners (likely to vote republican) and among those asked, less than a quarter actually replied (presumably voters with an axe to grind with the incumbent; see Bryson 1976).

Other fallacies arise in the context of interpreting or presenting the results of statistical analyses. There is the obvious area of confusing correlation and causation or of misreading the meaning of statistical tests of significance, where even professional statisticians have a hard time to correctly interpret a positive test result at – say – a 5% level of significance (there are even textbooks which state that this means: "The null hypothesis is wrong with 95% probability"). Another problem here is that true significance levels are in many applications much higher then nominal ones due to the fact that only "significant" outcomes are reported.

Such problems with interpreting statistical tests are tightly connected with the misuse of conditional probabilities, which is probably the both most widespread and most dangerous way that one can misread statistical evidence (Krämer and Gigerenzer 2005). One of these is to infer, from a conditional probability $P(A|B)$ that is seen as "large," that the conditional event A is "favorable" to the conditioning event B, in the sense that $P(B|A) > P(B)$.

This confusion occurs in various contexts and is possibly the most frequent logical error that is made in the interpretation of statistical information. Here are some examples from the German press (with the headlines translated into English):

- "Beware of German tourists" (According to *Der Spiegel* magazine, most skiers involved in accidents in a Swiss skiing resort came from Germany).
- "Boys more at risk on bicycles" (the newspaper *Hannoversche Allgemeine Zeitung* reported that among children involved in bicycle accidents, the majority were boys).
- "Soccer most dangerous sport" (the weekly magazine *Stern* commenting on a survey of accidents in sports).
- "Private homes as danger spots" (the newspaper *Die Welt* musing about the fact that a third of all fatal accidents in Germany occur in private homes).
- "German shepherd most dangerous dog around" (The newspaper *Ruhr-Nachrichten* on a statistic according to which German shepherds account for a record 31% of all reported attacks by dogs).
- "Women more disoriented drivers" (The newspaper *Bild* commenting on the fact that among cars that were found entering a one-way street in the wrong direction, most were driven by women).

These examples can easily be extended. Most of them result from unintentionally misreading the statistical evidence. When there are cherished stereotypes to conserve, such as the German tourist bullying his fellow vacationers, or women somehow lost in space, perhaps some intentional neglect of logic may have played a role as well. Also, not all of the above statements are necessarily false. It might, for instance, well be true that when 1,000 men and 1,000 women drivers are given a chance to enter a one-way street the wrong way, more women than men will actually do so, but the survey by *Bild* simply counted wrongly entering cars and this is certainly no proof of their claim. For example, what if there were no men on the street at that time of the day? And in the case of the Swiss skiing resort, where almost all foreign tourists came from Germany, the attribution of abnormally dangerous behavior to this class of visitors is clearly wrong.

In terms of favorable events, *Der Spiegel*, on observing that $P(\text{German tourist} \mid \text{skiing accident})$ was "large," concluded that the reverse conditional probability was also large, in particular, that being a German tourist increases the chances of being involved in a skiing accident:

$$P(\text{skiing accident}|\text{German tourist}) > P(\text{skiing accident}).$$

Similarly, *Hannoversche Allgemeine Zeitung* concluded from $P(\text{boy} \mid \text{bicycle accident}) = \text{large}$ that $P(\text{bicycle accident} \mid \text{boy}) > P(\text{bicycle accident})$ and so on. In all these examples, the point of departure was always a large value of $P(A|B)$, which then led to the – possibly unwarranted – conclusion that $P(B|A) > P(B)$. From the symmetry

$$P(B|A) > P(B) \Longleftrightarrow P(A|B) > P(A)$$

it is clear, however, that one cannot infer anything regarding A's favorableness for B from $P(A|B)$ alone, and that one needs information on $P(A)$ as well.

Another avenue through which the attribute of favorableness can be incorrectly attached to certain events is ▶Simpson's paradox, which in our context asserts that it is possible that B is favorable to A when C holds, B is also favorable to A when C does not hold, yet overall, B is unfavorable to A. Formally, one has

$$\begin{aligned} &P(A|B \cap C) > P(A) \qquad \text{and} \\ &P(A|B \cap \overline{C}) > P(A) \qquad \text{yet} \\ &P(A|B) < P(A). \end{aligned}$$

This paradox also extends to situations where $C_1 \cup \ldots \cup C_n = \Omega$, $C_i \cap C = \emptyset\,(i \neq j)$.

One instance where Simpson's paradox (to be precise: the refusal to take account of Simpson's paradox) has been deliberately used to mislead the public is the debate on the causes of cancer in Germany. The official and fiercely defended credo of the Green movement has it that the increase in cancer deaths from well below 20% of all deaths after the war to almost 30% today, is mostly due to industrial pollution and chemical waste of all sorts. However, as Table 1 shows, among women, the probability of dying from cancer has actually *decreased* for young and old alike! Similar results hold for men.

A final and more trivial example for faulty inferences from conditional probabilities concerns the inequality

$$P(A|B \cap D) > P(A|C \cap D).$$

Plainly, this does not imply

$$P(A|B) > P(A|C),$$

yet this conclusion is still sometimes drawn. A German newspaper once claimed that people get happier as they

Statistical Fallacies. Table 1 Probability of dying from cancer Number of women (among 100,000 in the respective age groups) who died from cancer in Germany

Age	1970	2001
0–4	7	3
5–9	6	2
10–14	4	2
15–19	6	2
20–24	8	4
25–29	12	6
30–34	21	13
35–39	45	25
40–44	84	51
45–49	144	98
50–54	214	161
55–59	305	240
60–64	415	321
65–69	601	468
70–74	850	656
75–79	1183	924
80–84	1644	1587

(Statistisches Jahrbuch für die Bundesrepublik Deutschland)

grow older. The paper's "proof" runs as follows: Among people who die at age 20–25, about 25% commit suicide. This percentage then decreases with advancing age; thus, for instance, among people who die over the age of 70, only 2% commit suicide. Formally, one can put these observations as

$$P(\text{suicide} \mid \text{age } 20-25 \text{ and death}) > P(\text{suicide} \mid \text{age} > 70 \text{ and death}),$$

and while this is true, it certainly does not imply

$$P(\text{suicide} \mid \text{age } 20-25) > P(\text{suicide} \mid \text{age} > 70).$$

In fact, a glance at any statistical almanac shows that quite the opposite is true.

Here is a more recent example from the US, where likewise $P(A|B)$ is confused with $P(A|B \cap D)$. This time the confusion is spread by renowned Harvard Law professor who advised the O. J. Simpson defense team. The prosecution had argued that Simpson's history of spousal abuse reflects a motive to kill, advancing the premise that "a slap is a prelude to homicide." The defence – in the end successfully – argued that the probability of the event K that a husband killed his wife if he battered her was rather small, so battering showed not be viewed as evidence of murder.

$$P(K \mid \text{battered}) = 1/2,500.$$

The relevant probability, however, is not this one. It is that of a man murdering his partner given that he battered her *and* that she was murdered:

$$P(K \mid \text{battered and murdered}).$$

This probability is about 8/9 (Good 1996). It must not of course be confused with the probability that O. J. Simpson is guilty. But it shows that battering is a fairly good predictor of guilt for murder.

About the Author

Dr. Walter Krämer is a Professor and Chairman of Department of Statistics, University of Dortmund. He is Editor of *Statistical Papers* and *German Economic Review*. He is a member of ISI and NRW Akademie der Wissenschaften. Professor Krämer is author of more than 100 articles and 30 books. His book *So lügt man mit Statistik* (in German), modelled after Durrel Huffs classic "How to lie with Statics" has been translated into several languages.

Cross References

►Fraud in Statistics
►Misuse of Statistics
►Questionnaire
►Simpson's Paradox
►Statistical Evidence
►Statistical Fallacies: Misconceptions, and Myths
►Telephone Sampling: Frames and Selection Techniques

References and Further Reading

Bryson MC (1976) The literary digest poll: making of a statistical myth. Am Stat 30:184–185
Cohen JB (1938) The misuse of statistics. J Am Stat Soc 33:657–674
Good IJ (1962) A classification of fallacious arguments and interpretations. Technometrics 4:125–132
Good IJ (1978) Fallacies, statistical. In: Kruskal WH, Tanar JM (eds) International encyclopedia of statistics, vol 1. pp 337–349
Good IJ (1996) When batterer becomes murderer. Nature 381:481
Krämer W (2008) So lügt man mit Statistik, 11th paperback edn. Piper-Verlag, München

Krämer W, Gigerenzer G (2005) How to confuse with statistics: the use and misuse of conditional probabilities. Stat Sci 20:223–230
Moran PAP (1973) Problems and mistakes in statistical analysis. Commun Stat 2:245–257

Statistical Fallacies: Misconceptions, and Myths

Shlomo Sawilowsky
Professor
Wayne State University, Detroit, MI, USA

Compilations and illustrations of statistical fallacies, misconceptions, and myths abound (e.g., Brewer 1985; Huck 2008; Huff 1954; Hunter and May 1993; King 1986; Sawilowsky 1993, 2003a, b, c, d, 2005, 2007a, b; Vandenberg 2006). The statistical faux pas is appealing, intuitive, logical, and persuasive, but demonstrably false. They are uniformly presented based on authority and supported based on assertion. Unfortunately, these errors spontaneously regenerate every few years, propagating in peer reviewed journal articles; popular college textbooks; and most prominently, in the alternate (e.g., qualitative), nonprofessional (e.g., Wikipedia), and dissident literature. Some of the most egregious and grievous are noted below.

1. *Law of Large Numbers, Central Limit Theorem (CLT), population normality, and asymptotic theory*. This quartet is asserted to inform the statistical properties (i.e., Type I and II errors, comparative statistical power) of parametric tests for small samples (e.g., n≤50 or so). In fact, much of what was asserted regarding small samples based on these eighteenth to nineteenth century theorems was wrong. Most of what is correctly known about the properties of parametric statistics has been learned through Monte Carlo studies and related methods conducted in the last quarter of the twentieth century to the present.

   Examples of wrong statements include (a) random selection is mooted by drawing a sufficiently large sample, (b) the CLT guarantees $\overline{X}$ is normally distributed, (c) the CLT safeguards parametric tests as long as n ≥ 30, and (d) asymptotic relative efficiency is a meaningful predictor of small sample power. A corollary that is particularly destructive is journal editor and reviewer bias in favor of this quartet over Monte Carlo evidence, relegating the inelegance of the latter to be a function of "anyone who has a personal computer and knowledge of Algebra I."

   (e) Perhaps the most pervasive myth is that real variables are normally distributed. Micceri (1989) canvassed authors of psychology and education research over a number of years and determined that less than 3% of their data sets (even those where n > 5,000) could be considered even remotely bell-shaped (e.g., symmetric with light tails). Not a single data set was able to pass any known statistical test of normality. Similar studies have been conducted in other disciplines with the same result. Population normality is not the norm.

   (f) Journal editors and reviewers mistakenly attach more importance to lemmas, theorems, and corollaries from this quartet than on evidence from small samples Monte Carlo studies and related methods.
2. *Random assignment*. It is commonly asserted that the lack of random assignment can be rehabilitated via matching, ANCOVA, regression, econometric simultaneous modeling, latent-variable modeling, etc. In truth, "*there is no substitute for randomization*" (Sawilowsky 2007b, p 214.)
3. *Control group*. It is frequently asserted by journal editors and referees, and funding agency reviewers, that science and rigorous experimental design demand the use of a control, comparison, or second treatment group. Actually, there are many designs that do not require this, such as factorial ANOVA, times series, and single subject repeated measures layouts.
4. *Data transformations*. (a) One reason for transforming data is to better meet a parametric test's underlying assumptions. The inexplicable pressure to shoehorn a parametric test into a situation where doesn't fit has prompted textbook authors to recommend transforming data to better meet underlying assumptions. For example, if the data are skewed then the square root transformation is recommended. The debate on the utility of transforming for this purpose is known as the Games-Levine controversy that was waged in the early 1980s, primarily recorded in *Psychological Bulletin*.

   There is a misguided presumption that the statistician has a priori knowledge of when or how best to transform. Also, it is a fallacy to interpret results from a transformation in the original metric. What does it mean to conclude that the arcsin of children's weight in the intervention group was statistically significantly higher than the arcsin of children's weight in the comparison group? When was the last time a patient chal-

lenged the physician's recommended medication by demanding to know the logarithm of the expected reduction in weight as predicted from the clinical trial?

(b) Another reason for transforming the data is to convert a parametric procedure into a nonparametric procedure. The rank transformation is the prime example. Based on asymptotic theory published in very prestigious journals, and subsequent recommendations from high profile statistical software companies, data analysts were encouraged to routinely run their data through a ranking procedure, and follow with the standard parametric test on those ranks.

Careful data analysts have shown through Monte Carlo studies that good results may be obtained for the two independent samples, one-way independent ANOVA, and two independent samples multivariate layouts. The myth persists, however, that this procedure is a panacea. Those same careful data analyst have also shown the rank transformation does not work in the context of two dependent samples, factorial ANOVA, factorial ANCOVA, MANOVA, or MANCOVA layouts, yielding Type I error rates as high as 1, and greatly suppressed power (e.g., Sawilowsky 1985a; Sawilowsky et al. 1989; Blair et al. 1987). Yet, software vendors continue to promote this procedure.

(c) It is also a myth that secondary transformations resolve this problem. The original data are transformed into ranks, and the ranks are in turn transformed into expected normal scores, random normal scores, or some other type of score. However, careful data analysts have also shown that secondary transformations fare no better than the rank transformation in terms of displaying poor Type I error control and severely depressed power (Sawilowsky 1985b).

5. *p values*. (a) Significance testing, as opposed to hypothesis testing, is mistakenly asserted to be scientific. Whereas hypothesis testing is objective due to the a priori stated threshold of what constitutes a rare event, significance testing is not objective. With the advent of easily obtained (and even exact) p values through statistical software, significance testing permits citing the resulting p value and letting the reader decide a posteriori if it is significant. Unfortunately, post and ad hoc significance testing obviates objectivity in interpreting the results, which is a fatal violation of a cornerstone of science. (b) Obtained p values are asserted to be transitory. For example, a p value that is close to nominal alpha (e.g., $\alpha = 0.05$ and $p = 0.06$) is incorrectly claimed to be approaching statistical significance, when in fact the result of the experiment is quite stationary. (c) The magnitude of the p value is asserted to inform the magnitude of the treatment effect. A p value of 0.0001 is erroneously claimed to mean the effect is of great practical importance. Although that may be true, it is not because of any evidence based on the magnitude of p.
6. *Effect Size*. Statistical philosophers stipulate that the null hypothesis can never literally be true. By virtue of all phenomena existing in a closed universe, at some part of the mantissa the population values must diverge from zero. Thus, it is claimed that effect sizes should be reported even if a hypothesis test was not conducted, or even if the result of a hypothesis test is not statistically significant.

   This viewpoint is presaged on an imputed meta-analytic intent that will arise in the future even if there is no such intent at the time the experiment was conducted. This fallacy arises, as do many errors in interpretation of statistics, by ignoring the null hypothesis being tested. Under the truth of the null hypothesis observed results for the sample are not statistically significantly different from zero, and thus the magnitude of the observed result is meaningless. Hence, effect sizes are only meaningfully reported in conjunction with a statistically significant hypothesis test.
7. *Experiment-wise Type I error*. It is universally recommended that prudent statisticians should conduct preliminary tests of underlying assumptions (e.g., homoscedasticity, normality) prior to testing for effects. It is asserted that this does no harm to the experiment-wise Type I error rate. However, Monte Carlo evidence demonstrates that the experiment-wise Type I error rate will inflate if preliminary tests are conducted without statistical adjustment for multiple testing. Moreover, there will be a Type I inflation even if the decision to proceed is based on eye-balling the data.
8. *Confidence Intervals*. Confidence intervals have recently been promoted over the use of hypothesis tests for a litany of unsupported reasons. (a) Among its supposed benefits is the assertion that confidence intervals provide more confidence than do hypothesis tests. This is based on the fallacy that confidence intervals are based on some system of probability theory other than that of hypothesis tests, when in fact they are the same. (b) Another prevalent misconception is confidence intervals must be symmetric.
9. ►*Robust statistics*. Typically, proposed expansions of descriptive robust statistics into inferential procedures are substantiated via comparisons with para-

metric methods. It is rare to find direct comparisons of inferential robust statistics with nonparametric procedures. (a) It is asserted that robust descriptive statistics maintain their robustness when evolved into inferential counterparts. This is a fallacy, however, because robust descriptive statistics were derived under parametric models confronted with perturbations. Therefore, Monte Carlo studies show they exhibit inflated Type I errors in many layouts. (b) It is similarly asserted that robust inferential statistics are high in comparative statistical power, but they are generally less powerful than rank based nonparametric methods when testing hypotheses for which the latter are intended.

10. ►*Permutation tests*. Permutation analogs to parametric tests are correctly stated to have equal power, and indeed can rehabilitate parametric tests' poor Type I error properties. However, it is incorrectly asserted that they are more powerful than nonparametric methods when testing for shift in location, when in fact the power spectrum of permutation tests generally follows (albeit somewhat higher) the power spectrum of their parametric counterparts, which is considerably less powerful than nonparametric procedures.
11. *Exact statistics*. Exact statistics, recently prevalent due to the advent of statistical software, are often advertised by software venders as being the most powerful procedure available to the statistician for the analysis of small samples. Actually, the advantage of exact statistics is that the p values are correct, but as often as not a smaller p value will result from the use of tabled asymptotic p values.
12. *Parametric tests*. The t and F tests are asserted to be (a) completely robust to Type I errors with respect to departures from population normality, (b) generally robust with respect to departures from population homoscedasticity, and (c) at least somewhat robust with respect to departures from independence. All three of these assertions are patently false. (d) Parametric tests are incorrectly asserted to trump the need for random selection or assignment of data, particularly due to Sir Ronald Fisher's paradigm of analysis on the data at hand.

    (e) Parametric tests (e.g., t, F) are asserted to be more powerful than nonparametric tests (e.g., Wilcoxon Rank Sum (see ►Wilcoxon–Mann–Whitney Test), Wilcoxon Signed Ranks (see ►Wilcoxon-signed-rank test)) when testing for shift in location. In fact, for skewed distributions, the nonparametric tests are often three to four times more powerful than their parametric counterparts. (f) As sample size increases, these parametric tests are asserted to increase their power advantages over nonparametric tests. In fact, the opposite is true until the upper part of the power spectrum is reached (e.g., the ceiling is 1) when the parametric tests eventually converge with the nonparametric test's statistical power.
13. *Nonparametric rank tests*. The assertions denigrating the Wilcoxon tests are so pervasive (to the extent that the two independent samples case is more frequently attributed as the Mann Whitney U, even though Wilcoxon had priority by 2 years) that the reader is referred to Sawilowsky (2005) for a listing of 22 frequently cited fallacies, misconceptions, and myths. Among the highlights are the incorrect beliefs that (a) the uniformly most powerful unbiased moniker follows the usage of the parametric t test for data sampled from nonnormally distributed populations, (b) the Wilcoxon tests should only be used with small data sets, (c) the Wilcoxon tests should only be used with ordinal scaled data, and (d) the Wilcoxon tests' power properties are oblivious to ►outliers.
14. χ^2. (a) We live in a χ^2 society due to political correctness that dictates equality of outcome instead of equality of opportunity. The test of independence version of this statistic is accepted *sans voire dire* by many legal systems as the single most important arbiter of truth, justice, and salvation. It has been asserted that any statistical difference between (often even nonrandomly selected) samples of ethnicity, gender, or other demographic as compared with (often even inaccurate, incomplete, and outdated) census data is *primae faciea* evidence of institutional racism, sexism, or other ism. A plaintiff allegation that is supportable by a significant χ^2 is often accepted by the court (judges and juries) *praesumptio iuris et de iure*. Similarly, the goodness of fit version of this statistic is also placed on an unwarranted pedestal.

    In fact, χ^2 is super powered for any arbitrary large number of observations. For example, in the goodness of fit application where the number of observed data points is very large and the obtained χ^2 can be of an order of magnitude greater than three, there is the custom not to even bother with the divisor E_i, and instead to proclaim a good fit if the new empirical process results in a reduced obtained value of the numerator. The converse is true where the number of observed data points are small (e.g., $N < 20$ or 30), in which case the χ^2 test of independence is among the least powerful methods available in a statistician's repertoire.

15. *Stepwise regression*. Stepwise (or "unwise", Leamer 1985) regression and replicability are two mutually exclusive concepts. It is asserted to be an appropriate data mining technique (see ►Data Mining). However, it is analogous to talking a walk in the dark in the park, tripping over a duffle bag, inspecting the bag and finding data sheets crumpled together, transcribing and entering the data into a statistical software program, having the software command the CPU to regress all possible combinations of independent variables on the dependent variable until the probability to enter has been met, reporting the results, and eyeballing the results to construct an explanation or prediction about an as yet unstated research hypothesis. There is nothing scientifically rigorous about Stepwise regression, even when it is adorned with the appellation of nonmodel-based regression. It is tantamount to a search for Type I errors.
16. *ANOVA main and interaction effects*. (a) It is asserted that because certain transformations can be invoked to make interaction effects apparently vanish, main effects are real and interaction effects are illusionary. Actually, it is easily demonstrated through symbolic modeling that main effects in the presence of interactions are spurious.

    (b) It is a misguided tendency to interpret significant main effects first and significant interaction effects second. The correct interpreting and stopping rules (see Sawilowsky 2007a) are to begin with the highest order effect, and cease with the highest order statistically significant effect(s) on that level.

    For example, in a 2 × 2 × 2 ANOVA layout, meaningful interpretation begins with the a × b × c interaction. Analysis should cease if it is statistically significant. If it is not, then the focus of analysis descends to the a × b, a × c, and b × c lower order interactions. If none are statistically significant, it is then appropriate to give attention to the a, b, and c main effects. (c) It is true that MANOVA is useful even when there are only univariate hypotheses, because the sole reason for invoking it is to provide increased statistical power. Thus, it is meaningful to follow with univartiate tests to provide further insight after a statistically significant MANOVA result. However, it is a misconception that so-called step-down univariate tests are necessary, or meaningful, to interpret a statistically significant MANOVA that was conducted to examine a multivariate hypothesis, which by definition is multivariate because it consists of hopelessly intertwined dependent variables (see Sawilowsky 2007a).
17. *ANCOVA*. (a) This procedure is the Catch-22 of statistical methods. Because it is erroneously assumed to correct for baseline differences, and baseline differences are concomitant with the lack of ►randomization, the myth has arisen that using ANCOVA rehabilitates the lack of randomization. Unfortunately, to be a legitimate test ANCOVA requires randomization, only after which it serves to decrease the error term in the denominator of the F ratio, and hence increase statistical power.

    (b) ANCOVA, even when legitimately applicable due to randomization, is used to control for unwanted effects. The logic of partitioning and then removing sums of squares of an effect known to be significant is nearly meritless. It is by far more realistic to retain and model the unwanted effects by entering it (by some technique other than dummy coding) into a general linear model (i.e., regression) than it is to remove it from consideration.

    Consider a hypothetical treatment for the fresh water fish disease ichthyophthirius multifilis (ich). Suppose to determine its effectiveness the following veterinarian prescribed treatment protocol must be followed: (1) Remove the water while the fish remain in the aquarium. (2) Wait ten days until all moisture is guaranteed to have evaporated from the fish. (3) Apply Sawilowsky's miracle *ich-b-gone*™®© salve to the fish. (4) Wait an additional ten days for the salve to completely dry. (5) Refill the aquarium with water. Results of the experiment show no evidence of ich. Hence, the salve is marketable as a cure for ich, controlling for water.

    (c) There is a propensity, especially among doctoral dissertation proposals, and proposals submitted to funding agencies, to invoke as many covariates into ANCOVA as possible, under the mistaken impression that any covariate will reduce the error term and result in a more powerful test. In fact, a covariate must be carefully chosen. If it is not highly correlated with the dependent variable the trivial sum of squares that it may remove from the residual in the denominator will not overcome the impact of the loss of the *df*, resulting in a less powerful test. See Sawilowsky (2007b) for other myths regarding ANCOVA.
18. *Readership's view on publication differs from retraction and errata*. One of the most unfortunate, and sometimes insidious, characteristics of peer reviewed statistical outlets is the propensity to publish new and exciting statistical procedures that were derived via elegant squiggles, but were never subjected to Monte Carlo or other real data analysis methodologies to

determine their small samples Type I error and power properties. It appears that the more prestigious the outlet, the greater is the reluctance in publishing subsequent notices to the readership that the statistic or procedure fails, is severely limited, or has no practical value. If an editor imagines an article is so important to the readership that it is publishable, it is a misconception for editors to presume that the same readership would be uninterested in subsequently learning that the article was erroneous.

Some editors and reviewers, in an effort to protect the prestige of the outlet, create great barriers to correcting previously published erroneous work, such as demanding that the critical manuscript also solve the original problem in order to be worthy of publication (e.g., Hyman 1995). For example, this removes oversight if an ineffective or counter-productive cure for cancer was published by demanding the rebuttal author first cure cancer in order to demonstrate the published cure was vacuous.

19. *Mathematical and applied statistics/data analysis*. It is a myth that mathematical statistics and applied statistics/data analysis share a common mission and toolkit. The former is a branch of mathematics, whereas the latter are not. The consumer of real world statistics rejoices over an innovation that increases the ability to analyze data to draw a practical conclusion that will improve the quality of life, even if the memoir in which it was enshrined will never appear in the American Mathematical Society's *Mathematical Reviews* and its *MathSciNet* online database.
20. *Statisticians, authors of statistical textbooks, and statistics*. The following are myths: (a) Statisticians are subject matter experts in all disciplines. (b) Statisticians are mathematician wannabes. (c) Anyone who has a cookbook of statistical procedures is a qualified statistician. Corollary: Only the British need to certify statisticians. (d) Anyone who has taken an undergraduate course in statistics is qualified to teach statistics or serve as an expert witness in court. (e) Statistics textbooks are free from computational errors. (f) Statistics textbook authors are consistent in their use of symbols. (g) If three randomly selected statistics textbook authors opine the same view it must be true. Corollary: It is a myth that if a statistical topic is examined in three randomly selected statistics textbooks the explanations will be *i.i.d.* (h) t, F, regression, etc., aren't statistics – they are data analysis. (i) It is a myth that statistics can be used to perform miracles.

About the Author

Biography of Shlomo Sawilowsky is in ►Frequentist Hypothesis Testing: A Defense

Cross References

►Analysis of Covariance
►Asymptotic Relative Efficiency in Estimation
►Box–Cox Transformation
►Confidence Interval
►Data Analysis
►Effect Size
►Frequentist Hypothesis Testing: A Defense
►Interaction
►Misuse of Statistics
►Monte Carlo Methods in Statistics
►Multivariate Analysis of Variance (MANOVA)
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Normal Scores
►Null-Hypothesis Significance Testing: Misconceptions
►Permutation Tests
►Power Analysis
►P-Values
►Randomization
►Rank Transformations
►Robust Statistics
►Scales of Measurement and Choice of Statistical Methods
►Statistical Fallacies
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Blair RC, Sawilowsky SS, Higgins JJ (1987) Limitations of the rank transform in factorial ANOVA. Communications in Statistics-Computations and Simulations B16:1133–1145

Brewer JK (1985) Behavioral statistics textbooks: Source of myths and misconceptions? J Educ Stat 10:252–268

Huck SW (2008) Statistical Misconceptions. Psychology Press, London

Huff D (1954) How to lie with statistics. Norton, New York

Hunter MA, May RB (1993) Some myths concerning parametric and nonparametric tests. Can Psychol 34(4):365–469

Hyman R (1995) How to critique a published article. Psychol Bull 118(2):178–182

King G (1986) How not to lie with statistics: avoiding common mistakes in quantitative political science. Am J Polit Sci 30(3):666–687

Leamer E (1985) Sensitivity analyses would help. Am Econ Rev 75:308–313

Micceri T (1989) The Unicorn, the normal curve, and other improbable creatures. Psychol Bull 105(1):156–166

Sawilowsky S (1985) Robust and power analysis of the 2 × 2 × 2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests. Unpublished doctoral dissertation, University of South Florida

Sawilowsky S (1985b) A comparison of random normal scores test under the F and Chi-square distributions to the 2x2x2 ANOVA test. Florida J Educ Res 27:83–97
Sawilowsky S (1990) Nonparametric tests of interaction in experimental design. Rev Educ Res 60(1):91–126
Sawilowsky SS (1993) Comments on using alternatives to normal theory statistics in social and behavioral sciences. Can Psychol 34(4):432–439
Sawilowsky S (2003a) A different future for social and behavioral science research. J Mod Appl Stat Meth 2(1):128–132
Sawilowsky SS (2003b) You think you've got trivials? J Mod Appl Stat Meth 2(1):218–225
Sawilowsky SS (2003c) Trivials: The birth, sale, and final production of meta-analysis. J Mod Appl Stat Meth 2(1):242–246
Sawilowsky S (2003d) Deconstructing arguments from the case against hypothesis testing. J Mod Appl Stat Meth 2(2):467–474
Sawilowsky S (2005) Misconceptions leading to choosing the t test over the Wilcoxon Mann-Whitney U test for shift in location parameter. J Mod Appl Stat Meth 4(2):598–600
Sawilowsky S (2007a) ANOVA: effect sizes, simulation interaction vs. main effects, and a modified ANOVA table. In: Sawilowsky S (ed) Real data analysis, Ch. 14, Information Age Publishing, Charlotte, NC
Sawilowsky S (2007b) ANCOVA and quasi-experimental design: the legacy of campbell and stanley. In: Sawilowsky S (ed) Real data analysis, Ch. 15, Information Age Publishing, Charlotte, NC
Sawilowsky S, Blair RC, Higgins JJ (1989) An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. J Educ Stat 14:255–267
Thompson B (1995) Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. Educ Psychol Meas 55(4):525–534
Vandenberg RJ (2006) Statistical and methodological myths and urban legends: Where, pray tell, did they get this idea? Organ Res Meth 9:194–201

Statistical Genetics

Susan R. Wilson
Professor, Faculty of Medicine and Faculty of Science
University of New South Wales, Sydney, NSW, Australia

Statistical genetics broadly refers to the development and application of statistical methods to problems arising in genetics. Genetic data analysis covers a broad range of topics, from the search for the genetic background affecting manifestation of human diseases to understanding genetic traits of economic importance in domestic plants and animals. The nature of genetic data has been evolving rapidly, particularly in the past decade, due mainly to ongoing advancements in technology.

The work over a century ago of Gregor Mendel, using inbred pea lines that differed in easily scored characteristics, marks the start of collecting and analysing genetic data. Today we can easily, and relatively inexpensively, obtain many thousands, even millions or more, of genetic and phenotypic, as well as environmental, observations on each individual. Such data include high-throughput gene expression data, single nucleotide polymorphism (SNP) data and high-throughput functional genomic data, such as those that examine genome copy number variations, chromatin structure, methylation status and transcription factor binding. The data are being generated using technologies like microarrays, and very recently, next-generation sequencing. In the next few years, it is anticipated that it will be possible to sequence an entire human genome for $100, in a matter of days or even hours. The sheer size and wealth of these new data are posing many, ongoing, challenges.

Traditionally there have been close links between developments in genetics and in statistics. For example Sir RA Fisher's proposal of ►analysis of variance (ANOVA) can be traced back to the genetic problems in which he was interested. It is not widely known that probabilistic graphical models have their origins at about the same time in S Wright's genetic path analysis. A current thrust of modern statistical science concerns research into methods for dealing with data in very high dimensional space, such as is being generated today in molecular biology laboratories. New opportunities abound for analysing extremely complex biological data structures.

Basic analyses of genetic data include estimation of allele and haplotype frequencies, determining if Hardy-Weinberg equilibrium holds, and evaluating linkage disequilibrium. Statistical analyses of sequence, structure and expression data cover a range of different types of data and questions, from mapping, to finding sequence homologies and gene prediction, and to finding protein structure. Although many tools appear ad hoc, often it is found that there are some solid, statistical underpinnings. For example, the very widely used heuristic computational biology tool, Basic Local Alignment Sequence Tool (BLAST) is based on random walk theory (see ►Random Walk).

In animal and plant breeding, there are a range of approaches to finding and mapping quantitative trait loci, in both inbred lines and outbred pedigrees. Population genetics is a large topic in its own right, and is concerned with the analysis of factors affecting the genetic composition of a population. Hence it is centrally concerned with evolutionary questions, namely the change in the genetic composition of a population over time due to

natural selection, mutation, migration, and other factors. The knowledge of the structure of genes as DNA sequences has completely changed population genetics, including retrospective theory, in which a sample of genes is taken, DNA sequence determined, and the questions relate to the way in which, through evolution, the population has arrived at its presently observed state. For intrapopulation genetic inferences, coalescent theory (whereby from a sample of genes one traces ancestry back to the common ancestor) is fundamental. Evolutionary genetics is another, huge, topic. Many approaches have been developed for phylogenetic analyses, from applying likelihood methods, to use of parsimony and distance methods. In forensics, the use of DNA profiles for human identification often requires statistical genetic calculations. The probabilities for a matching DNA profile can be evaluated under alternative hypotheses about the contributor(s) to the profile, and presented as likelihood ratios. Conditional probabilities are needed, namely the probabilities of the profiles given that they have already been seen, and these depend on the relationships between known and unknown people.

Genetic epidemiology is a growing area, especially with current research to find the genes underpinning complex genetic diseases. "Methodological research in genetic epidemiology (is developing) at an ever-accelerating pace, and such work currently comprises one of the most active areas of methodological research in both ►biostatistics and epidemiology. Through an understanding of the underlying genetic architecture of common, complex diseases modern medicine has the potential to revolutionize approaches to treatment and prevention of disease" (Elston et al. 2002). Pharmacogenetics research is concerned with the identification and characterization of genes that influence individual responses to drug treatments and other exogenous stimuli. Modern pharmacogenetics involves the evaluation of associations between genetic polymorphisms and outcomes in large-scale clinical trials typically undertaken to evaluate the efficacy of a particular drug in the population at large. Meta-analysis methods (see ►Meta-Analysis) are an increasingly important tool for modern genetic analysis.

A starting point for the whole area of statistical genetics is the "Handbook" (Balding et al. 2004) that is also available online. Interestingly, the final chapter addresses ethics in the use of statistics in genetics. An encyclopaedic approach is used in the reference text of Elston et al. (2002). Software also is proliferating, and a good starting point is the suite of R packages in the Comprehensive R Archive Network (CRAN) Task View: Statistical Genetics (http://cran.r-project.org/web/views/Genetics.html) and in Bioconductor (http://www.bioconductor.org), an open source and open development software project for the analysis of genomic data.

About the Author

For biography *see* the entry ►Biostatistics.

Cross References

►Analysis of Variance
►Bioinformatics
►Biostatistics
►Forensic DNA: Statistics in
►Medical Statistics

References and Further Reading

Balding DJ, Bishop M, Cannings C (2004) Handbook of statistical genetics, 2nd edn. Wiley

Elston RC, Olson JM, Palmer L (2002) Biostatistical genetics and genetic epidemiology. Wiley, New York

Weir BS (1996) Genetic data analysis II. Sinaur Assocs, Sunderland, MA

Statistical Inference

Richard A. Johnson
Professor Emeritus
University of Wisconsin, Madison, WI, USA

At the heart of statistics lie the ideas of statistical inference. Methods of statistical inference enable the investigator to argue from the particular observations in a sample to the general case. In contrast to logical deductions from the general case to the specific case, a statistical inference can sometimes be incorrect. Nevertheless, one of the great intellectual advances of the twentieth century is the realization that strong scientific evidence can be developed on the basis of many, highly variable, observations.

The subject of statistical inference extends well beyond statistics' historical purposes of describing and displaying data. It deals with collecting informative data, interpreting these data, and drawing conclusions. Statistical inference includes all processes of acquiring knowledge that involve fact finding through the collection and examination of data. These processes are as diverse as opinion polls, agricultural field trials, clinical trials of new medicines, and the studying of properties of exotic new materials. As a consequence, statistical inference has permeated all fields of human endeavor in which the evaluation of information must be grounded in data-based evidence.

A few characteristics are common to all studies involving fact finding through the collection and interpretation of data. First, in order to acquire new knowledge, relevant data must be collected. Second, some variability is unavoidable even when observations are made under the same or very similar conditions. The third, which sets the stage for statistical inference, is that access to a complete set of data is either not feasible from a practical standpoint or is physically impossible to obtain.

To more fully describe statistical inference, it is necessary to introduce several key terminologies and concepts. The first step in making a statistical inference is to model the population(s) by a *probability distribution* which has a numerical feature of interest called a *parameter*. The problem of statistical inference arises once we want to make generalizations about the *population* when only a *sample* is available.

A *statistic*, based on a sample, must serve as the source of information about a parameter. Three salient points guide the development of procedures for statistical inference

1. Because a sample is only part of the population, the numerical value of the statistic will not be the exact value of the parameter.
2. The observed value of the statistic depends on the particular sample selected.
3. Some variability in the values of a statistic, over different samples, is unavoidable.

The two main classes of inference problems are *estimation* of parameter(s) and *testing hypotheses* about the value of the parameter(s). The first class consists of point estimators, a single number estimate of the value of the parameter, and interval estimates. Typically, the interval estimate specifies an interval of plausible values for the parameter but the subclass also includes prediction intervals for future observations. A test of hypotheses provides a yes/no answer as to whether the parameter lies in a specified region of values.

Because statistical inferences are based on a sample, they will sometimes be in error. Because the actual value of the parameter is unknown, a test of hypotheses may yield the wrong yes/no answer and the interval of plausible values may not contain the true value of the parameter.

Statistical inferences, or generalizations from the sample to the population, are founded on an understanding of the manner in which variation in the population is transmitted, via sampling, to variation in a statistic. Most introductory texts (see Johnson and Bhattacharyya 2010; Johnson, Freund, and Miller 2011) give expanded discussions of these topics.

There are two primary approaches, *frequentist* and *Bayesian*, for making statistical inferences. Both are based on the *likelihood* but their frameworks are entirely different.

The frequentist treats parameters as fixed but unknown quantities in the distribution which governs variation in the sample. Then, the frequentist tries to protect against errors in inference by controlling the probabilities of errors. The long-run relative frequency interpretation of probability then guarantees that if the experiment is repeated many times only a small proportion of times will produce incorrect inferences. Most importantly, using this approach in many different problems keeps the overall proportion of errors small.

Frequentists are divided on the problem of testing hypotheses. Some statisticians (Cox 2006) follow R. A. Fisher and perform *significance tests* where the decision to reject a *null hypothesis* is based on values of the statistic that are extreme in directions considered important by subject matter interest. It is more common to take a *Neyman–Pearson* approach where an *alternative hypothesis* is clearly specified together with the corresponding distributions for the statistic. *Power*, the probability of rejecting the null hypothesis when it is false, can then be optimized. A definitive account of Neyman–Pearson theory is given in Lehmann and Casella (2003) and Lehmann and Romano (2008).

In contrast, Bayesians consider unknown parameters to be random variables and, prior to sampling, assign a *prior distribution* for the parameters. After the data are obtained, the Bayesian takes the product prior times likelihood and obtains the *posterior distribution* of the parameter after a suitable normalization. Depending on the goal of the investigation, a pertinent feature or features of the posterior distribution are used to make inferences. The mean is often a suitable point estimator and a suitable region of highest posterior density gives an interval of plausible values. See Box and Tiao (1973) and Gelman et al. (2004) for discussions of Bayesian approaches.

A second phase of statistical inference, *model checking*, is required for both frequentist and Bayesian approaches. Are the data consonant with the model or must the model be modified in some way? Checks on the model are often subjective and rely on graphical diagnostics.

D. R. Cox, gives an excellent introduction to statistical inference in Cox (2006) where he compares Bayesian and frequentist approaches and highlights many of the important issues.

Statistical inferences have been extended to semiparametric and fully nonparametric models where functions are the infinite dimension parameters.

About the Author

Richard A. Johnson is Professor Emeritus at the University of Wisconsin following 42 years on the regular faculty of the Department of Statistics (1966–2008). He served as Chairman (1981–1984). Professor Johnson has coauthored six books including (a) *Applied Multivariate Statistical Analysis* (1982, 6th edition 2007), Prentice-Hall with D. W. Wichern, (b) *Probability and Statistics for Engineers* (1990, 8th edition 2011), Prentice-Hall, with I. Miller and J. E. Freund, and (c) *Statistics-Principles and Methods* (1985, 6th edition 2010), J. Wiley and Sons, with G. K. Bhattacharyya. Richard is a recipient of the *Technometrics* Frank Wilcoxon Prize (1991), the Institute of Mathematical Statistics Carver Award (2008), and the American Statistical Association, San Antonio Chapter, Don Owen Award (2009). He is founding editor of *Statistics and Probability Letters* and served as editor for the first 25 years (1992–2007). Professor Johnson is a Fellow of the American Statistical Association, Fellow of the Institute of Mathematical Statistics, Fellow of the Royal Statistical Society and an Elected member of the International Statistical Institute. He has published 125 research papers and has lectured in more than 22 countries. His research interests include multivariate analysis, reliability and life testing, and large sample theory.

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Confidence Interval
►Estimation
►Estimation: An Overview
►Likelihood
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Robust Inference
►Significance Testing: An Overview
►Statistical Inference: An Overview

References and Further Reading

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Addison-Wesley
Cox DR (2006) Principles of statistical inference, Cambridge University Press, Cambridge
Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, 2nd edn. Chapman and Hall/CRC Press, Boca Raton, FL
Johnson R, Freund J, Miller I (2011) Miller and Freund's probability and statistics for engineers, 8th edn. Prentice-Hall, Upper Saddle River
Johnson R, Bhattacharyya GK (2010) Statistics – principles and methods, 6th edn. Wiley, Hoboken, NJ
Lehmann EL, Casella GC (2003) Theory of point estimation, 2nd edn. Springer, New York
Lehmann EL, Romano JP (2008) Testing of statistical hypotheses, 3rd edn. Springer, New York

Statistical Inference for Quantum Systems

Alexander S. Helevo
Professor
Steklov Mathematical Institute, Moscow, Russia

With the advent of lasers and optical communication it was realized that specific restrictions on the fidelity of information transmission due to quantum-mechanical nature of a communication channel need be taken into account and require a special approach. In the 1960–1970s this led to creation of a consistent quantum statistical decision theory which gave the framework for investigation of fundamental limits for detection and estimation of the states of quantum systems (Helstrom; Holevo 1976; 1982). In this theory statistical uncertainty is described by using mathematical apparatus of quantum mechanics – operator theory in a Hilbert space. Thus, the quantum statistical decision theory is a "noncommutative" counterpart of the classical one which was based on the Kolmogorov probability model and both of them can be embedded into a general framework (Holevo 1976). The interest to quantum statistical inference got the new impetus at the turn of the century (Barndorff-Nielsen et al. 2003). In high precision and quantum optics experiments researchers became able to operate with elementary quantum systems such as single ions, atoms and photons leading to potentially important applications such as quantum cryptography and novel communication protocols. In currently discussed proposals for quantum computing, the information is written into states of elementary quantum cells – qubits, and is read off via quantum measurements. Therefore the issue of extracting the maximum statistical information from the state of a given quantum system becomes important. On the other hand, building a consistent statistical theory of quantum measurement has significant impact onto foundations of quantum mechanics resulting in clarification of several subtle points. Last but not the least, quantum statistical inference has a number of appealing specifically

noncommutative features which open new perspectives for avantgarde research in the mathematical statistics.

As in the classical statistical decision theory, there is a set Θ of values of an unknown parameter θ, a set $\mathcal{X}$ of decisions x and a loss function $L_\theta(x)$, defining the quality of the decision x for a given value of parameter θ. The difference comes with the description of statistical uncertainty: here to each θ corresponds a density operator ρ_θ in the separable Hilbert space $\mathcal{H}$ of the system. *Density operator* ρ is a positive operator in $\mathcal{H}$ with unit trace, describing *state* of the quantum system. In physical problems the quantum system is the information carrier such as coherent electromagnetic field, prepared by transmitter in a state which depends on the signal θ.

A *decision rule* is defined by a quantum *measurement* with outcomes $x \in \mathcal{X}$. In the case of finite set $\mathcal{X}$ corresponding to hypotheses testing (detection), decision rule is described mathematically by a *resolution of the identity* in $\mathcal{H}$, i.e., the family of operators $M = \{M_x; x \in \mathcal{X}\}$ satisfying

$$M_x \geq 0, \quad \sum_{x\in\mathcal{X}} M_x = I, \tag{1}$$

where I is the identity operator. The probability of making decision x in the state ρ_θ is defined by the basic formula generalizing the Born-von Neumann statistical postulate

$$P_M(x|\theta) = \mathrm{Tr}\rho_\theta M_x.$$

Decision rule is implemented by a receiver making a quantum measurement and the problem is to find the optimal measurement performance.

The mean risk corresponding to the decision rule M is given by the usual formula

$$R_\theta\{M\} = \sum_{x\in\mathcal{X}} L_\theta(x) P_M(x|\theta). \tag{2}$$

In this way one has a family $\{R_\theta\{M\}, \theta \in \Theta\}$ of affine functionals defined on the convex set $\mathfrak{M}(\mathcal{X})$ of decision rules (1). The notions of admissible, minimax, Bayes decision rule are then defined as in the classical Wald's theory. The profound difference lies in the much more complicated convex structure of the sets of quantum states and decision rules.

The *Bayes risk* corresponding to a priori distribution π on Θ is

$$R_\pi\{M\} = \int_{\theta\in\Theta} R_\theta\{M\} d\pi(\theta) = \mathrm{Tr} \sum_{x\in\mathcal{X}} \hat{L}(x) M_x, \tag{3}$$

where

$$\hat{L}(x) = \int_{\theta\in\Theta} \rho_\theta L_\theta(x) d\pi(\theta) \tag{4}$$

is the operator-valued posterior loss function. Bayes decision rule minimizing $R_\pi\{M\}$ always exists and can be found among extreme points of the convex set $\mathfrak{M}(\mathcal{X})$. An illustration of the effect of noncommutativity is the following analog of the classical rule saying that Bayes procedure minimizes posterior loss: M is Bayes if and only if there exists Hermitian trace-class operator Λ such that

$$\Lambda \leq \hat{L}(x), \quad (\hat{L}(x) - \Lambda) M_x = 0, \quad x \in \mathcal{X}. \tag{5}$$

The operator Λ plays here the role of the minimized posterior loss.

The Bayes problem can be solved explicitly in a number of important cases, notably in the case of two hypotheses and for the families of states with certain symmetry. In general, symmetry and invariance play in quantum statistical inference much greater role; on the other hand, the concept of sufficiency has less applicability because of the severe restrictions onto existence of conditional expectations in the noncommutative probability theory (Petz 2008).

The optimum is found among the extreme points of the convex set of decision rules which therefore play a central role. In the classical case the extreme points are precisely deterministic decision rules. Their quantum analog are *orthogonal resolutions of the identity* satisfying $M_x M_y = \delta_{xy} M_x$ in addition to (1). However in the noncommutative case these form only a subset of all extreme decision rules. According to a classical result of Naimark, any resolution of the identity can be extended to an orthogonal one in a larger Hilbert space. In statistical terms, such an extension amounts to an outer quantum randomization. Consequently, there are quantum Bayes problems in which the optimal rule is inherently "randomized" (Holevo 1982). This paradoxical fact has a profound physical background, namely, the measurement *entanglement* between the system and the outer randomizer, which is a kind of intrinsically quantum correlation due to tensor product structure of the composite systems in quantum theory. Notably, in standard approach to quantum mechanics only orthogonal resolutions of the identity (namely, spectral measures of self-adjoint operators) were considered as representing *observables* (i.e., random variables). Thus, quantum statistical decision theory gives a strong argument in favor of the substantial generalization of the fundamental notion of quantum observable.

As in the classics, the case of two simple hypotheses ρ_0, ρ_1 is the most tractable one: there are quantum counterparts of the Neumann-Pearson criterion and of the asymptotics for the error probability and for the Bayes risk (the quantum Chernoff bound). However the derivation of these asymptotics is much more involved due to possible noncommutativity of the density operators ρ_0, ρ_1 (Hayashi 2006).

In estimation problems Θ and $\mathcal{X}$ are parametric varieties (typically $\mathcal{X} = \Theta \subset \mathbb{R}^s$) and the decision rules are given by *positive operator-valued measures* on Θ which are (generalized) spectral measures for operators representing the estimates. Solution of the Bayes estimation problem can be obtained by generalizing results for finite $\mathcal{X}$ with appropriate integration technique (Holevo 1976). Explicit solutions are obtained for problems with symmetry and for estimation of the mean value of Bosonic Gaussian states. The last is quantum analog of the classical "signal+noise" problem, however with the noise having quantum-mechanical origin and satisfying the canonical commutation relations (Holevo 1982).

Quantum statistical treatment of models with the shift or rotation parameter provides a consistent approach to the issue of canonical conjugacy and nonstandard uncertainty relations in quantum mechanics, such as time-energy, phase-number of quanta, as well as to approximate joint measurability of incompatible observables. In the quantum case estimation problems with multidimensional parameter are inherently more complex than those with one-dimensional parameter. This is due to the possible non-commutativity of the components reflecting existence of *incompatible* quantities that in principle cannot be measured exactly in one experiment. This sets new statistical limitations to the components of multidimensional estimates, absent in the classical case, and results in essential non-uniqueness of logarithmic derivatives and of the corresponding quantum Cramér–Rao inequalities (Helstrom 1976; Holevo 1982).

Another special feature of quantum statistical inference appears when considering series of i.i.d. quantum systems: the statistical information in quantum models with independent observations can be strictly superadditive. This means that the value of a measure of statistical information for a quantum system consisting of independent components can be strictly greater than the sum of its values for the individual systems. The property of strict superadditivity is again due to the existence of entangled (collective) measurements over the composite system (Hayashi 2005).

One of the most important quantum estimation models is the *full model*, in which the state is assumed completely unknown. In the case of finite dimensionality d this is a parametric model with a specific group of symmetries (the unitary group), in particular, for $d = 2$ it is the model of unknown qubit state (i.e., 2×2-density matrix), with the three-dimensional Stokes parameter varying inside the Bloch sphere. The most advanced results here concern the asymptotic estimation theory for the i.i.d. observations, culminating in the noncommutative analog of Le Cam's local asymptotic normality for estimation of an arbitrary mixed state of a finite dimensional quantum system (Guta and Kahn 2009; Hayashi 2005). The full model in infinite dimensions belongs to nonparametric quantum mathematical statistics, which is at present in a stage of development. In this connection the method of *homodyne tomography* of a density operator widely used in quantum optics is particularly important (Artiles et al. 2005).

Quantum statistical decision theory provides powerful general methods for computing fundamental limits to accuracy of physical measurements, which serve as benchmarks for evaluating the quality of existing physical measurement procedures. It also gives the mathematical description of the optimal decision rule; however the quantum theory in principle provides no universal recipe for constructing a measuring device from the corresponding resolution of the identity and such kind of problems have to be treated separately in each concrete situation. Still, in several cases methods of quantum statistical inference give important hints towards the realization (based, e.g., on covariance with respect to the relevant symmetries) and can provide an applicable description of the required (sub)optimal measurement procedure (Artiles et al. 2005; Hayashi 2005; Helstrom 1976) .

Acknowledgment

Supported in part by RFBR grant 09-01-00424 and the program "Mathematical control theory" of Russian Academy of Sciences.

About the Author

Alexander S. Holevo (Kholevo) is Professor at the Steklov Mathematical Institute. He is also a Professor at the Moscow State University and Moscow Institute for Physics and Technology. A. S. Holevo has been awarded the Markov Prize of Russian Academy of Sciences (1997) for his work in the noncommutative probability, the International Quantum Communication Award (1996) and A. von Humboldt Research Award (1999) for the development of mathematical theory of quantum information systems. He is the author of five monographs and more than 160 research articles in classical and quantum probability, statistics and information theory and in the mathematical foundations of quantum mechanics. He is currently Co-editor-in-chief of the journal *Theory of Probability and Its Applications*.

Cross References

►Astrostatistics
►Bayesian Statistics
►Chernoff Bound

►Decision Theory: An Introduction
►Decision Theory: An Overview
►Loss Function
►Markov Chain Monte Carlo
►Random Matrix Theory
►Statistical Inference: An Overview
►Stochastic Processes

References and Further Reading

Artiles L, Gill RD, Guta M (2005) An invitation to quantum tomography. J Roy Stat Soc B 67:109–134
Barndorff-Nielsen OE, Gill RD, Jupp PE (2003) On quantum statistical inference. J Roy Stat Soc B 65:775–816
Guta M, Kahn J (2009) Local asymptotic normality for finite dimensional quantum systems. Commun Math Phys 289(2):597–652
Hayashi M (ed) (2005) Asymptotic theory of quantum statistical inference. Selected papers. World Scientific, New York
Hayashi M (2006) Quantum information: an introduction, Springer, New York
Helstrom CW (1976) Quantum detection and estimation theory. Academic, New York
Holevo AS (1976) Investigations in the general theory of statistical decisions. Proc Steklov Math Inst 124:1–140 (AMS Translation, 1978, Issue 3)
Holevo AS (1982) Probabilistic and statistical aspects of quantum theory. North-Holland, Amsterdam
Petz D (2008) Quantum information theory and quantum statistics. Springer, Berlin

Statistical Inference for Stochastic Processes

M. B. Rajarshi
President of the International Indian Statistician Association (Indian Chapter)
Professor
University of Pune, Pune, India

Statistical inference for ►stochastic processes deals with dependent observations made at time points in $\{0, 1, 2, \cdots\}$ or $[0, \infty)$. Thus, the time parameter can be either discrete or continuous in nature.

Markov Chains and Sequences

Let $\{X_t, t = 0, 1, 2, \ldots\}$ be a time-homogeneous L-order Markov sequence with the state-space S. Let $p_\theta(x_t|x_{t-1}, x_{t-2}, \cdots x_{t-L})$ be the conditional probability mass function(p.m.f.) or probability density function(p.d.f.) of X_t given $X_{t-1}, X_{t-2}, \cdots X_{t-L}$, θ being an unknown parameter in Θ, an open set in the K-dimensional Euclidean space. The (conditional) log-likelihood (given $(X_{(1)}, X_{(2)}, \cdots, X_{(L)})$) is given by $\ln(L_T(\theta)) = \sum_{t=L,T} \ln[p_\theta(x_t|x_{t-1}, x_{t-2}, \cdots, x_{t-L})]$, $T > L$. We assume that the conditional p.m.f./p.d.f. satisfies the Cramer regularity conditions and that $\{X_t, t = 0, 1, 2, \ldots\}$ is a strictly stationary and ergodic sequence. The Fisher Information matrix is defined by

$$I(\theta) = ((\, -E[\partial^2 \ln(p_\theta(X_t|X_{t-1}, X_{t-2}, \cdots X_{t-L}))/\partial\theta_i \partial\theta_j] \,))$$

and is assumed to be positive definite (the expectation is with respect to the joint distribution of $(X_t, X_{t-1}, \cdots, X_{t-L})$ and is computed under the assumption of stationarity). Under these conditions, it can be shown that there exists a consistent solution $\widehat{\theta}$ of the likelihood equations, such that $\sqrt{T}(\widehat{\theta} - \theta) \to N_K(0, [I(\theta)]^{-1})$ in distribution (Billingsley 1961). We apply the ►martingale central limit theorem to the score function (i.e., the vector of $\partial \ln(L_T(\theta))/\partial\theta_i$, $i = 1, 2, \cdots, K$) (Billingsley 1961; Hall and Heyde 1980) and the Strong Law of Large numbers for various sample averages of stationary and ergodic sequences to prove this result. The large-sample distribution theory of Likelihood Ratio Tests (LRTs) and confidence sets follows in a manner similar to the case of independently and identically distributed (i.i.d.) observations.

Some of the assumptions made above can be relaxed, cf. Basawa and Prakasa Rao (1980), Chap. 7. The LRT can be used for selecting the order of a model by testing a model against the alternatives of a higher order model. However, the ►Akaike's Information Criterion (AIC) and Bayes criterion (BIC), respectively given by $AIC = -2 \ln L_T(\hat{\theta}) + K$ and $BIC = -2 \ln L_T(\hat{\theta}) + K \ln(T)$ are more appropriate for selection of a model and an order. The model with the least AIC/BIC is selected. When S is finite, the procedure based on BIC yields a consistent estimator of the true order, cf. Katz (1981). The AIC is an inconsistent procedure, cf. Davison (2003), Sect. 4.7. For finite Markov chains, Pearson's χ^2-statistic can be used in place of the LRT for various hypotheses of interest. In moderate samples, the chi-square approximation to Pearson's χ^2-statistic is better than the same to LRT.

First order Markov models offer a satisfactory fit to observations somewhat infrequently. Lindsey (2004, p. 113) discusses approaches based on ►logistic regression and log-linear models (contingency table analysis) for higher order finite ►Markov chains. A distinct advantage of such a modeling is that both time-dependent and time-independent covariates can be incorporated, see discussion of Generalized Auto-Regressive Moving Average (GARMA) models below. A limitation of such models is that the conditional probabilities depend upon the numerical values (coding) assigned to the states, which is not suitable for models for data without any numerical structure, such as linguistic classes.

Higher order Markov chains and sequences can be handicapped by a large number of parameters. An important Markov model of order L with a substantially small number of parameters is due to Raftery (1985) and it is given by

$$p_\theta(x_t|x_{t-1}, x_{t-2}, \cdots x_{t-L}) = \sum_{l=1,L} \lambda_l q_{x_{t-l},x_t}, \quad \lambda_l \geq 0, \quad \sum_l \lambda_l = 1.$$

Here, $q_{x,y}$ is a transition probability matrix (t.p.m.) or a transition density. The model is known as Mixture Transition Density (MTD) model. For an M-state chain, the number of parameters of the MTD model is $M(M-1) + L - 1$, far less (particularly for $M > 2$) than $(M^L)(M-1)$, the number of parameters in the corresponding saturated Markov chain. In the MTD models, like the Auto-Regressive (AR) time series models, we need to add only a single parameter to the r-order model to get the $(r+1)$-order model. We may note that if the state-space is continuous or countably infinite, the transition density $q_{x,y}$ is a specified function of K unknown parameters.

Non-Markovian Models

Hidden Markov Model (HMM). HMM was introduced in speech recognition studies. It has a very wide range of applications. Let $\{Y_t, t = 0, 1, 2, \ldots\}$ be a first-order Markov chain with the state-space $S_y = \{1, 2, \cdots M\}$ and the one-step t.p.m. P. The Markov chain $\{Y_t, t = 0, 1, 2, \ldots\}$ is not observable. Let $\{X_t, t = 0, 1, 2, \ldots\}$ be an observable process taking values in S_x with M_1 elements such that $P[X_t = j|Y_t = i, Y_{t-1}, \cdots Y_0, X_{t-1} \cdots X_{t-2}, \cdots, X_0] = q_{ij}$, $i \in S_x$, $j \in S_y$. Thus, if $M_1 = M$, the number of parameters of a Hidden Markov chain is $2M(M-1)$ which is considerably smaller than a higher order Markov chain. For estimation of unobserved states $\{Y_t, t = 0, 1, 2, \ldots T\}$ and estimation of parameters, the Baum-Welsch algorithm is widely used, which is an early instance of the Expectation-Maximization (EM) algorithm.

For a discussion of Hidden Markov chains, we refer to MacDonald and Zucchini (1997) and Elliot et al. (1995). Cappe et al. (2005) give a thorough and more recent account of a general state-space HMM.

ARMA Models for integer valued random variables. A non-negative Integer-valued ARMA (INARMA) sequence is defined as follows. The binomial operator $\gamma \circ W$ is defined by a binomial random variable with W as the number of trials and γ as the success probability (if $W = 0$, $\gamma \circ W = 0$). Let $\{Z_t, t = 0, \pm 1, \pm 2, \cdots\}$ be a sequence of i.i.d. non-negative integer valued random variables with a finite variance. Then, the $INARMA(p, q)$ process is defined by $X_t = \sum_{i=1,p} \alpha_i \circ X_{t-i} + \sum_{j=1,q} \beta_j \circ Z_{t-j} + Z_t$. All the binomial experiments required in the definition of the process are independent. The process $\{Z_t\}$ is not observable. The process $\{X_t\}$ is (second order) stationary if $\sum \alpha_i < 1$ and is invertible if $\sum \beta_j < 1$. An excellent review of such processes has been given in McKenzie (2003). Interesting special cases such as AR, MA and Poisson, Binomial, Negative Binomial as the stationary distributions are reported therein.

GARMA models. These are extensions of the ▶Generalized Linear Models based on an exponential family of distributions and can incorporate vector of time-dependent covariates z_t along with past observations. The conditional mean of X_t given the past is given by $h(\eta_t)$ where $h^{-1} = g$ (say) is the link function of the chosen exponential family and $\eta_t = z_t'\gamma + \sum_{i=1,p} \phi_i \left[g(x_{t-i}) - z_{t-i}'\gamma\right] + \sum_{j=1,q} \theta_j [g(x_{t-j}) - \eta_{t-j}]$. The parameters $\{\phi_i\}$ and $\{\theta_j\}$ denote the auto-regressive and moving average parameters respectively. The parameter γ explains the effect of covariates. A modification of the mean function is required to take care of the range of the observations. A limitation of this class of models is that in the absence of regressors or when the vector γ is null, it may not be possible to have a stationary series. We refer to Benjamin et al. (2003) and Fahrmeir and Tutz (2004), Chap. 6 for more details.

Bienayme-Galton-Watson Branching Process

Billingsley's work based on martingale methods for deriving asymptotic properties of the maximum likelihood estimator paved the way for many interesting theoretical developments for non-ergodic models such as a Bienayme-Galton-Watson (BGW) branching process.

Let $\{X_t, t = 0, 1, \cdots\}$ be a BGW Branching process with the state-space $S = \{0, 1, \cdots\}$ and the off-spring distribution $p_k, k = 0, 1, \cdots$. Parameters of interest are the offspring distribution and its functions such as the mean μ and the variance σ^2. A number of estimators for μ have been suggested: Lotka's estimator X_T/X_{T-1} (taken to be 1 if $X_{T-1} = 0$), Heyde's estimator $(X_T)^{1/T}$ and the nonparametric maximum likelihood estimator $\widehat{\mu}_T = (Y_T - X_0)/Y_{T-1}$ with $Y_t = X_0 + X_1 + \ldots + X_t$. The maximum likelihood estimator has a natural interpretation that it is the ratio of the total number of off-springs (in the realization) born to the total number of parents. By using the Scott central limit theorem for martingales (Scott 1978), it can be shown that, on the non-extinction path, $\sqrt{Y_{T-1}}(\hat{\mu}_T - \mu)/\sigma$ is asymptotically standard Normal. A natural estimator of σ^2, resulting from regression considerations, is given by $(1/T)\sum_t X_{t-1}(X_t/X_{t-1} - \hat{\mu}_T)^2$. This can be shown to be consistent and asymptotic normal with $\sqrt{T}$-norming, if

the fourth moment of the offspring distribution is finite. These results are useful to construct tests and confidence intervals for μ.

Based on a single realization, only μ and σ^2 are estimable on the non-extinction path of the process (i.e., consistent estimators exist for these parameters), if no parametric form of the offspring distribution is assumed. A good account of inference for branching processes, their extensions and related population processes along with applications can be found in Guttorp (1991).

Non-parametric Modeling Based on Functional Estimation

For a stationary process, where every finite dimensional distribution is absolutely continuous, we may opt for a non-parametric approach. We estimate the conditional density of X_t given $X_{t-1}, X_{t-2}, \ldots, X_{t-L}$ by the ratio of estimators of appropriate joint densities. The joint density of p consecutive random variables is estimated by a kernel-based estimator as follows. Let $K_p(x)$ be a probability density function, where $x \in R^p$, the p-dimensional Euclidean space. Let h_T be a sequence of positive constants such that $h_T \to 0$ and $Th_T^p \to \infty$ as $T \to \infty$. The estimator of joint density of consecutive p observations at $(x_1, x_2, \cdots, x_p)$ is then given by $\widehat{f}(x_1, x_2, \cdots, x_p) = \left[1/\left(Th_T^p\right)\right] \sum_{j=1,T-p} K((x_1 - X_j)/h_T, (x_2 - X_{j+1})/h_T, \cdots, (x_p - X_{j+p})/h_T)$. Based on the estimator of the conditional p.d.f., one can estimate the conditional mean (or other parameters such as conditional median or mode).

Properties of conditional density estimators are established assuming that the random sequence $\{X_t, t = 0, 1, 2, \cdots\}$ satisfies certain mixing conditions. We discuss strong or α-mixing, since most of the other forms of mixing imply the strong mixing. Let $F_{0,s}$ be the σ-field generated by the random variables $(X_0, X_1, \cdots, X_s)$ and let $F_{s+t,\infty}$ be the σ-field generated by the collection of random variables $\{X_{s+t}, X_{s+t+1}, \cdots\}$. The stationary sequence $\{X_t, t = 0, 1, 2, \cdots\}$ is said to be strong mixing if $sup_{A \in F_{0,s}, B \in F_{s+t,\infty}} \{|P(A \cap B) - P(A)P(B)|\} \leq \alpha(t)$ and $\alpha(t) \to 0$ as $t \to \infty$. For most of the results, we need faster rates of decay of $\alpha(t)$. Asymptotic properties of the kernel-based estimator have been established in Robinson (1983) who also illustrates how plots of conditional means can be helpful in bringing out nonlinear relationships. Prakasa Rao (1996) discusses, in detail, non-parametric analysis of time series based on functional estimation.

Non-parametric inference. Tests for median or tests and estimation procedures based on order or rank statistics, like the widely used tests in the case of i.i.d. observations can be suggested. However, the exact distribution is neither free from the unknown parameters, nor it is known, except in some special cases. Thus, such procedures for stationary observations lack simplicity and elegance of the rank-based tests. Further, robustness of an estimator is much more complex for dependent observations, since the effect of a spurious observation or an outlier (which can be an innovation outlier in an ARMA model) spreads over a number of succeeding observations. In an important paper, Martin and Yohai (1986) discuss influence functions of estimators obtained from ARMA Time Series model.

Bootstrap. Efron's Bootstrap (see ▶Bootstrap Methods) for i.i.d. samples is now routinely used to estimate the variance or the sampling distributions of estimators, test statistics and approximate pivotals. In most of the situations of practical interests, it gives a more accurate estimator of the sampling distribution than the one obtained by the traditional methods based on the Central Limit Theorem. In the i.i.d. case, we obtain B bootstrap samples, each sample being a Simple Random Sample With Replacement (SRSWR) of size T from the observed sample. This generates B values of a statistic or pivotal of interest.

For a stationary AR model of order L, the first L values of a bootstrap series may be the same as those of the observed time series. We take a SRSWR sample of size $T-L$ from residuals. The randomly selected residuals are then successively used to generate a bootstrap time series. We then have B time series, each of length T. For stationary and invertible MA or ARMA models, a bootstrap series is constructed from a SRSWR sample of the residuals. Rest of the methodology is the same as the usual bootstrap procedure. Bose (1988) (AR models) and (1990) (MA models) has shown that such a bootstrap approximation to the sampling distribution of the least square estimators is superior to the traditional normal approximation.

Bootstrap procedures for (strictly) stationary and ergodic sequence are based on blocks of consecutive observations. Bootstrap procedure is a boon for stochastic models, since in most of the cases, working out the variance of a statistic or its sampling distribution is very complex. By and large, it is beyond the reach of an end-user of statistics. (Consider, for example, computing the variance of a 10 per cent trimmed mean computed from stationary observations.) In a Moving Blocks Bootstrap (MBB)(Kunsch 1989; Liu and Singh 1992), we form K blocks of L consecutive observations to capture the dependence structure of the process. There are $N = T - L + 1$ blocks of L consecutive observations. We obtain a SRSWR of size K from these N *blocks* to get a bootstrap sample of size $T^* = KL$. If T is divisible by L, $K = T/L$, otherwise, it can be taken to be the

integer nearest to T/L. Let F_T be the empirical distribution function of T observations and let H be a functional on the space of distribution functions, computed at F_T (such as the trimmed mean or a percentile). A bootstrap statistic H^* is computed from the empirical distribution function of T^* bootstrap observations. Other procedures are NBB (Non-overlapping Blocks Bootstrap) and CBB (Circular Blocks Bootstrap), cf. Lahiri (2003), Chap. 2. Carlstein (1986) considers non-overlapping subseries of size L.

Let us assume that $L \to \infty$, $T \to \infty$ such that $T/L \to \infty$. Kunsch has shown that the bootstrap estimator of the variance of the normalized sample mean ($\sqrt{T}\overline{X}$) is consistent. (He further discusses jackknife procedures wherein we delete a block at a time.) The MBB procedure correctly estimates the sampling distribution of the sample mean. This property holds for a large number of mean-like statistics and smooth (continuously differentiable) functions of the mean vector, see Lahiri (2003 p. 177). Statistics based on averages of consecutive observations or their smooth functions (such as serial correlation coefficients) can be similarly bootstrapped. Second-order properties of the bootstrap estimator of the sampling distribution of the normalized/Studentized smooth functions of the sample mean (vector) have been obtained by Lahiri (1991) and Gotze and Kunsch (1996). Let $G(\mu)$, a third order differentiable function of the population mean vector μ, be the parameter of interest. While constructing the bootstrap version of the pivotal, we need to consider $G(\overline{X}^*) - G(\hat{\mu}_T)$, where $\hat{\mu}_T = E^*(\overline{X}^*)$. If the block length L is of the order $T^{1/4}$, the best possible error rate of the MBB approximation for estimation of the distribution function is $O(T^{-3/4})$. Though it is not as good as the accuracy that we have in the case of i.i.d. or residual based ARMA bootstrap, it is still better than the normal approximation to an asymptotic pivotal. Optimal block lengths for estimator of variance and the sampling distribution of a smooth statistics have been discussed in Chap. 7 of Lahiri (2003).

Under certain conditions, it is possible to bootstrap the empirical process, cf. Radulovic (2002). Such results as well as those discussed above for block based bootstrap, assume that the underlying process is strong mixing with a specified rate of decay of the mixing coefficients along with the block lengths L. We can construct confidence bands for the distribution function, by using the bootstrap distribution of the empirical process. Further, a number of statistics such as natural estimators of a compactly differentiable functional of the distribution function can be bootstrapped. Such a class of estimators include most of the estimators that we use in practice.

Kulperger and Prakasa Rao (1989) discuss bootstrap estimation of the sampling distribution of the estimator of a suitable function of P, the one-step t.p.m. of a finite ergodic irreducible Markov chain. They consider the expected value of time taken to reach a state from another state of a Markov chain, as a parametric function P. Computing the variance of such an estimator is very tedious. Bootstrap samples are generated by regarding the maximum likelihood estimate of the t.p.m. P as the underlying parameter.

State-space models (Doubly stochastic processes/Randomly driven stochastic processes). Let $\{X_t, t = 0, 1, \cdots\}$ be an unobservable process. Let $\{Y_t, t = 0, 1, \cdots\}$ be an observable process with the conditional p.m.f. or p.d.f. $f(y_0, y_1, \cdots, y_t \mid x_0, x_1, \cdots, x_t)$. In practice, often the process $\{X_t, t = 0, 1, \cdots\}$ is a Markov sequence and the conditional distribution of Y_t given $(y_0, y_1, \cdots, y_{t-1}, x_0, x_1, \cdots, x_t)$ depends upon x_t and y_{t-1} only. Such models are useful in situations where parameters vary slowly over time. It may be noted that models such as HMM, MTD or ARMA among others can be conveniently viewed as state-space models. Varying parameters can be modeled by a random process, see Guttorp (1995, p. 111) for an example involving a two state Markov chain.

Counting and Pure Jump Markov Processes

Let $\{X(t), t \geq 0\}$ be a counting process with $X(0) = 0$. Let $F(t_-)$ be the complete history up to t but not including t (technically the σ-field generated by the collection of random variables $\{X(u), u < t\}$). The intensity function $\lambda(t)$ can be stochastic (a random variable with respect to $F(t_-)$). It is characterized by the properties that $P[X(t + dt) - X(t) = 1 | F(t_-)] = \lambda(t)dt + o(dt)$, $P[X(t + dt) - X(t) = 0 | F(t_-)] = 1 - \lambda(t)dt + o(dt)$ and $P[X(t + dt) - X(t) \geq 1 | F(t_-)] = o(dt)$ for small dt. We assume that $E(X(t)) < \infty$ for every t. Let $M(t) = X(t) - E[X(t)|F(t_-)]$. It can be shown that $\{M(t), t > 0\}$ is a continuous time martingale with respect to $F(t_-)$, i.e., $E[M(t + s) - M(t) \mid F(t_-)] = 0$ for every $s > 0$. Time-dependent or time independent regressors can be included in the intensity function $\lambda(t)$.

Let the intensity $\lambda(t)$ be $\lambda(t, \theta)$, a specified function of the time and the parameters θ. In practice, to informally compute the likelihood, a partition $t_0 = 0, t_1, t_2, \ldots, t_N = T$ of $[0, T]$ is selected and the likelihood for such a partition is computed first. One then allows the norm of this partition to converge to 0. It turns out that the likelihood is given by $\ln(L(\theta) = \int \ln(\lambda(u, \theta))dX(u) - \int \lambda(u, \theta)I(u)du$, where $I(t) = 1$, if there is a jump at t and 0, otherwise. Such a general formulation linking counting processes inference with martingales in continuous time is due to Aalen (1978).

Important special cases include (a) Poisson process (see ▶Poisson process) with $\lambda(t) = \lambda$ for all t; (b) a Non-homogeneous Poisson Process where $\lambda(t)$ is a deterministic function, (c) Pure birth process $\lambda(t) = \lambda X(t_-)$, and (d) Renewal process (see ▶Renewal Processes) $\lambda(t) = h[t - t(x(t))]$ where $h(t)$ is the failure rate or hazard function of the absolutely continuous lifetime distribution of the underlying i.i.d. lifetimes and $t(x(t))$ is the time epoch at which the last failure before t takes place. (d) Semi-Markov or Markov renewal process. Here the intensity function depends on the state of the process at $t(x(t))$ and the state observed at t (assuming that there is an event at t).

Inference for counting processes and asymptotic properties of the maximum likelihood estimators have been discussed in Karr (1986) and Andersen et al. (1993).

Likelihood of a time-homogeneous continuous time *Pure Jump Markov process* follows similarly. Let, for $i \neq j$, $P[X(t + dt) = j|X(t) = i] = \lambda_{ij}dt + o(dt)$ and let $P[X(t + dt) = i|X(t) = i] = 1-Q_{ii}dt+o(dt)$. The probability of other events is $o(dt)$. Here, $Q_{ii} = -\sum_{j\neq i}\lambda_{ij}$. If the state space is finite, each of the row-sums of the matrix $Q = ((Q_{ij}))$ is 0. The transition function $P[X(t) = j|X(0) = i]$ of the process is assumed to be differentiable in t for every i, j. The log-likelihood, conditional on $X(0) = x(0)$, is given by $\ln L = \sum_{i\neq j} N_{ij} \ln Q_{ij} - \sum_i Q_{ii}\tau_i$, where N_{ij} is the number of direct transitions from i to j and τ_i is the time spent in the state i, both during $[0, T]$. If the number of states is finite, the non-parametric maximum likelihood estimator of Q_{ij} is given by Nij/τ_i. Properties of maximum likelihood estimators have been discussed in Adke and Manjunath (1984) and Guttorp (1995, Chap. 3). Important cases include (Linear or Non-linear) Birth-Death-Immigration-Emigration processes and Markovian Queuing models.

Goodness of fit procedures are both graphical and formal. The Q-Q plot of the times spent in a state i scaled by the maximum likelihood estimates of their expected values, reveals departures from the exponential distribution. Since N_{ij}'s form transition counts of the embedded Markov chain, one can check whether such transitions have any memory. If the model under study has a stationary distribution, the observed frequencies of the test can be compared with the fitted stationary distribution, see Keiding (1975) who analyzes a Birth-Death-Immigration process model.

Diffusion Processes

Let $\{X(t), t \geq 0\}$ be a diffusion process with $\mu(x, \theta)$ and $\sigma^2(x)$ as the trend and diffusion functions respectively. The likelihood for the observed path $\{X(t), 0 \leq t \leq T\}$ is the Radon-Nikodym derivative of the probability measure of $\{X(t), 0 \leq t \leq T\}$ under the assumed diffusion process with respect to the probability measure of $\{X(t), 0 \leq t \leq T\}$ under the assumption of a diffusion process with the mean function equal to 0 for all x and the variance function $\sigma^2(x)$. It is assumed that $\sigma^2(x)$ is a known function. The log-likelihood is given by

$$\ln(L(\theta)) = \int_{0,T} \mu(x(t), \theta)/(\sigma(x(t))dx(t) - 1/2 \int_{0,T} \mu^2(x(t), \theta)/(\sigma(x(t))dt.$$

(If the variance functions is unknown, a time transformation is used to reduce the process with a known variance function.) Some special cases are (a) Brownian motion, (b) Geometric Brownian Motion, and (c) Ornstein-Uhlenbeck process. $\sqrt{T}-$ consistency and ▶asymptotic normality of the estimator of the mean of the process can be shown under the assumption that the process is non-null persistent (i.e., the process almost surely returns to any bounded set and the corresponding mean return time is finite). In this case, we can obtain non-parametric estimators of the common distribution function and the probability density function of $X(t)$. We refer to Prakasa Rao (1999a) and Kutoyants (2004) for details. Kutoyants (2004) also discusses asymptotic distributions of the estimator of the mean of the process in the null persistent case.

Observing a continuous time process may not be always feasible. We choose a partition of $[0, T]$, write the likelihood of such a partially observed process and then take the limit as the norm of the partition tends to 0. Validity of such operations has been established in Kutoyants (2004). Sorensen (2004) gives an extensive review for inference for stationary and ergodic diffusion processes observed at discrete points. The following techniques are discussed therein: (a) estimating functions with special emphasis on martingale estimating functions and so-called simple estimating functions, (b) analytical and numerical approximations of the likelihood function which can, in principle, be made arbitrarily accurate, (c) Bayesian analysis and MCMC methods, and (d) indirect inference and Generalized Method of Moments which both introduce auxiliary (but wrong) models and correct for the implied bias by simulation.

Statistical analysis and theoretical derivation of diffusion processes (as well as counting processes) is based on the theory of semimartingales. A semimartingale is a sum of a local martingale and a function of bounded variation. A class of diffusion processes and counting processes form

a subclass of the family of submartingales. A unified theory of statistical inference for semimartingales is presented in Prakasa Rao (1999b).

A fractional diffusion process is driven by the fractional Brownian motion (see ►Brownian Motion and Diffusions), which is not a semimartingale. Such processes can be useful in modeling phenomena with long range dependence, but the earlier techniques based on the theory of semimartingales are not applicable. Statistical inference for fractional diffusion processes has been discussed in Prakasa Rao (2010).

Concluding Remarks

Computational aspects. Computation of likelihood and its subsequent maximization are involved for most of the stochastic models. There are many procedures such as Kalman Filter, EM algorithm and Monte Carlo EM algorithm (which is based on Markov Chain Monte Carlo methods, see ►Markov Chain Monte Carlo), to compute the likelihood and the maximum likelihood estimator. From a computer programming view-point, implementation of the EM algorithm and its stochastic versions, require a special routine for each model. The conditional expectation step may require extensive simulations from a joint density, the constant of integration of which is not known. For state-space models, one needs to carry out a T-tuple integral (or a sum) to compute the likelihood. It seems that various methods based on numerical analysis to get a good approximation to the likelihood, its maximization and derivatives (which are needed to compute standard error of the maximum likelihood estimator), are preferred to other procedures. Possibly this is due to a very slow rate for convergence of the EM algorithm (and its stochastic versions) and yet another round of computations required to compute the estimator of the variance of the maximum likelihood estimator.

Efficiency of Estimators

(a) *Finite sample optimality.* Godambe's criterion (Godambe 1985) of a finite sample optimality of an estimator is based on optimality of the estimating equation it solves. Under the usual differentiability-based regularity conditions, an estimating function g^* is said to be optimal in G, if it minimizes $E(g(A)^2)/(E(\partial g(A)/\partial\theta))^2$. Let F_t be the σ–field generated by the collection of random variables $\{X_s, s = 0, 1, \cdots, t\}$. Let $g(t,\theta)$ be an F_t measurable random variable involving θ, a real parameter, such that $E[g(t,\theta) \mid F_{t-1}] = 0$ and $Var[g(t,\theta) \mid F_{t-1}] = V(t)$. Let $g(A) = \sum_t A(t)g(t,\theta)$, where $A(t)$ is an F_{t-1} measurable random variable, $t \geq 1$. Let $G = \{g(A)\}$ be the class of estimating functions $g(A)$ which satisfy the regularity conditions together with the assumptions that $E(g(A)^2) < \infty$ and $E(\partial g(A)/\partial\theta) \neq 0$. Godambe proves that the optimal choice of $A(t)$ is given by $E[\partial g(t,\theta)/\partial\theta \mid F_{t-1}]/V(t)$. In practice, we need to assume that such optimal weights do not involve other (incidental or nuisance) parameters.

A number of widely used estimators turn out to be solutions of such an optimal estimating equations $g^* = 0$. Further, Godambe's result justifies the estimator for each finite sample size and in addition, it broadens the class of parametric models to a larger class of semi-parametric models, for which the estimating function is optimal. The score function is optimal in a class of regular estimating functions, justifying use of the maximum likelihood estimator in finite samples. Continuous time analogues of these results with applications to counting processes have been discussed in a number of papers in a volume edited by Godambe (1991) and Prakasa Rao and Bhat (1996).

Optimality of an estimating function in a class is also equivalent to an optimal property of confidence intervals based on it. In large samples, the optimal g^* leads to a shortest confidence interval for θ at a given confidence coefficient. In a number of situations, the confidence interval, obtained from a Studentized estimating function, is typically better than the approximate pivotal obtained by Studentizing the corresponding estimator, in the sense that the true coverage rate of the procedure based on estimating function is closer to the nominal confidence coefficient. Bootstrapping the Studentized estimating function further improves performance of the corresponding confidence interval.

(b) *Asymptotic efficiency*. In non-ergodic models such as a BGW process, large-sample efficiency issues are rather complex. Though the random norming is convenient from an application view-point, the non-random norming is more appropriate and meaningful for efficiency issues. Further, notions of asymptotic efficiency based on variance of an estimator are no more applicable, since the variance of the asymptotic distribution for a large number of estimators does not exist. The W-efficiency of the maximum likelihood estimator, under certain regularity conditions, has been established by Hall and Heyde (1980) and Basawa and Scott (1983). Estimators based on other criteria can also be W-efficient. The Bayes estimator, under certain conditions, is asymptotically distributed like the maximum likelihood estimator. This result is known as the Bernstein-von Mises theorem and for its proof in the case of stochastic processes, we refer to Chap. 10 of Basawa and Prakasa Rao (1980).

Inference problems in stochastic processes have enriched both theoretical investigations and applied statistics.

Theoretical research in bootstrap, estimating functions, functional estimation and non-Gaussian non-Markov processes has widened scope of stochastic models. Use of fast and cheap computing has been helpful in computing likelihood, maximum likelihood estimators and Bayes estimators in very complicated stochastic models.

Acknowledgment

I am thankful to Professor B.L.S. Prakasa Rao and two other reviewers for a number of helpful suggestions.

About the Author

Dr. M. B. Rajarshi retired in 2009, as a Professor of statistics from the University of Pune. His areas of of interests are inference for stochastic processes, applied probability and stochastic modeling. He has published about 35 papers, some of which have appeared in *Annals of Statistics*, *Journal of Applied Probability*, *Journal of the American Statistical Association*, *Annals of the Institute of Statistical Mathematics*, *Naval Logistic Quarterly*, *Statistics and Probability Letters*, *Communications in Statistics*, *Ecology and Theoretical Population Biology*. He has held visiting appointments at Penn State University, University of Waterloo and Memorial University of Newfoundland (Canada). He was elected as Member of the International Statistical Institute (1998). He was the Chief Editor of the *Journal of the Indian Statistical Association* (2000–2006). At present, Dr. Rajarshi is the President of the International Indian Statistician Association-Indian Chapter and the Vice-President of the Indian Statistical Association.

Cross References

►Akaike's Information Criterion
►Asymptotic Relative Efficiency in Estimation
►Bootstrap Methods
►Brownian Motion and Diffusions
►Central Limit Theorems
►Generalized Linear Models
►Kalman Filtering
►Likelihood
►Markov Chain Monte Carlo
►Markov Chains
►Markov Processes
►Martingale Central Limit Theorem
►Martingales
►Methods of Moments Estimation
►Nonparametric Estimation
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Poisson Processes
►Renewal Processes
►Statistical Inference: An Overview
►Stochastic Processes
►Stochastic Processes: Classification

References and Further Reading

Aalen OO (1978) Nonparametric inference for a family of counting processes. Ann Stat 6:701–726
Adke SR, Manjunath SM (1984) An introduction to finite Markov processes. Wiley Eastern, New Delhi
Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical Models based on counting processes. Springer, New York
Basawa IV, Prakasa Rao BLS (1980) Statistical inference for stochastic processes. Academic, London
Basawa IV, Scott DJ (1983) Asymptotic optimal inference for non-ergodic models. Springer, New York
Benjamin M, Rigby R, Stasinopoulos M (2003) Generalized autoregressive moving average models. J Am Stat Assoc 461:214–223
Billingsely P (1961) Statistical inference for Markov processes. Chicago University Press, Chicago, IL
Bose A (1988) Edgeworth correction by bootstrap in autoregressions. Ann Stat 1709–1722
Bose A (1990) Bootstrap in moving average models. Ann Inst Stat Math 42:753–768
Cappe O, Moulines E, Ryden T (2005) Inference in Hidden Markov models. Springer, New York
Carlstein E (1986) The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann Stat 14:1172–1179
Davison AC (2003) Statistical models. Cambridge University Press, Cambridge
Elliot RJ, Aggaoun L, Moore JB (1995) Hidden Markov models: estimation and control. Springer, New York
Fahrmeir L, Tutz G (2004) Multivariate statistical modelling based on generalized linear models. Springer, New York
Godambe VP (1985) The foundation of finite sample estimation in stochastic processes. Biometrika 72:419–428
Godambe VP (ed) (1991) Estimating Functions. Oxford Science Publications, New York
Gotze F, Kunsch HR (1996) Second-order correctness of the block-wise bootstrap for stationary observations. Ann Stat 24: 1914–1933
Guttorp P (1991) Statistical inference for branching processes. Wiley, New York
Guttorp P (1995) Stochastic modeling of scientific data. Chapman & Hall, London
Hall P, Heyde CC (1980) Martingale limit theory and its applications. Academic, New York
Karr AF (1986) Point processes and their statistical inference. Marcel Dekker, New York
Katz RW (1981) On some criteria for estimating the order of a Markov chain. Technometrics 23:243–249
Keiding N (1975) Maximum likelihood estimation in birth and death process. Ann Stat 3:363–372
Kulperger RJ, Prakasa Rao BLS (1989) Bootstrapping a finite Markov chain. Sankhya A 51:178–191
Kunsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann Stat 17:1217–1261
Kutoyants YA (2004) Statistical inference for ergodic diffusion processes. Springer, New York

Lahiri SN (1991) Second order optimality of stationary bootstrap. Statist Probab Lett 11:335–341
Lahiri SN (2003) Resampling methods for dependent data. Springer, New York
Lindsey JK (2004) Statistical analysis of stochastic processes in time. Cambridge University Press, Cambridge
Liu RY, Singh K (1992) Moving blocks jackknife and bootstrap capture weak dependence. In: Lepage R, Billard L (eds) Exploring the limits of bootstrap. Wiley, New York, pp 225–248
MacDonald I, Zucchini W (1997) Hidden Markov and other models for discrete valued time series. Chapman & Hall, London
Martin RD, Yohai VJ (1986) Influence functionals for time series. Ann Stat 14:781–818
McKenzie E (2003) Discrete variate time series. In: Shanbhag DN, Rao CR (eds) Stochastic processes: modeling and simulation. Handbook of statistics. North-Holland, Amsterdam, pp 573–606
Prakasa Rao BLS (1996) Nonparametric Approach in Time Series Analysis. In: Prakasa Rao BLS, Bhat BR (eds) Stochastic processes and statistical inference. New Age International New Delhi, pp 73–89
Prakasa Rao BLS (1999a) Statistical inference for diffusion type processes. Arnold, London
Prakasa Rao BLS (1999b) Semimartingales and their statistical inference. CRC Press, Boca Raton, FL
Prakasa Rao BLS (2010) Statistical inference for fractional diffusion processes. Wiley, New York
Prakasa Rao BLS, Bhat BR (eds) (1996) Stochastic processes and statistical inference. New Age International, New Delhi
Radulovic D (2002) On the bootstrap and the empirical processes for dependent sequences. In: Dehling H, Mikosch T, Sorensen M (eds) Empirical process techniques for dependent data, Birkhauser, Boston, pp 345–364
Raftery AE (1985) A model for high order Markov chains. J Roy Stat Soc B 47:528–539
Robinson PM (1983) Nonparametric estimators for time series. J Time Ser Anal 4:185–207
Scott DJ (1978) A central limit theorem for martingales and an application to branching processes. Stoch Processes Appl 6:241–252
Sorensen H (2004) parametric inference for diffusion processes observed at discrete points in time: a survey. Internat Statist Rev 72:337–354

Statistical Inference in Ecology

Subhash R. Lele[1], Mark L. Taper[2]
[1]Professor
University of Alberta, Edmonton, AB, Canada
[2]Research Scientist
Montana State University, Bozeman, MT, USA

Researchers in ecology and evolution have long recognized the importance of understanding randomness in nature in order to distinguish the underlying pattern. Sir Francis Galton developed regression analysis to answer questions about heredity; Karl Pearson's systems of distributions were motivated by the desire to fit evolutionary data on the size of crab claws. Fisher's contributions from the fundamental theorem of evolution to fields of quantitative genetics, species abundance distributions and measurement of diversity are legendary. Studies on the geographic distribution of species led to the study of spatial statistics in ecology in the early part of the 20th century. The classification and discrimination methods developed by Fisher and others for numerical taxonomy and community ecology are still commonly used in ecology.

Unfortunately, Karl Pearson believed that causation was an illusion of scientific perception, stating in the introduction to the 1911 3rd edition of *The Grammar of Science*, "Nobody believes now that science explains anything; we all look upon it as a shorthand description, as an economy of thought." Under Pearson's influence, statistical techniques in ecology tended, until recently, to be more descriptive than predictive with a major early exception of path analysis developed by Sewall Wright in the first decades of the 20th century.

In curious contradiction, mathematical models used by ecologists to model population dynamics and related processes were highly sophisticated and predictive in nature. For example, Lotka–Volterra models were developed in the 1930s. Generalization of these models to multispecies cases such as the Predator-Prey, Host-Parasitoid and other systems of models were available soon after that. Skellam (1951) pioneered the use of spatial diffusion processes to model spread of invasive species.

Gause's work (Gause 1934) was unique in that he tried to validate the mathematical models using experimental data. He used non-linear regression to fit Logistic growth model to the population growth series for paramecia. Most of this work was based on the assumption that error comes into the process only through observational inaccuracies, and thus he missed the modern nuance of inherent randomness or process variation.

Statistical ecology received a large impetus in the 1970s after the publication of Professor E.C. Pielou's numerous classic books (e.g., Pielou 1977) and number of conferences and the resultant edited volumes by Professor G.P. Patil (e.g., Patil et al. 1971). These provided nice summaries of what was known then and also indicated future directions. Driven by the passage of the 1973 Endangered Species Act (ESA) and the dozens of other environmental laws passed in the United States during the 1970's the field of ecology gained substantial prominence in the context of managing and not simply describing ecosystems. This necessitated the development of models that were predictive and not simply descriptive.

Population Viability Analysis (PVA) where one uses stochastic models to predict the distribution of extinction times for a population or species of concern became an important tool for studying the effect of various human activities on nature. Political decisions regarding the conservation of species are often legally required by the ESA to consider the results of a PVA. The importance of demographic and environmental stochasticity as well as the measurement error in forecasting became apparent. Expanding beyond a single population focus, the development of meta-population theory was based on probabilistic models for spatial dispersal and growth. Ecologists became more familiar and comfortable with the idea of modeling randomness and studying its impact on prediction. While much of what is modeled as random in ecology undoubtedly represents unrecognized deterministic influences, it seems likely that true stochasticity is as much a fundamental part of ecology as it is in physics. For example, demographic events such as the sex of offspring are truly random, and not simply the consequence unrecognized deterministic influences. Such demographic stochasticity strongly influences population dynamics when population size is low.

Although stochastic models became prominent in the 1970s and 80s, statistical inference, the methods that connect theoretical models to data, or inductive inference, was still limited. Most of the statistical techniques used were based on linear regression and its derivatives such as the ►Analysis of Variance. The main hurdles were limited data, limited computational power and mathematical nature of the statistical inferential tools. Dennis et al. (1991) and Dennis and Taper (1994) made a major advance by incorporating stochastic population dynamic models as the skeleton for a full likelihood based inference in ecological time series.

The rapid rise in computational power available to ecologists, coupled with the development of computational statistical techniques especially the bootstrap (see ►Bootstrap Methods) and Monte-Carlo approaches have reduced the threshold of mathematical expertise necessary to apply sophisticated statistical inference techniques making the analysis of complex ecological models feasible. This has provided significant impetus for developing strong inferential tools in ecology.

Following are some of the important examples of the application of statistical thinking in ecology.

1. *Sampling methods for estimation of population abundances and occurrences*: Mark-Capture-Recapture (Seber 2002) methods have formed an important tool in the statistical ecology toolbox, but have also led to development of new statistical methods that have found applications in epidemiology and other sciences. Capture probabilities may change temporally or spatially. ►Generalized Linear models and mixed models have proved their usefulness in these situations. Biases due to visibility are adjusted using distance based sampling methods. In many instances, it is too expensive to conduct abundance estimation and one has to settle for site occupancy models based on presence-absence data. Site occupancy data and methods have made a broader range of ecologists aware of the ubiquitous nature of measurement error. Although a species may be present, it may not be detected because of various factors such as lack of visibility, time of the day when birds may not be singing etc. (MacKenzie et al. 2006). This is an active area of research.
2. *Resource selection by animals*: Ecologists need to know what resources animals select and how does this selection affect their fitness and survival. Human developments such as dams or a gas pipe line across a habitat that might be critical to the animals can doom their survival. Recent technological advances such as GPS collars and DNA analysis help in collecting information on where animals spend their time and what they eat. The resource selection probability function (RSPF) (Manly et al. 2002; Lele and Allen 2006) and habitat suitability maps (Hirzel et al. 2006) have been essential tools for environmental impact assessments (EIA) for studying impact of various developments.
3. *Model identification and selection*: The statistical models used for prediction can be either process driven or phenomenological, "black box", models (Breiman 2001). Predictions from ecological models are often made for the distant and not the immediate future. This extrapolation makes it essential that ecological models be process driven. The use of powerful likelihood methods for analyzing population time series models is a relatively new development. The predictions are strongly affected by the particular process based model chosen. This has forced ecologist to consider many models simultaneously and to search for good methods for ►model selection. Information based model selection (Burnham and Anderson 2002) has received considerable attention in this context. Although alternative methods and modifications are constantly being suggested and tested (Taper et al. 2008).
4. *Hierarchical models*: This is one of the most exciting developments in statistical ecology. General hierarchical models are also known as latent variable models, random effects models, mixed models and ►mixture models. These models are natural models to account

for the hierarchical structure inherent in many ecological processes. They also simplify statistical analysis in the presence of missing data, sampling variability, covariates measured with error and other problems commonly faced by ecologists. Reviews of the use of hierarchical models in ecology are available in Royle and Dorazio (2009), Cressie et al. (2009) or Clark and Gelfand (2006). Survival analysis methods and random effects models have found important applications in avian nest survival studies (Natarajan and McCulloch 1999). Linear mixed effects models have been used in evolution and animal breeding since the 1940's. However, generalization of those ideas to more complex models was not possible until recently. Writing down the likelihood function for general hierarchical models is difficult (Lele et al. 2007) and hence use of standard likelihood based inference is not popular. On the other hand, non-informative Bayesian inference using Markov Chain Monte Carlo algorithm (see ►Markov Chain Monte Carlo) is computationally feasible. These calculations are simulation based and replicate the causal processes that ecologists seek to understand. Due to its simplicity, the non-informative Bayesian approach has become quite popular in ecology. However, there are important philosophical and pragmatic issues that should be considered before using this approach (Lele and Allen 2006, Lele and Dennis 2009). Moreover, the recent development of the data-cloning algorithm (Lele et al. 2007; Ponciano et al. 2009) has removed the computational obstacle to likelihood inference for general hierarchical models.

Powerful statistical methods are being developed for ecology, generally coupled with software. The development of accessible tools has greatly facilitated the application of complex statistical analysis to ecological problems. These advances have come at a cost. Researchers are under pressure to be cutting edge and consequently tend to use techniques because they are convenient and fashionable not necessarily because they are appropriate.

Ecological statistics is vibrant and contributing greatly to the advancement of the science, but what are the future directions? One clear recommendation that can be made is in the realm of teaching. Education in ecological statistics has not kept pace with statistical practice in ecology, and improvements are desperately needed (Lele and Taper 2002, Dennis 2004). While methods instruction will always be essential, what is needed most by young ecologists is the development of strong foundational thinking about the role of statistical inference in ecological research.

On the other hand, recommendations regarding the development of new statistics are less clear. Techniques generally follow the questions that need to be answered. However, we are confident that while descriptive statistics and black box prediction will have their place, the greatest advances to knowledge in ecology will come from challenging the probabilistic predictions from explicit models of ecological process with data from well-designed experiments and surveys.

About the Authors

For biographies of both authors *see* the entry ►Statistical Evidence.

Cross References

►Bayesian Statistics
►Distance Sampling
►Factor Analysis and Latent Variable Modelling
►Linear Mixed Models
►Marine Research, Statistics in
►Modeling Survival Data
►Multilevel Analysis
►Nonlinear Mixed Effects Models
►Non-probability Sampling Survey Methods
►Proportions, Inferences, and Comparisons
►Statistical Ecology

References and Further Reading

Breiman L (2001) Statistical modeling: the two cultures. Stat Sci 16:199–215

Burnham KP, Anderson DR (2002) Model selection and multi-model inference: a practical information-theoretic approach, 2nd edn. Springer-Verlag, New York

Clark JS, Gelfand A (eds) (2006) Hierarchical modelling for the environmental sciences: statistical methods and applications. Oxford university press, Oxford, U.K

Cressie N, Calder CA, Clark JS, Ver Hoef JM, Wikle CK (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol Appl 19:553–570

Dennis B (2004) Statistics and the scientific method in ecology. In: Taper ML, Lele SR (eds) The nature of scientific evidence: statistical, empirical and philosophical considerations. University of Chicago Press, USA, pp 327–378

Dennis B, Taper ML (1994) Density dependence in time series observations of natural populations: estimation and testing. Ecol Monogr 64:205–224

Dennis B, Munholland PL, Scott JM (1991) Estimation of growth and extinction parameters for endangered species. Ecol Monogr 61:115–143

Gause GF (1934) The struggle for existence. Williams and Wilkins, Baltimore, MD, USA

Hirzel AH, LeLay G, Helfer V, Randin C, Guisan A (2006) Evaluating the ability of habitat suitability models to predict species presence. Ecol Model 199:142–152

Lele SR, Allen KL (2006) On using expert opinion in ecological analyses: a frequentist approach. Environmetrics 17:683–704
Lele SR, Dennis B (2009) Bayesian methods for hierarchical model: are ecologists making a Faustian bargain? Ecol Appl 19:581–584
Lele SR, Keim JL (2006) Weighted distributions and estimation of resource selection probability functions. Ecology 87:3021–3028
Lele SR, Taper ML (2002) What shall we teach in environmental statistics? Discussion. Environ Ecol Stat 9(2):145–146
Lele S, Dennis B, Lutscher F (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecol Lett 10:551–563
MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE (2006) Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Academic Press, NY
Manly BFJ, McDonald LL, Thomas DL, McDonald TL, Erickson WP (2002) Resource selection by animals: statistical analysis and design for field studies, 2nd edn. Kluwer Press, Boston, USA
Natarajan R, McCulloch CE (1999) Modeling heterogeneity in nest survival data. Biometrics 55:553–559
Patil GP, Pielou EC, Waters WE (1971) Statistical ecology: proceedings of the 1969 International Conference on Statistical Ecology, New Haven. Pennsylvania State University Press, PA, USA
Pielou EC (1977) Mathematical ecology. Wiley, New York
Ponciano J, Taper ML, Dennis B, Lele SR (2009) Hierarchical models in ecology: confidence intervals, hypothesis testing and model selection using data cloning. Ecology 90:356–362
Royle A, Dorazio R (2009) Hierarchical models and inference in ecology: the analysis of data from populations, metapopulations and communities. Elsevier Inc., UK
Seber GAF (2002) The estimation of animal abundance and related parameters, 2nd edn. Blackburn Press, Caldwel, NJ
Skellam JG (1951) Random dispersal in theoretical populations. Biometrika 38:196–218
Taper ML, Staples DF, Shepard BB (2008) Model structure adequacy analysis: selecting models on the basis of their ability to answer scientific questions. Synthese 163:357–370

Statistical Inference: An Overview

Aris Spanos
Wilson Schmidt Professor
Virginia Tech, Blacksburg, VA, USA

Introduction

Statistical inference concerns the application and appraisal of methods and procedures with a view to *learn from data* about observable stochastic phenomena of interest using probabilistic constructs known as *statistical models*. The basic idea is to construct statistical models using probabilistic assumptions that "capture" the chance regularities in the data with a view to adequately account for the underlying data-generating mechanism; see **?** (**?**). The discussion that follows focuses primarily on frequentist inference, and to a lesser extent on Bayesian inference.

The perspective on statistical inference adopted here is broader than earlier accounts, such as: "making inferences about a population from a random sample drawn from it" (Dodge 2003), in so far as it extends its intended scope beyond *random samples* and static *populations*, to include dynamic phenomena giving rise to observational (non-experimental) data. In addition, the discussion takes into account the fact that the demarcation of the intended scope of statistical inference is intrinsically challenging because it is commonly part of broader scientific inquiries; see Lehmann (1990). In such a broader context statistical inference is often *preceded* with substantive questions of interest, combined with the selection of data pertaining to the phenomenon being studied, and *succeeded* with the desideratum to relate the inference results to the original substantive questions.

This special placing of statistical inference raises a number of crucial methodological problems pertaining to the adequateness of the statistical model to provide a well-grounded link between the phenomenon of interest, at one end of the process, and furnishing evidence for or against the substantive hypotheses of interest, at the other. The link between the phenomenon of interest and the statistical model – thru the data – raises several methodological issues including: the role of substantive and statistical information (Lehmann 1990), as well as the criteria for selecting a statistical model and establishing its adequacy Spanos (2007). The link between the data – construed in the context of a statistical model – and evidence for or against particular substantive claims also raises a number of difficult problems including the fact that "accept" or "reject" the null hypothesis (or a small p-value) does not mean that there is evidence for the null or the alternative, respectively. Indeed, one can make a case that most of the foundational problems bedeviling statistical inference since the 1930s stem from its special place in this broader scientific inquiry; see Mayo (2006).

Frequentist Statistical Inference

Modern statistical inference was founded by Fisher (1922) who initiated a change of paradigms in statistics by recasting the then dominating *Bayesian-oriented induction*, relying on large sample size (n) approximations (Pearson 1920), into a frequentist *statistical model-based induction*, relying on *finite sampling distributions*, inspired by Gosset's (1908) derivation of the Student's t distribution for any sample size $n > 1$. Before Fisher, the notion of a statistical model was implicit, and its role was primarily confined to the *description* of the distributional features

of *the data in hand* using the histogram and the first few sample moments. Unlike Karl Pearson who would commence with data $\mathbf{x}_0$: = $(x_1, \ldots, x_n)$ in search of a frequency curve to describe the histogram of $\mathbf{x}_0$, he proposed to begin with (a) a prespecified model (a hypothetical infinite population), and (b) view $\mathbf{x}_0$ as a realization thereof. Indeed, he made the initial choice (specification) of the prespecified statistical model a response to the question: "Of what population is this a random sample?" (Fisher 1922, p. 313), emphasizing that: "the adequacy of our choice may be tested a posteriori" (ibid., p. 314).

The Notion of a Statistical Model

Fisher's notion of a prespecified statistical model can be formalized in terms of the stochastic process $\{X_k,\ k\in\mathbb{N}\}$, underlying data $\mathbf{x}_0$. This takes the form of parameterizing the probabilistic structure of $\{X_k,\ k\in\mathbb{N}\}$ to specify a *statistical model*:

$$\mathcal{M}_\theta(\mathbf{x}) = \{f(\mathbf{x};\theta),\ \theta\in\Theta\},\ \mathbf{x}\in\mathbb{R}_X^n,\ \text{for } \Theta\subset\mathbb{R}^m,\ m<n. \tag{1}$$

$f(\mathbf{x};\theta)$ denotes the joint *distribution of the sample* $\mathbf{X}$: = $(X_1, \ldots, X_n)$ that encapsulates the whole of the probabilistic information in $\mathcal{M}_\theta(\mathbf{x})$, by giving a general description of the probabilistic structure of $\{X_k,\ k\in\mathbb{N}\}$ (Doob 1953). $\mathcal{M}_\theta(\mathbf{x})$ is chosen to provide an idealized description of the mechanism that generated data $\mathbf{x}_0$ with a view to appraise and address the substantive questions of interest.

The quintessential example of a statistical model is *the simple Normal model*:

$$\mathcal{M}_\theta(\mathbf{x})\text{: } X_k \backsim \text{NIID}(\mu,\sigma^2),\ \theta\text{: } = (\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ k=1,2,\ldots,n,\ldots, \tag{2}$$

where "$\backsim$ NIID(μ,σ^2)" stands for "distributed as Normal, Independent and Identically Distributed, with mean μ and variance σ^2".

The statistical model $\mathcal{M}_\theta(\mathbf{x})$ plays a pivotal role in statistical inference in so far as it determines what constitutes *a legitimate*:

(a) Event — any well-behaved (Borel) functions of the sample $\mathbf{X}$ —
(b) Assignment of probabilities to legitimate events via $f(\mathbf{x};\theta)$
(c) Data $\mathbf{x}_0$ for inference purposes
(d) Hypothesis or inferential claim
(e) Optimal inference procedure and the associated error probabilities

Formally an event is legitimate when it belongs to the σ-field generated by $\mathbf{X}$ (Billingsley 1995). Legitimate data come in the form of data $\mathbf{x}_0$ that can be realistically viewed as a truly typical realization of the process $\{X_k,\ k\in\mathbb{N}\}$, as specified by $\mathcal{M}_\theta(\mathbf{x})$. Legitimate hypotheses and inferential claims are invariably about the data-generating mechanism and framed in terms of the unknown parameters θ. Moreover, the optimality (effectiveness) of the various inference procedures depends on the validity of the probabilistic assumptions constituting $\mathcal{M}_\theta(\mathbf{x})$; see Spanos (1999).

The interpretation of probability underlying frequentist inference associates probability with the *limit* of relative frequencies anchored on the Strong Law of Large Numbers (SLLN). "Stable relative frequencies" (Neyman 1952), i.e., one's that satisfy the SLLN, constitute a crucial feature of real-world phenomena we call stochastic. The *long-run* metaphor associated with this interpretation enables one to conceptualize probability in terms of viewing $\mathcal{M}_\theta(\mathbf{x})$, $\mathbf{x}\in\mathbb{R}_X^n$ as an idealized description of the data-generating mechanism. The appropriateness of this interpretation stems primarily from its capacity to facilitate the task of bridging the gap between stochastic phenomena and the mathematical underpinnings of $\mathcal{M}_\theta(\mathbf{x})$, as well as elucidate a number of issues pertaining to modeling and inference; see Spanos (2009).

Different Forms of Statistical Inference

Fisher (1925), almost single-handedly, put forward a frequentist theory of *optimal estimation,* and Neyman and Pearson (1933) modified Fisher's significance testing to propose an analogous theory for *optimal testing*; see Cox and Hinkley (1974). Optimality of inference in frequentist statistics is defined in terms of the capacity of different procedures to give rise to valid inferences, evaluated in terms of the associated *error probabilities*: how often these procedures lead to erroneous inferences.

The main forms of statistical inference in frequentist statistics are: (a) point estimation, (b) interval estimation, (c) hypothesis testing, and (d) prediction.
All these forms share the following features:

(a) Assume that the prespecified statistical model $\mathcal{M}_\theta(\mathbf{x})$ is valid vis-à-vis data $\mathbf{x}_0$.
(b) The objective of inference is always to learn about the underlying data-generating mechanism, and it is framed in terms of the unknown parameter(s) θ.
(c) An inference procedure is based on a *statistic* (estimator, test statistic, predictor), say $Y_n = g(X_1, X_2, \ldots, X_n)$,

whose sampling distribution provides the relevant error probabilities that calibrate its reliability. In principle, the sampling distribution of Y_n is derived via:

$$P(Y_n \leq y) = \underbrace{\int\int \cdots \int}_{\{\mathbf{x}:\, g(x_1,\ldots,x_n)\leq y\}} f(\mathbf{x};\theta)dx_1dx_2\cdots dx_n. \quad (3)$$

Point estimation centers on a mapping: $h(.): \mathbb{R}_X^n \to \Theta$, say $\widehat{\theta}_n(\mathbf{X}) = h(X_1, X_2, \ldots, X_n)$, known as an estimator of θ. The idea underlying optimal estimation is to select a mapping $h(.)$ that locates, as closely as possible, the *true* value of θ; whatever that happens to be. The qualification "as closely as possible" is quantified in terms of certain features of the sampling distribution of $\widehat{\theta}_n(\mathbf{X})$, known as estimation properties: unbiasedness, efficiency, sufficiency, consistency, etc.; see Cox and Hinkley (1974).

A *key concept* in Fisher's approach to inference is the *likelihood function*:

$$L(\theta;\mathbf{x}) = \ell(\mathbf{x}) \cdot f(\mathbf{x};\theta),\ \theta \in \Theta, \quad (4)$$

where $\ell(\mathbf{x}) > 0$ denotes a proportionality constant. Fisher (1922) defined the *Maximum Likelihood* (ML) estimator $\widehat{\theta}_{ML}(\mathbf{X})$ of θ to be the one that maximizes $L(\theta;\mathbf{x})$. He was also the first to draw a sharp distinction between the *estimator* $\widehat{\theta}(\mathbf{X})$ and the *estimate* $\widehat{\theta}(\mathbf{x}_0)$, and emphasized the importance of using the sampling distribution of $\widehat{\theta}(\mathbf{X})$ to evaluate the reliability of inference in terms of the relevant error probabilities.

Example In the case of the simple Normal model, the statistics:

$$\overline{X}_n = \frac{1}{n}\Sigma_{k=1}^n X_k \backsim \mathsf{N}\left(\mu, \frac{\sigma^2}{n}\right),$$
$$s^2 = \frac{1}{(n-1)}\Sigma_{k=1}^n (X_k - \overline{X}_n)^2 \backsim \left(\frac{\sigma^2}{n-1}\right)\chi^2(n-1), \quad (5)$$

where $\mathsf{N}(.,.)$ and $\chi^2(.)$ denote the Normal and chis-square distributions, constitute "good" estimators of (μ, σ^2) is terms of satisfying most of the above properties.

Point estimation is often considered *inadequate* for the purposes of scientific inquiry because a "good" point estimator $\widehat{\theta}_n(\mathbf{X})$, by itself, does not provide any measure of the reliability and precision associated with the estimate $\widehat{\theta}_n(\mathbf{x}_0)$. This is the reason why $\widehat{\theta}_n(\mathbf{x}_0)$ is often accompanied by some significance test result (e.g., p-value) associated with the *generic* hypothesis $\theta = 0$.

Interval estimation rectifies this crucial weakness of point estimation by providing the relevant error probabilities associated with inferences pertaining to "covering" the true value of θ. This comes in the form of the Confidence Interval (CI):

$$\mathbb{P}(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) = 1 - \alpha, \quad (6)$$

where the statistics $L(\mathbf{X})$ and $U(\mathbf{X})$ denote the lower and upper (random) bounds that "covers" the true value θ^* with probability $(1-\alpha)$, or equivalently, the "coverage error" probability is α.

Example In the case of the simple Normal model:

$$\mathbb{P}\left(\overline{X}_n - c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{X}_n + c_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)\right) = 1 - \alpha, \quad (7)$$

provides a $(1-\alpha)$ Confidence Interval (CI) for μ. The evaluation of the coverage probability $(1-\alpha)$ is based on the following sampling distribution result:

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s} \backsim \mathsf{St}(n-1), \quad (8)$$

where $\mathsf{St}(n-1)$ denotes the Student's t distribution with $(n-1)$ degrees of freedom, attributed to Gosset (1908).

What is often not appreciated sufficiently about estimation in general, and CIs in particular, is the underlying reasoning that gives rise to sampling distribution results such as (5) and (8). The reasoning that underlies estimation is *factual,* based on evaluating the relevant sampling distributions "under the True State of Nature" (TSN), i.e., the *true* data-generating mechanism: $\mathcal{M}^*(\mathbf{x}) = \{f(\mathbf{x};\theta^*)\}$, $\mathbf{x} \in \mathbb{R}_X^n$, where θ^* denotes the true value of the unknown parameter(s) θ. Hence, the generic CI in (6) is more accurately stated as:

$$\mathbb{P}\left(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X});\ \theta = \theta^*\right) = 1 - \alpha, \quad (9)$$

where $\theta = \theta^*$ denotes 'evaluated under the TSN'. The remarkable thing about factual reasoning is that one can make probabilistic statements like (9), with a precise error probability (α), *without* knowing the true θ^*.

Example In the case of the simple Normal model, the distributional results (5) and (8) are more accurately stated as:

$$\overline{X}_n \overset{\text{TSN}}{\backsim} \mathsf{N}\left(\mu_*, \frac{\sigma_*^2}{n}\right), \quad \frac{(n-1)s^2}{\sigma_*^2} \overset{\text{TSN}}{\backsim} \chi^2(n-1),$$
$$\frac{\sqrt{n}(\overline{X}_n - \mu^*)}{s} \overset{\text{TSN}}{\backsim} \mathsf{St}(n-1), \quad (10)$$

where $\theta^*{:=} (\mu_*, \sigma_*^2)$ denote the "true" values of the unknown parameters $\theta{:=}(\mu, \sigma^2)$.

Prediction is similar to estimation in terms of its underlying factual reasoning, but it differs from it in so far as it is concerned with finding the most representative

value of X_k beyond the observed data, say X_{n+1}. An optimal predictor of X_{n+1} is given by:

$$\widehat{X}_{n+1} = \overline{X}_n, \tag{11}$$

whose reliability can be calibrated using the sampling distribution of the prediction error:

$$\widehat{u}_{n+1} = (X_{n+1} - \overline{X}_n) \overset{\text{TSN}}{\backsim} \mathsf{N}\left(0, \sigma_*^2\left(1+\frac{1}{n}\right)\right), \tag{12}$$

to construct a $(1-\alpha)$ prediction interval:

$$\mathbb{P}\left(\overline{X}_n - c_{\frac{\alpha}{2}}\left(s\sqrt{\left(1+\frac{1}{n}\right)}\right) \leq X_{n+1} \leq \overline{X}_n + c_{\frac{\alpha}{2}}\left(s\sqrt{\left(1+\frac{1}{n}\right)}\right); \theta=\theta^*\right) = 1-\alpha. \tag{13}$$

Hypothesis testing. In contrast to estimation, the reasoning underlying hypothesis testing is *hypothetical*. The sampling distribution of a test statistic is evaluated under several hypothetical scenarios concerning the statistical model $\mathcal{M}_\theta(\mathbf{x})$, referred to as "under the null" and "under the alternative" hypotheses of interest.

Example Consider testing the hypotheses in the context of (2):

$$H_0\text{: } \mu \leq \mu_0 \text{ vs. } H_1\text{: } \mu > \mu_0. \tag{14}$$

What renders the hypotheses in (14) legitimate is that: (a) they pose questions concerning the underlying data-generating mechanism, (b) they are framed in terms of the unknown parameter θ, and (c) in a way that partitions $\mathcal{M}_\theta(\mathbf{x})$. In relation to (c), it is important to stress that even in cases where substantive information excludes or focuses exclusively on certain subsets (or values) of the parameter space, the entire Θ is relevant for statistical inference purposes. Ignoring this, and focusing only on the substantively relevant subsets of Θ, gives rise to fallacious results.

The N-P test for the hypotheses (14) $T_\alpha := \{\tau(\mathbf{X}), C_1(\alpha)\}$, where:

$$\tau(\mathbf{X}) = \frac{\sqrt{n}(\overline{X}_n-\mu_0)}{s},\ C_1(\alpha) = \{\mathbf{x}\text{: } d(\mathbf{x}) > c_\alpha\}, \tag{15}$$

can be shown to be Uniformly Most Powerful (UMP) in the sense that, its type I error probability (significance level) is:

$$\begin{aligned}(\text{a})\ \ \alpha &= \max_{\mu\leq\mu_0}\mathbb{P}(\mathbf{x}\text{: } \tau(\mathbf{X}) > c_\alpha; H_0)\\ &= \mathbb{P}(\mathbf{x}\text{: } \tau(\mathbf{X}) > c_\alpha; \mu=\mu_0),\end{aligned} \tag{16}$$

and among all the α-level tests T_α has highest *power* (Lehmann 1986):

$$\begin{aligned}(\text{b})\ &\mathbb{P}(\mathbf{x}: \tau(\mathbf{X}) > c_\alpha; \mu=\mu_1), \textit{ for all } \mu_1 > \mu_0,\\ &\mu_1 = \mu_0+\gamma,\ \gamma \geq 0;\end{aligned} \tag{17}$$

In this sense, a UMP test provides the most effective α-level probing procedure for detecting any discrepancy ($\gamma \geq 0$) of interest from the null.

To evaluate the error probabilities in (16) and (17) one needs to derive the sampling distribution of $\tau(\mathbf{X})$ under several *hypothetical* values of μ relating to (14):

$$(\text{a})\ \tau(\mathbf{X}) \overset{\mu=\mu_0}{\backsim} \mathsf{St}(n-1),\quad (\text{b})\ \tau(\mathbf{X}) \overset{\mu=\mu_1}{\backsim} \mathsf{St}(\delta(\mu_1); n-1),\quad \text{for any } \mu_1 > \mu_0, \tag{18}$$

where $\delta(\mu_1) = \frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$ is known as the non-centrality parameter. The sampling distribution in (18a) is also used to evaluate Fisher's (1935) p-value:

$$p(\mathbf{x}_0) = \mathbb{P}(\mathbf{x}\text{: } \tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu=\mu_0), \tag{19}$$

where a small enough $p(\mathbf{x}_0)$ can be interpreted as indicating discordance with H_0.

Remark It is unfortunate that most statistics books use the vertical bar (|) instead of the semi-colon (;) in formulae (16)–(17) to denote the evaluation *under* H_0 or H_1, as it relates to (18), encouraging practitioners to misinterpret error probabilities as being *conditional* on H_0 or H_1; see Cohen (1994). It is worth emphasizing these error probabilities are: (1) never conditional, (2) always assigned to inference procedures (never to hypotheses), and (3) invariably depend on the sample size $n > 1$.

Comparing the sampling distributions in (18) with those in (10) brings out the key difference between hypothetical and factual reasoning: in the latter case there is only one unique scenario, but in hypothetical reasoning there is usually an infinity of scenarios. The remarkable thing about hypothetical reasoning is that one can pose sharp questions by comparing $\mathcal{M}_\theta(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}_X^n$, for different hypothetical values of θ, with $\mathcal{M}^*(\mathbf{x}_0)$, to learn about $\mathcal{M}^*(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}_X^n$. This often elicits more informative answers from $\mathbf{x}_0$ than factual reasoning. This difference is important in understanding the nature of the error probabilities associated with each type of inference as well as in interpreting the results of these procedures.

In particular, factual reasoning can only be used pre-data to generate the relevant error probabilities, because when data $\mathbf{x}_0$ is observed (i.e., post-data) the unique factual scenario has been realized and the sampling distribution in question becomes degenerate. This is the reason why the p-value in (19) is a well-defined post-data error probability, but one cannot attach error probabilities to an observed CI: $(L(\mathbf{x}_0) \leq \theta \leq U(\mathbf{x}_0))$; see the exchange between Fisher (1955) and Neyman (1956). In contrast, the scenarios in hypothetical reasoning are equally relevant to both pre-data and post-data assessments. Indeed, one can go a long

way towards delineating some of the confusions surrounding frequentist testing, as well as addressing some of the criticisms leveled against it – statistical vs. substantive significance, with a large enough n one can reject any null hypothesis, no evidence against the null is *not* evidence for it – using post-data error probabilities to provide an evidential interpretation of frequentist testing based on the severity rationale; see Mayo and Spanos (2006) for further discussion.

Bayesian Inference

Bayesian inference also begins with a prespecified statistical model $\mathcal{M}_\theta(\mathbf{x})$, as specified in (1), but modifies it in three crucial respects:

(1) Probability is now interpreted as (subjective or rational) *degrees of belief* (not as the limit of relative frequencies).
(2) The unknown parameter(s) θ are now viewed as *random variables* (not as constants) with their own distribution $\pi(\theta)$, known as the *prior distribution*.
(3) The distribution of the sample is now viewed as *conditional* on θ, and denoted by $f(\mathbf{x} \mid \theta)$ instead of $f(\mathbf{x};\theta)$.

All three of these modifications have been questioned in the statistics literature, but the most prominent controversies concern the nature and choice of the prior distribution. There are ongoing disputes concerning subjective vs. default (reference) priors, informative vs. non-informative (invariant) priors, proper vs. improper priors, conjugate vs. non-conjugate, matching vs. non-matching priors, and how should these choices be made in practice; see Kass and Wasserman (1996) and Roberts (2007).

In light of these modifications, one can use the definition of conditional probability distribution between two jointly distributed random vectors, say (Z,W):

$$f(\mathbf{z} \mid \mathbf{w}) = \frac{f(\mathbf{z},\mathbf{w})}{f(\mathbf{w})} = \frac{f(\mathbf{z},\mathbf{w})}{\int_{\mathbf{z}} f(\mathbf{z},\mathbf{w})d\mathbf{z}} = \frac{f(\mathbf{w} \mid \mathbf{z})f(\mathbf{z})}{\int_{\mathbf{z}} f(\mathbf{w} \mid \mathbf{z})f(\mathbf{z})d\mathbf{z}},$$

to define *Bayes formula* that determines the *posterior* distribution of θ:

$$\pi(\theta \mid \mathbf{x}_0) = \frac{f(\mathbf{x}_0|\theta)\cdot\pi(\theta)}{\int_\theta f(\mathbf{x}_0|\theta)\cdot\pi(\theta)d\theta} \propto \pi(\theta)\cdot L(\theta \mid \mathbf{x}_0),\ \theta \in \Theta, \tag{20}$$

where $L(\theta \mid \mathbf{x}_0)$ denotes the *reinterpreted* likelihood function, not (4).

Bayesian inference is based exclusively on the posterior distribution $\pi(\theta \mid \mathbf{x}_0)$ which is viewed as the *revised* (from the initial $\pi(\theta)$) degrees of belief for different values of θ in light of the summary of the data by $L(\theta \mid x_0)$. A Bayesian point estimate of θ specified by selecting the *mean* ($\widehat{\theta}_B(\mathbf{x}_0) = E(\pi(\theta \mid \mathbf{x}_0))$) or the *mode* of the posterior. A Bayesian interval estimate for θ is given by finding two values $a < b$ such that:

$$\int_a^b \pi(\theta \mid \mathbf{x}_0)d\theta = 1-\alpha, \tag{21}$$

known as a $(1-\alpha)$ *posterior* (or credible) *interval*.

Bayesian *testing of hypotheses* is more difficult to handle in terms of the posterior distribution, especially for point hypotheses, because of the technical difficulty in attaching probabilities to particular values of θ, since the parameter space Θ is usually *uncountable*. There have been numerous attempts to address this difficulty, but no agreement seems to have emerged; see Roberts (2007). Assuming that one adopts his/her preferred way to sidestep this difficulty, Bayesian testing for H_0: $\theta = \theta_0$ vs. H_1: $\theta = \theta_1$ relies on comparing their respective degrees of belief using the *posterior ratio*:

$$\frac{\pi(\theta_0|\mathbf{x}_0)}{\pi(\theta_1|\mathbf{x}_0)} = \frac{L(\theta_0|\mathbf{x}_0)\cdot\pi(\theta_0)}{L(\theta_1|\mathbf{x}_0)\cdot\pi(\theta_1)}, \tag{22}$$

or, its more widely used modification in the form of the *Bayes Factor* (BF):

$$BF(\mathbf{x}_0) = \left(\frac{\pi(\theta_0|x_0)}{\pi(\theta_1|\mathbf{x}_0)}\right) / \left(\frac{\pi(\theta_0)}{\pi(\theta_1)}\right) = \frac{L(\theta_0|\mathbf{x}_0)}{L(\theta_1|\mathbf{x}_0)}, \tag{23}$$

together with certain rules of thumb, concerning the *strength* of the degrees of belief *against* H_0 based on the magnitude of $\ln BF(\mathbf{x}_0)$: for $0 \leq \ln BF(\mathbf{x}_0) \leq .5$, $.5 < \ln BF(\mathbf{x}_0) \leq 1$, $1 < \ln BF(\mathbf{x}_0) \leq 2$ and $\ln BF(\mathbf{x}_0) > 2$, the degree of belief against H_0 is *poor*, *substantial*, *strong* and *decisive*, respectively; see Roberts (2007). Despite their intuitive appeal, these rules of thumb have been questioned by Kass and Raftery (1995) inter alia.

The question that naturally arises at this stage concerns the nature of the reasoning underlying *Bayesian inference*. In Bayesian inference *learning* is about revising one's degrees of belief pertaining to $\theta \in \Theta$, from $\pi(\theta)$ (pre-data) to $\pi(\theta \mid \mathbf{x}_0)$ (post-data). In contrast to frequentist inference — which pertains to the *true* data-generating mechanism $\mathcal{M}^*(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}_X^n$ — Bayesian inference is concerned with more or less appropriate (in terms of $\pi(\theta \mid \mathbf{x}_0)$) models within $\mathcal{M}_\theta(\mathbf{x}_0)$, $\theta \in \Theta$. In terms of the underlying reasoning the Bayesian is similar to the decision theoretic inference which is also about selecting among more or less cost (or utility)-appropriate models. This questions attempts to present N-P testing as naturally belonging to the decision theoretic approach.

The problem with the inference not pertaining to the underlying data-generating mechanism can be brought out more clearly when Bayesian inference is viewed in

the context of the broader scientific inquiry. In that context, one begins with substantive questions pertaining to the phenomenon of interest, and the objective is to learn about the phenomenon itself. Contrasting frequentist with Bayesian inference, using interval estimation as an example, Wasserman (2008) argued: "Frequentist methods have coverage guarantees; Bayesian methods don't. In science, coverage matters" (p. 463).

About the Author

Dr. Aris Spanos is Professor of Economics and former Chair of the Department of Economics (2001–2006) at Virginia Tech, USA. Previously he has taught at London University (England), Cambridge University (England), University of California (USA) and the University of Cyprus. He is the author of two textbooks entitled: *Statistical Foundations of Econometric Modelling* (Cambridge University Press, 1986), and *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data* (Cambridge University Press, 1999). He has published over 70 papers in leading econometric, economic, philosophical and statistical journals.

Cross References

►Bayes' Theorem
►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Nonparametric Statistics
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bayesian vs. Classical Point Estimation: A Comparative Overview
►Causation and Causal Inference
►Confidence Interval
►Degrees of Freedom in Statistical Inference
►Empirical Likelihood Approach to Inference from Sample Survey Data
►Estimation
►Estimation: An Overview
►Exact Inference for Categorical Data
►Fiducial Inference
►Frequentist Hypothesis Testing: A Defense
►Generalized Quasi-Likelihood (GQL) Inferences
►Inference Under Informative Probability Sampling
►Likelihood
►Multi-Party Inference and Uncongeniality
►Neyman-Pearson Lemma
►Nonparametric Predictive Inference
►Nonparametric Statistical Inference
►Null-Hypothesis Significance Testing: Misconceptions
►Optimal Statistical Inference in Financial Engineering
►Parametric Versus Nonparametric Tests
►Philosophical Foundations of Statistics
►Proportions, Inferences, and Comparisons
►P-Values
►Ranking and Selection Procedures and Related Inference Problems
►Robust Inference
►Sampling Distribution
►Significance Testing: An Overview
►Significance Tests: A Critique
►Statistical Evidence
►Statistical Inference
►Statistical Inference for Quantum Systems
►Statistical Inference for Stochastic Processes
►Statistical Inference in Ecology

References and Further Reading

Billingsley P (1995) Probability and measure, 4th edn. Wiley, New York
Cohen J (1994) The earth is round (p < .05). Am Psychol 49:997–1003
Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman & Hall, London
Dodge Y (ed) (2003) The Oxford dictionary of statistical terms. The International Statistical Institute, Oxford University Press, Oxford
Doob JL (1953) Stochastic processes. Wiley, New York
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans Roy Soc A 222:309–368
Fisher RA (1925) Theory of statistical estimation. Proc Cambridge Philos Soc 22:700–725
Fisher RA (1935) The design of experiments. Oliver & Boyd, Edinburgh
Fisher RA (1955) Statistical methods and scientific induction. J Roy Stat Soc B 17:69–78
Gosset WS (1908) The probable error of the mean, Biometrika 6:1–25
Kass RE, Raftery AE (1995) Bayes factor and model uncertainty. J Am Stat Assoc 90:773–795
Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules. J Am Stat Assoc 91:1343–1370
Lehmann EL (1986) Testing statistical hypotheses, 2nd edn. Wiley, New York
Lehmann EL (1990) Model specification: the views of Fisher and Neyman, and later developments. Stat Sci 5:160–168
Mayo DG (1996) Error and the growth of experimental knowledge, The University of Chicago Press, Chicago
Mayo DG, Spanos A (2006) Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. Br J Philos Sci 57:323–357
Neyman J (1952) Lectures and conferences on mathematical statistics and probability, 2nd edn. U.S. Department of Agriculture, Washington, DC
Neyman J (1956) Note on an article by Sir Ronald Fisher. J Roy Stat Soc B 18:288–294
Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. Philos Trans Roy Soc A 231: 289–337
Pearson K (1920) The fundamental problem of practical statistics. Biometrika XIII:1–16

Roberts CP (2007) The Bayesian choice, 2nd edn. Springer, New York
Spanos A (1999) Probability theory and statistical inference: econometric modeling with observational data. Cambridge University Press, Cambridge
Spanos A (2007) Curve-Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach, Philosophy of Science, 74:1046–1066
Spanos A (2009) Model-based inference and the frequentist interpretation of probability. Working Paper, Virginia Tech
Student (pseudonym for Gosset W) (1908) The probable error of the mean. Biometrika 6:1–25
Wasserman L (2008) Comment on article by Gelman. Bayesian Anal 3:463–466

Statistical Literacy, Reasoning, and Thinking

Joan Garfield
Professor
University of Minnesota, Minneapolis, MN, USA

Statistics educators often talk about their desired learning goals for students, and invariably, refer to outcomes such as being statistically literate, thinking statistically, and using good statistical reasoning. Despite the frequent reference to these outcomes and terms, there have been no agreed upon definitions or distinctions. Therefore, the following definitions were proposed by Garfield (2005 and have been elaborated in Garfield and Ben-Zvi (2008).

Statistical literacy is regarded as a key ability expected of citizens in information-laden societies, and is often touted as an expected outcome of schooling and as a necessary component of adults' numeracy and literacy. Statistical literacy involves understanding and using the basic language and tools of statistics: knowing what basic statistical terms mean, understanding the use of simple statistical symbols, and recognizing and being able to interpret different representations of data (Garfield 1999; Rumsey 2002; Snell 1999).

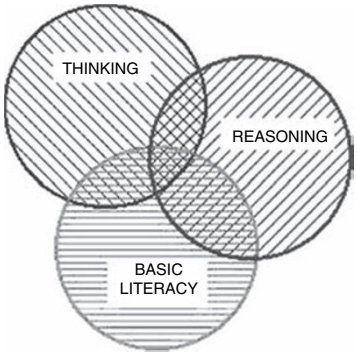
There are other views of statistical literacy such as Gal's (2000, 2002), whose focus is on the data consumer: Statistical literacy is portrayed as the ability to interpret, critically evaluate, and communicate about statistical information and messages. Gal (2002) argues that statistically literate behavior is predicated on the joint activation of five interrelated knowledge bases (literacy, statistical, mathematical, context, and critical), together with a cluster of supporting dispositions and enabling beliefs. Watson and Callingham (2003) proposed and validated a model of three levels of statistical literacy (knowledge of terms, understanding of terms in context, and critiquing claims in the media).

Statistical reasoning is the way people reason with statistical ideas and make sense of statistical information. Statistical reasoning may involve connecting one concept to another (e.g., understanding the relationship between the mean and standard deviation in a distribution) or may combine ideas about data and chance (e.g., understanding the idea of confidence when making an estimate about a population mean based on a sample of data). Statistical reasoning also means understanding and being able to explain statistical processes, and being able to interpret statistical results (Garfield 2002). For example, being able to explain the process of creating a sampling distribution for a statsistics and why this distribution has particular features. Statistical reasoning invovles the mental representations and connections that students have regarding statistical concepts. Another examples is being able to see how and why an outlier makes the mean and standard deviation larger than when that outlier is removed, or reasoning about the effect of an influential data value on the correlation coefficient.

Statistical thinking involves a higher order of thinking than statistical reasoning. Statistical thinking is the way professional statisticians think (Wild and Pfannkuch 1999). It includes knowing how and why to use a particular method, measure, design or statistical model; deep understanding of the theories underlying statistical processes and methods; as well as understanding the constraints and limitations of statistics and statistical inference. Statistical thinking is also about understanding how statistical models are used to simulate random phenomena, understanding how data are produced to estimate probabilities, recognizing how, when, and why existing inferential tools can be used, and being able to understand and utilize the context of a problem to plan and evaluate investigations and to draw conclusions (Chance 2002). Finally, statistical thinking is the normative use of statistical models, methods, and applications in considering or solving statistical problems.

Statistical literacy, reasoning, and thinking are unique learning outcomes, but there is some overlap as well as a type of hierarchy, where statistical literacy provides the foundation for reasoning and thinking (see Fig. 1). A summary of additional models of statistical reasoning and thinking can be found in Jones et al. (2004).

There is a growing network of researchers who are interested in studying the development of students' statistical literacy, reasoning, and thinking (e.g., SRTL – The



Statistical Literacy, Reasoning, and Thinking. Fig. 1 The overlap and hierarchy of statistical literacy, reasoning, and thinking (Artist Website, https://app.gen.umn.edu/artist)

International Statistical Reasoning, Thinking, and Literacy Research Forums, http://srtl.stat.auckland.ac.nz/). The topics of the research studies conducted by members of this community reflect a shift in emphasis in statistics instruction, from developing procedural understanding, i.e., statistical techniques, formulas, computations and procedures; to developing conceptual understanding and statistical literacy, reasoning, and thinking.

Words That Characterize Assessment Items for Statistical Literacy, Reasoning, and Thinking

One way to distinguish between these related outcomes is by examining the types of words used in assessment of each outcome. Table 1 (modified from delMas (2002)) lists words associated with different assessment items or tasks.

Statistical Literacy, Reasoning, and Thinking. Table. 1 Typical words associated with different assessment items or tasks

Basic Literacy	Reasoning	Thinking
Identify	Explain why	Apply
Describe	Explain how	Critique
Translate		Evaluate
Interpret		Generalize
Read		
Compute		

The following three examples (from Garfield and Ben-Zvi 2008) illustrate how statistical literacy, reasoning, and thinking may be assessed.

Example of an Item Designed to Measure Statistical Literacy

A random sample of 30 first-year students was selected at a public university to estimate the average score on a mathematics placement test that the state mandates for all freshmen. The average score for the sample was found to be 81.7 with a sample standard deviation of 11.45. Describe to someone who has not studied statistics what the standard deviation tells you about the variability of placement scores for this sample.

This item assesses statistical literacy because it focuses on understanding (knowing) what the term "standard deviation" means.

Example of an Item Designed to Measure Statistical Reasoning

The following stem plot displays the average annual snowfall amounts (in inches, with the stems being tens and leaves being ones) for a random sample of 25 American cities:

0	000000024
1	028
2	00228
3	8
4	2248
5	48
6	0

Without doing any calculations, would you expect the mean of the snowfall amounts to be larger, smaller, or about the same as the median? Why?

This item assess statistical reasoning because students need to connect and reason about how shape of a distribution affects the relative locations of measures of center, in

this case, reasoning that the mean would be larger than the mean because of the positive skew.

Example of an Item Designed to Assess Statistical Thinking

A random sample of 30 first year students was selected at a public university to estimate the average score on a mathematics placement test that the state mandates for all freshmen. The average score for the sample was found to be 81.7 with a sample standard deviation of 11.45.

A psychology professor at a state college has read the results of the university study. The professor wants to know if students at his college are similar to students at the university with respect to their mathematics placement exam scores. This professor collects information for all 53 first year students enrolled this semester in a large section (321 students) of his "Introduction to Psychology" course. Based on this sample, he calculates a 95% confidence interval for the average mathematics placement scores exam to be 69.47 to 75.72. Below are two possible conclusions that the psychology professor might draw. For each conclusion, state whether it is valid or invalid. Explain your choice for both statements. Note that it is possible that neither conclusion is valid.

(a) The average mathematics placement exam score for first year students at the state college is lower than the average mathematics placement exam score of first year students at the university.
(b) The average mathematics placement exam score for the 53 students in this section is lower than the average mathematics placement exam score of first year students at the university.

This item assesses statistical thinking because it asks students to think about the entire process involved in this research study in critiquing and justifying different possible conclusions.

Comparing Statistical Literacy, Reasoning, and Thinking to Bloom's Taxonomy

These three statistics learning outcomes also seem to coincide somewhat with Bloom's more general categories of learning outcomes (1956). In particular, some current measurement experts feel that Bloom's taxonomy is best used if it is collapsed into three general levels (knowing, comprehending, and applying). Statistical literacy may be viewed as consistent with the "knowing" category, statistical reasoning as consistent with the "comprehending" category (with perhaps some aspects of application and analysis) and statistical thinking as encompassing many elements of the top three levels of Bloom's taxonomy (application, analysis, and synthesis).

About the Author

Dr. Joan Garfield is Professor of Educational Psychology and Head of a unique graduate program in Statistics Education at the University of Minnesota, USA. She is Associate Director for Research and co-founder of the Consortium for the Advancement of Undergraduate Statistics Education (CAUSE), Past Chair of the American Statistical Association Section on Statistical Education (ASA), and past Vice President of the International Association for Statistical Education. She has co-authored or co-edited 5 books including *Developing Students' Statistical Reasoning: Connecting Research and Teaching practice* (Garfield and Ben-Zvi, Springer, 2008) as well as numerous journal articles. Professor Garfield has received the ASA Founders' Award, the CAUSE Lifetime Achievement Award, is a fellow of ASA and AERA, and has received both Post- baccalaureate and Undergraduate Outstanding Teaching awards given by the University of Minnesota. She has helped found three journals in statistics Education (JSE, SERJ and TISE) and currently serves as Associate Editor for SERJ and TISE. Finally, she is co-founder and co-chair of the biennial International Research Forum on Statistical Reasoning, Thinking, and Literacy.

Cross References

►Decision Trees for the Teaching of Statistical Estimation
►Learning Statistics in a Foreign Language
►Online Statistics Education
►Promoting, Fostering and Development of Statistics in Developing Countries
►Role of Statistics in Advancing Quantitative Education
►Statistical Consulting
►Statistics Education

References and Further Reading

Bloom BS (ed) (1956) Taxonomy of educational objectives: the classification of educational goals: handbook I, cognitive domain. Longmans, Green, New York

Chance BL (2002) Components of statistical thinking and implications for instruction and assessment. J Stat Educ 10(3), from http://www.amstat.org/publications/jse/v10n3/chance.html Retrieved 15 July2007

delMas RC (2002) Statistical literacy, reasoning, and learning: a commentary. J Stat Educ 10(3), from http://www.amstat.org/publications/jse/v10n3/delmas_intro.html. Retrieved 6 November 2006

Gal I (ed) (2000) Adult numeracy development: theory, research, practice. Hampton Press, Cresskill, NJ
Gal I (2002) Adults' statistical literacy: meaning, components, responsibilities. Int Stat Rev 70(1):1–25
Garfield J (1999) Thinking about statistical reasoning, thinking, and literacy. Paper presented at the first international research forum on statistical reasoning, thinking, and literacy (STRL-1), Kibbutz Be'eri, Israel
Garfield J (2002) The challenge of developing statistical reasoning. J Stat Educ 10(3), from http://www.amstat.org/publications/jse/v10n3/garfield.html Retrieved 15 July 2007
Garfield J, delMas R, Chance B (2005) The Web-Based Assessment Resource for Improving Statistics Thinking (ARTIST) Project. Project funded by the National Science Foundation. Accessed 1 Aug 2010
Garfield J, Ben-Zvi D (2008) Developing students' statistical reasoning: connecting research and teaching practice. Springer, Dordrecht, The Netherlands
Jones GA, Langrall CW, Mooney ES, Thornton CA (2004) Models of development in statistical reasoning. In: Ben-Zvi D, Garfield J (eds) The challenge of developing statistical literacy, reasoning, and thinking. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp 97–117
Rumsey DJ (2002) Statistical literacy as a goal for introductory statistics courses. J Stat Educ 10(3), from http://www.amstat.org/publications/jse/v10n3/rumsey2.html Retrieved 15 July 2007
Snell L (1999) Using chance media to promote statistical literacy. Paper presented at the 1999 Joint Statistical Meetings, Dallas, TX
Watson JM, Callingham R (2003) Statistical literacy: a complex hierarchical construct. Stat Educ Res J 2:3–46, from http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(2)_Watson_Callingham.pdf Retrieved 26 April 2008
Wild CJ, Pfannkuch M (1999) Statistical thinking in empirical enquiry. Int Stat Rev 67(3):223–265

Statistical Methods for Non-Precise Data

Reinhard Viertl
Professor and Head
Vienna University of Technology, Vienna, Austria

Non-Precise Data

Real data obtained from measurement processes are not precise numbers or vectors, but more or less non-precise, also called fuzzy. This uncertainty is different from measurement errors and has to be described formally in order to obtain realistic results from data analysis. A real life example is the water level of a river at a fixed time. It is typically not a precise multiple of the scale unit for height measurements. In the past this kind of uncertainty was mostly neglected in describing such data. The reason for that is the idea of the existence of a "true" water level which is identified with a real number times the measurement unit. But this is not realistic. The formal description of such non-precise water levels can be given using the intensity of the wetness of the gauge to obtain the so called *characterizing functions* from the next section. Further examples of non-precise data are readings on digital measurement equipments, readings of pointers on scales, color intensity pictures, and light points on screens.

Remark 1 Non-precise data are different from measurement errors because in error models the observed values y_i are considered to be numbers, i.e., $y_i = x_i + \epsilon_i$, where ϵ_i denotes the error of the i-th observation.

Historically non-precise data were not studied sufficiently. Some earlier work was done in interval arithmetics. General non-precise data in form of so called fuzzy numbers were considered in the 1980s and first publications combining fuzzy imprecision and stochastic uncertainty came up, see Kacprzyk and Fedrizzi (1988). Some of these approaches are more theoretically oriented. An applicable approach for statistical analysis of non-precise data is given in Viertl (1996).

Characterizing Functions of Non-Precise Data

In case of measurements of one-dimensional quantities non-precise observations can be reasonably described by so-called *fuzzy numbers* $x^\star$. Fuzzy numbers are generalizations of real numbers in the following sense. Each real number $x \in I\!R$ is characterized by its indicator function $I_{\{x\}}(\cdot)$. A fuzzy number is characterized by its so-called *characterizing function* $\xi(\cdot)$ which is a generalization of an indicator function. A characterizing function is a real function of a real variable obeying the following:

1. $\xi : I\!R \longrightarrow [0,1]$
2. $\forall\, \delta \in (0,1]$ the so called δ-cut $C_\delta(x^\star) := \{x \in I\!R : \xi(x) \geq \delta\}$ is a non-empty and closed bounded interval

Remark 2 A characterizing function is describing the imprecision of *one* observation. It should not be confused with a probability density which is describing the stochastic variation of a random quantity X.

A fundamental problem is how to obtain the characterizing function of a non-precise observation. This depends on the area of application. Some examples can be given.

Example 1 For data in form of gray intensities in one dimension as boundaries of regions the gray intensity $g(x)$ as an increasing function of one real variable x can be used to obtain the characterizing function $\xi(\cdot)$ in the following way. Take the derivative $\frac{d}{dx}g(x)$ and divide it by its maximum then the resulting function or its convex hull can be used as characterizing function of the non-precise observation.

Non-Precise Samples

Taking observations of a one-dimensional continuous quantity X in order to estimate the distribution of X usually a finite sequence $x_1^\star, \cdots, x_n^\star$ of non-precise numbers is obtained. These non-precise data are given in form of n characterizing functions $\xi_1(\cdot), \cdots, \xi_n(\cdot)$ corresponding to $x_1^\star, \cdots, x_n^\star$. Facing this kind of samples even the most simple concepts like *histograms* have to be modified. This is necessary by the fact that for a given class K_j of a histogram in case of a non-precise observation $x_i^\star$ with characterizing function $\xi_i(\cdot)$ obeying $\xi_i(x) > 0$ for an element $x \in K_j$ and $\xi_i(y) > 0$ for an element $y \in K_j^c$ it is not possible to decide if $x_i^\star$ is an element of K_j or not.

A generalization of the concept of histograms is possible by so-called *fuzzy histograms*. For those histograms the height of the histogram over a fixed class K_j is a fuzzy number $h_j^\star$. For the definition of the characterizing function of $h_j^\star$ compare Viertl (2006). For other concepts of statistics in case of non-precise data compare Viertl (2006).

Fuzzy Vectors

In case of multivariate continuous data $\boldsymbol{x} = (x_1, \cdots, x_n)$, for example the position of an object on a radar screen, the observations are non-precise vectors $\boldsymbol{x}^\star$. Such non-precise vectors are characterized by so called *vector-characterizing functions* $\zeta_{\boldsymbol{x}^\star}(\cdot, \cdots, \cdot)$. These vector-characterizing functions are real functions of n real variables $x_1, \cdots, x_n$ obeying the following:

(1) $\zeta_{\boldsymbol{x}^\star} : I\!R^n \longrightarrow [0,1]$

(2) $\forall\, \delta \in (0,1]$ the δ-cut $C_\delta(\boldsymbol{x}^\star) := \{\boldsymbol{x} \in I\!R^n : \zeta_{\boldsymbol{x}^\star}(\boldsymbol{x}) \geq \delta\}$ is a non-empty, closed and star shaped subset of $I\!R^n$ with finite n-dimensional content

In order to generalize statistics $t(x_1, \cdots, x_n)$ to the situation of fuzzy data the fuzzy sample has to be combinded into a fuzzy vector called *fuzzy combined sample*.

Generalized Classical Inference

Based on combined fuzzy samples point estimators for parameters can be generalized using the so-called *extension principle* from fuzzy set theory. If $\vartheta(x_1, \cdots, x_n)$ is a classical point estimator for θ, then $\vartheta\left(x_1^\star, \cdots, x_n^\star\right) = \vartheta(\boldsymbol{x}^\star)$ yields a fuzzy element $\hat{\theta}^\star$ of the parameter space Θ.

Generalized confidence regions for θ can be constructed in the following way. Let $\kappa(x_1, \cdots, x_n)$ be a classical confidence function for θ with coverage probability $1 - \alpha$, i.e., $\Theta_{1-\alpha}$ is the corresponding confidence set. For fuzzy data $x_1^\star, \cdots, x_n^\star$ a generalized confidence set $\Theta_{1-\alpha}^\star$ is defined as the fuzzy subset of Θ whose membership function $\varphi(\cdot)$ is given by its values

$$\varphi(\theta) = \left\{ \begin{array}{ll} \sup\left\{\zeta(\boldsymbol{x}) : \boldsymbol{x} \in M_X^n, \theta \in \kappa(\boldsymbol{x})\right\} & \text{if } \exists\, \boldsymbol{x} : \theta \in \kappa(\boldsymbol{x}) \\ 0 & \text{if } \nexists\, \boldsymbol{x} : \theta \in \kappa(\boldsymbol{x}) \end{array} \right\} \quad \forall\, \theta \in \Theta.$$

Statistical tests are mostly based on so-called *test statistics* $t(x_1, \cdots, x_n)$. For non-precise data the values $t\left(x_1^\star, \cdots, x_n^\star\right)$ become non-precise numbers. Therefore test decisions are not as simple as in the classical (frequently artificial) situation. There are different generalizations possible. Also in case of non-precise values of the test statistic it is possible to find ▶*p*-values and the test decision is possible similar to the classical case. Another possibility is to define fuzzy *p*-values which seems to be more problem adequate. For details see Viertl (2006).

There are other approaches for the generalization of classical inference procedures to the situation of fuzzy data. References for that are Gil et al. (1988) and Näther (1997).

Generalized Bayesian Inference

In Bayesian inference for non-precise data, besides the imprecision of data there is also imprecision of the a-priori distribution. So ▶Bayes' theorem is generalized in order to take care of this. The result of this generalized Bayes' theorem is a so-called *fuzzy a-posteriori distribution* $\pi^\star\left(\cdot \mid x_1^\star, \cdots, x_n^\star\right)$ which is given by its so-called δ-*level functions* $\underline{\pi}_\delta(\cdot \mid \boldsymbol{x}^\star)$ and $\overline{\pi}_\delta(\cdot \mid \boldsymbol{x}^\star)$ respectively.

From the fuzzy a-posteriori distributions generalized Bayesian confidence regions, fuzzy highest a-posteriori density regions, and fuzzy predictive distributions can be constructed. Moreover also decision analysis can be generalized to the situation of fuzzy utilities and non-precise data.

Applications

Whenever measurements of continuous quantities have to be modeled non-precise data appear. This is the case with initial conditions for differential equations, time dependent description of quantities, as well as in statistical analysis of environmental data.

About the Author

Professor Reinhard Viertl is Past Head of the Austrian Statistical Society, 1987 and 1991. He had founded the Austrian Bayesian Society in 1981. Dr. Viertl organized an International Symposium on Statistics with Non-precise Data at Innsbruck, 1993. He is an Elected Member of the New York Academy of Science (1997). He is author or co-author of

more than 100 papers and 10 books, including *Statistical Methods for Non-Precise Data* (CRC Press, Boca Raton, Florida, 1996)

Cross References

►Bayesian Statistics
►Fuzzy Logic in Statistical Data Analysis
►Fuzzy Sets: An Introduction

References and Further Reading

Bandemer H (1993) Modelling uncertain data. Akademie Verlag, Berlin
Bandemer H (2006) Mathematics of uncertainty. Springer, Berlin
Dubois D, Lubiano M, Prade H, Gil M, Grzegorzewski P, Hryniewicz O (eds) (2008) Soft methods for handling variability and imprecision. Springer, Berlin
Gil M, Corral N, Gil P (1988) The minimum inaccuracy estimates in χ^2-tests for goodness of fit with fuzzy observations. J Stat Plan Infer 19:95–115
Kacprzyk J, Fedrizzi M (eds) (1988) Combining fuzzy imprecision with probabilistic uncertainty in decision making. Lecture notes in economics and mathematical systems, vol 310, Springer, Berlin
Näther W (1997) Linear statistical inference for random fuzzy data. Statistics 29(3):221–240
Ross T, Booker J, Parkinson W (eds) (2002) Fuzzy logic and probability applications – bridging the gap. SIAM, Philadephia, PA
Viertl R (1996) Statistical methods for non-precise data. CRC Press, Boca Raton, FL
Viertl R (2006) Univariate statistical analysis with fuzzy data. Comput Stat Data Anal 51:133–147
Viertl R (2008) Foundations of fuzzy Bayesian inference. J Uncertain Syst 2:3
Viertl R, Hareter D (2006) Beschreibung und Analyse unscharfer Information – statistische Methoden für unscharfe Daten. Springer, Wien

Statistical Methods in Epidemiology

Giovanni Filardo[1], John Adams[2], Hon Keung Tony Ng[3]
[1]Director of Epidemiology
Baylor Health Care System, Dallas, TX, USA
[2]Epidemiologist
Baylor Health Care System, Dallas, TX, USA
[3]Associate Professor
Southern Methodist University, Dallas, TX, USA

Introduction

Epidemiology is the study of the distribution and determinants of health-related states or events in specified populations and the translation of study results to control health problems at the group level. The major objectives of epidemiologic studies are to describe the extent of disease in the community, to identify risk factors (factors that influence a persons risk of acquiring a disease), to determine etiology, to evaluate both existing and new preventive and therapeutic measures (including health care delivery), and to provide the foundation for developing public policy and regulatory decisions regarding public health practice. Epidemiologic studies provide research strategies for investigating public health questions in a systematic fashion relating a given health outcome to the factors that might cause and/or prevent this outcome in human populations. Statistics informs many decisions in epidemiologic study design and statistical tools are used extensively to study the association between risk factors and health outcomes.

When analyzing data for epidemiologic research, the intent is usually to extrapolate the findings from a sample of individuals to the population of all similar individuals to draw generalizable conclusions. Despite the enormous variety of epidemiologic problems and statistical solutions, there are two basic approaches to statistical analysis: regression and non-regression methods.

Types of Epidemiologic Studies and Related Risk Measures

Epidemiologist, in conceptualizing basic types of epidemiologic studies, often group them as experimental (e.g., randomized control trials) and observational (cohort, case-control, and cross-sectional) studies. This manuscript will focus on cohort and ►case-control studies. The study design determines how risk is measured (e.g., person-time at risk, absolute risk, odds) and which probability model should be employed.

Cohort Studies

In a cohort study, a group of persons are followed over a period of time to determine if an exposure of interest is associated with an outcome of interest. The key factor identifying a cohort study is that the exposure of interest precedes the outcome of interest. Depending on the exposure, different levels of exposure are identified for each subject and the subjects are subsequently followed over a period of time to determine if they experienced the outcome of interest (usually, health-related). Cohort studies are also called prospective studies, retrospective cohort studies, follow-up studies or longitudinal studies. Among all the observational studies (which includes cohort, case-control, and cross-section studies), cohort studies are the "gold standard." However, the major limitation of cohort studies is that they may require a large number of study

participants and usually many years of follow-up (which can be expensive). Loss to follow-up is another concern for cohort studies. Disease prevalence in the population under study may also determine the practicality of conducting a cohort study. Should the prevalence of an outcome be very low, the number of subjects needed to determine if there is an association between an exposure and outcome may be prohibitive within that population.

Cohort studies may result in counts, incidence (cumulative incidence or incidence proportion), or incidence rate of the outcome of interest. Suppose each subject in a large population-based cohort study is classified as exposed or unexposed to a certain risk factor and positive (case) or negative (noncase) for some disease state. Due to the loss-to-follow-up or late entry in the study, the data are usually presented in terms of number of diseases developed per person-years at risk.

The incidence rate in the exposed group and unexposed groups are then expressed as $\pi_1 = y_1/t_1$ per person-year and $\pi_2 = y_2/t_2$ per person-year, respectively (Table 1). In this situation, the numbers of disease developed in exposed and unexposed groups are usually modeled assuming a Poisson distribution when the event is relatively rare (see, Haight 1967; Johnson et al. 2005).

If there is no loss-to-follow-up or late entry in the study (closed cohort in which all participants contribute equal follow-up time), it may be convenient to present the data in terms of proportion experiencing the outcome (i.e., cumulative incidence or incidence proportion). A 2 × 2 table of sample person-count data in a cohort study is presented in Table 2.

Let p_1 and p_2 be the probabilities denoting risks for developing cases in the population for exposed and unexposed groups, respectively. The most commonly used sample estimates for p_1 and p_2 are obtained as

$$\pi_1 = \frac{x_{11}}{n_1} \text{ and } \pi_2 = \frac{x_{12}}{n_2}.$$

Statistical Methods in Epidemiology. Table 1 Data presented in terms of person-year at risk and the number of diseases developed

	Exposed	Unexposed
Disease develops	y_1	y_2
Person-year at risk	t_1	t_2
Incidence rate	y_1/t_1	y_2/t_2

Statistical Methods in Epidemiology. Table 2 2 × 2 table of sample person-count data

	Exposed	Unexposed	Total
Cases	x_{11}	x_{12}	m_1
Noncases	x_{21}	x_{22}	m_2
Total	n_1	n_2	N

Note that p_1 and p_2 are the incidence proportion in the exposed and unexposed groups, respectively. In this situation, the probability of disease in exposed and unexposed groups are usually modeled assuming a ▶binomial distribution. Statistical estimation and related inference for incidence can be found in Lui (2004) and Sahai and Khurshid (1995).

It is oftentimes the goal in epidemiologic studies to measure the association between an exposure and an outcome. Depending upon how subjects are followed, in regard to time, different measures of risk are used. Relative risk (RR) is defined as

$$RR = \frac{\text{incidence proportion (or rate) in exposed group}}{\text{incidence proportion (or rate) in unexposed group}} = \frac{\pi_1}{\pi_2}.$$

The relative risk is a ratio, therefore, it is dimensionless and without unit. It is a measure of the strength of an association between an exposure and a disease, and is the measure used in etiologic studies. In most real-world situations, subjects enter the study at different times and they are follow for variable lengths of time. In this situation, we should consider the number of cases per the total person-time contributed and the relative rate that approximates the RR defined as

$$\text{Relative rate} = \frac{\text{incidence rate in exposed group}}{\text{incidence rate in unexposed group}} = \frac{\pi_1}{\pi_2}.$$

Note that the units for π_1 and π_2 are per person-year. As it is a ratio, it is also unitless. Another measure of risk is the attributable risk (AR) which is defined as:

$$AR = \text{incidence rate in exposed group} - \text{incidence rate in unexposed group} = \pi_1 - \pi_2.$$

In the rare event of a closed cohort study framework, π_1 and π_2 can be replaced by p_1 and p_2. Attributable risk is the magnitude of disease incidence attributable to a specific exposure. It tells us the most we can hope to accomplish in reducing the risk of disease among the exposed if we totally eliminated the exposure. In other words, AR is a measure of how much of the disease incidence is attributable to the exposure. It is useful in assessing the exposures public health importance. Attributable risk

percent (ARP) in exposed group, the percent of disease incidence attributable to a specific exposure, is also used to measure the risk of disease

$$ARP = \frac{(RR-1)}{RR} \times 100.$$

ARP tells us what percent of disease in the exposed population is due to the exposure. The statistical inference on these measures of risk is discussed extensively in the literature, see, for example, Lui (2004) and Sahai and Khurshid (1995).

Case-Control Studies

Case-control studies (*see* also ►Case-Control Studies) compare a group of persons with a disease (cases) with a group of persons without the disease (controls) with respect to history of past exposures of interest. In contrast to a cohort study where an exposure of interest is determined preceding the development of future outcome, in a case-control, the disease status is known a priori while the exposure of interest is subsequently assessed among cases and controls.

Although the underlying concept of case-control studies is different from cohort study, the data for case-control study can be summarized as in a 2 × 2 table in Table 2. We can calculate the probability that cases were exposed as

$$\Pr(\text{exposed}|\text{case}) = \frac{x_{11}}{m_1}$$

and the probability that cases were not exposed as

$$\Pr(\text{unexposed}|\text{case}) = \frac{x_{12}}{m_1}.$$

We can also calculate the odds of a case being exposed as

$$\frac{x_{11}/m_1}{x_{12}/m_1} = \frac{x_{11}}{x_{12}}$$

and the odds of a case not being exposed as x_{21}/x_{22}. In case-control studies, although risk factors might contribute to the development of the disease, we cannot distinguish between risk factors for the development of the disease and risk factors for cure or survival. A major weakness in case control studies is that they are inherently unable to discern whether the exposure of interest precedes the outcome (with few exceptions). Additionally, there is some difficulty in the selection of controls. It is often the case that selected controls are not necessarily from the source population that gave rise to the cases. Therefore, measurement of association can be problematic. We cannot measure incidence rate (or proportion) in the exposed and non-exposed groups, and therefore cannot calculate rate ratios or relative risk directly. Because direct measures of risk are not applicable here, it is necessary to describe the relationship between an exposure and outcome using odds of exposure. The odds ratio (*OR*), ratio of the odds of exposure in cases and the odds of exposure in controls, is

$$OR = \frac{x_{11}/x_{12}}{x_{21}/x_{22}} = \frac{x_{11}x_{22}}{x_{12}x_{21}}.$$

The odds ratio is the cross-product ratio in the 2 × 2 table. The odds ratio is a good approximation of the relative risk when the disease being studied occurs infrequently in the population under study (case-control studies are conducted most frequently in this situation). An $OR = 1$ indicates that there is no association between exposure and outcome. When $OR > 1$ ($OR < 1$), it indicates a positive (negative) association between the exposure and disease and the larger (smaller) the *OR*, the stronger the association. An example of the calculation and interpretation of the odds ratio is given by Bland and Altman (2000).

Note that there are other variations in case-control studies and related statistical techniques which are applicable in particular situations. For instance, McNemar's test is used in matched case-control studies. For an extensive review on major development on statistical analysis of case-control studies, one can refer to Breslow (1996).

Regression vs. Non-Regression Methods

In analyzing data from epidemiologic studies, non-regression and regression methods are often used to study the relationship between an outcome and exposure. Non-regression methods of analysis control for differences in the distribution of covariates among subjects in exposure groups of interest by stratifying, while regression methods control for covariates by including possible confounders (*see* ►Confounding and Confounder Control) of the association of interest in a regression model. In some situations, regardless of whether regression techniques are used, stratification may still be necessary.

Statistical techniques used in epidemiologic studies are determined by the study design and data type. For cohort or case-control studies dealing with proportions, non-regression statistical methods based on binomial or negative binomial distribution could be applied, depending on the sampling method used (if any). Mantel-Haenszel procedures and ►Chi-square tests are the common approaches to access the association between the disease and risk factor with or without stratification. **Logistic regression** and **generalized linear models** are other possible regression methods that can be used for observational studies (see, for example, Harrell 2001). For stud-

ies with count data, statistical methods based on Poisson distribution could be applied (Cameron and Trivedi 1998).

Study designs that employ matched pairs or one-to-one matching are often approached by methods that assume a certain uniqueness of each member of the pair. The rationale for matching resembles that of blocking in statistical design, in that each stratum formed by the matching strategy is essentially the same with respect to the factors being controlled. When matching in cohort or case-control studies, McNemar's test, Mantel-Haenszel test and conditional logistic regression are normally used for analysis.

When the outcome variable is time-to-event, non-regression statistical estimation techniques for survival curves and log-rank tests can be applied, for example, the well-known **Kaplan-Meier estimator** can be used to estimate the survival curve. Lifetime parametric or **semiparametric regression models**, such as the Weibull regression model and Cox proportional hazard model (*see* ►Hazard Regression Models), can be used to model time-to-event data while controlling for possible confounders.

Cross References

►Binomial Distribution
►Biostatistics
►Case-Control Studies
►Confounding and Confounder Control
►Geometric and Negative Binomial Distributions
►Hazard Regression Models
►Incomplete Data in Clinical and Epidemiological Studies
►Medical Statistics
►Modeling Count Data
►Poisson Regression
►Time Series Models to Determine the Death Rate of a Given Disease

References and Further Reading

Bland JM, Altman DG (2000) Statistics notes: the odds. BMJ 320:1468
Breslow NE (1996) Statistics in epidemiology: the case-control study. J Am Stat Assoc 91:14–28
Cameron AC, Trivedi PK (1998) Regression analysis of count data. Cambridge University Press, New York
Haight FA (1967) Handbook of the Poisson distribution. Wiley, New York
Harrell FE (2001) Regression modeling strategies. Springer, New York
Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, New York
Lui KJ (2004) Statistical estimation of epidemiological risk. Wiley, New York
Rothman KJ, Greenland S (1998) Modern epidemiology. Lippincott Williams & Wilkins, Philadelphia, PA
Sahai H, Khurshid A (1995) Statistics in epidemiology: methods, techniques, and applications. CRC Press, Boca Raton, FL

Statistical Modeling of Financial Markets

Mhamed-Ali El-Aroui
Associate-Professor of Quantitative Methods
ISG de Tunis, Bardo, Tunisia

Overview

Optimal investment strategies and efficient risk management often need high-performance predictions of market evolutions. These predictions are usually provided by statistical models based on both statistical analyses of financial historical data and theoretical modeling of financial market working.

One of the pioneering works of financial market statistical modeling is the Ph.D. thesis of Bachelier (1900) who was the first to note that financial stock prices have unforecastable and apparently random variations. Bachelier introduced the Brownian process to model the price movements and to assess contingent claims in financial markets. He also introduced the random walk assumption (see ►Random Walk) according to which future stock price movements are generally unforecastable. More precisely, he assumed that the price evolves as a continuous homogeneous Markov process (see ►Markov Processes). Then, by considering the price process as a limit of random walks, he showed that this process satisfies the Chapman–Kolmogorov equation and that the Gaussian distribution with the linearly increasing variance solves this equation.

Between the 1920s and the 1960s, many economists and statisticians (Coles, Working, Kendall, Samuelson, etc.) analyzed several historical stock prices data and supported the random walk assumption.

In the 1960s, Samuelson and Fama gave both theoretical and empirical proofs of the random walk assumption. They introduced the important *efficient market hypothesis* stating that, in efficient markets, price movements should be unforecastable since they should fully incorporate the expectations and informations of all market participants.

Mandelbrot in 1963 criticized the Bachelier Gaussian assumption and stated that "*Despite the fundamental importance of the Brownian motion, (see ►Brownian Motion and Diffusions) it is now obvious that it does not account for the abundant data accumulated since 1900 by empirical economists, simply because the empirical distributions of price changes are usually too peaked to be relative to samples from Gaussian population.*" It is consensually assumed now that financial returns are generally *leptokurtic* and should be modeled by heavy tailed probability distributions. Many mathematical tools were suggested to model

this heavy tailed property: Levy process (see ►Lévy Processes), alpha-stable processes, Pareto-type distributions, Extreme value theory, long memory processes, GARCH time series, etc. Leptokurtosis and heteroskedasticity are stylized facts observed in log-returns of a large variety of financial data (security prices, stock indices, foreign exchange rates, etc.).

In the following, it will be assumed that a market economy contains N financial assets, S_{jt} and R_{jt} will denote, respectively, the daily price and log-return of the j-th asset on day t ($R_{jt} = \log(S_{jt}/S_{jt-1})$). $R_t^{(m)}$ will denote the log-return on day t of the market portfolio. It will also be assumed that there exists a single deterministic lending and borrowing risk-free rate denoted r.

Markowitz in 1952 developed the mean-variance portfolio optimization, where it is assumed that rational investors choose among risky assets purely on the basis of expected return and risk (measured as returns variance). Sharpe in 1964 presented the Capital Asset Pricing Model (CAPM) where the excess return over the risk-free rate r of each asset j is, up to noise, a linear function of the excess return of the market portfolio. In other words, for each asset j: $R_{jt} - r = \alpha_j + \beta_j(R_t^{(m)} - r) + \epsilon_{jt}$; where the noise sequence ϵ_{jt} is uncorrelated with the market portfolio return.

A third major step in the history of statistical modeling of financial markets concerns the problem of pricing derivative securities. Merton, Black, and Scholes introduced a reference paradigm for pricing and hedging derivatives on financial assets. Their paradigm, known as the *Black–Scholes formula*, is based on continuous time modeling of asset price movements. It gave an explicit formula for pricing European options and got tremendous impact on the financial engineering field. Since 1973, the Black–Scholes model was used to develop several extensions combining financial, mathematical, and algorithmic refinements.

Alternative statistical modeling approaches used time series statistical tools. Since the 1980s, time series tools are very frequently used in everyday manipulations and statistical analysis of financial data. Statistical Time series models, such as ARMA, ARIMA, ARCH, GARCH, state space models, and the important Granger cointegration concept, are often used to analyze the statistical internal structure of financial time series. These models, and especially the Engel Auto-Regressed Conditionally Heteroskedastic (ARCH) model, are well suited to the nature of financial markets, they capture time dependencies, volatility clustering, comovements, etc.

In the 1990s, the statistical modeling of financial markets data was linked to the rich literature of Extreme Value Theory (EVT). Many researchers found that EVT is well suited to model maxima and minima of financial returns. This yielded a more efficient assessment of financial market risks. New EVT-based methods were developed to estimate the *Value-at-Risk* (VaR), which is now one of the most used quantitative benchmarks for managing financial risk (recommended by the Basel international committee of banking supervision).

In the last 10 years, *copula functions* (see ►Copulas and ►Copulas: Distribution Functions and Simulation) have been used by many finance researchers to handle observed comovements between markets, risk factors, and other relevant dependent financial variables. The use of copula for modeling multivariate financial series open many challenging methodological questions to statisticians, especially concerning the estimation of copula parameters and the choice of the appropriate copula function.

It is worth noting that many works combining statistical science and market finance were rewarded by Nobel prizes in economics: Samuelson in 1970, Markowitz and Sharpe in 1990, Merton and Scholes in 1997, and Engle and Granger in 2003.

Due to space limitations, only two selected topics will be detailed in the following: Black–Scholes modeling paradigm and the contribution of Extreme Value Theory to the market risk estimation.

Black–Scholes Model

The Black–Scholes model is one of the most used option-pricing models in the trading rooms. For liquid securities, quotations could occur every 30 sec; continuous time models could therefore give good approximations to the variations of asset prices. Price evolution of a single asset is modeled here by a continuous time random process denoted $\{S_t\}_{t\in\mathbb{R}_+}$. Black and Scholes assume that the studied market has some ideal conditions: Market efficiency, no transaction costs in buying or selling the stock, the studied stock pays no dividend, and known and constant risk-free interest-rate r.

The basic modeling equation of Black, Scholes, and Merton, comes from the updating of a risky investment in a continuous time modeling: $(S_{t+\,dt} - S_t)/S_t = \mu\, dt + \sigma(\mathbb{B}_{t+\,dt} - \mathbb{B}_t)$, where μ is a constant parameter called *drift* giving the global trend of the stock price; σ is a nonnegative constant called *volatility* giving the magnitude of the price variations and $\mathbb{B}_{t+\,dt} - \mathbb{B}_t$ are independent increments (the independence results from the market efficiency assumption) from a Brownian motion, i.e., random centered Gaussian variables. So in Black–Scholes dynamics, the stock price $\{S_t\}_{t\in\mathbb{R}_+}$ satisfies the following stochastic differential equation :$dS_t/S_t = \mu\, dt + \sigma d\mathbb{B}_t$.

Using Itô lemma on Black–Scholes equation gives the explicit solution of the previous stochastic differential equation: $S_t = S_0 \exp\left[\left(\mu - \sigma^2/2\right)t + \sigma\mathbb{B}_t\right]$, which is a geometric Brownian motion. The model parameters μ and σ are easily estimated from data.

The Black–Scholes model is still a reference tool for pricing financial derivatives. Its simple formula makes it an everyday benchmark tool in all trading rooms. But its restrictive assumptions contradict many stylized facts recognized by all financial analysts (volatility clustering, leptokurtosis, and left asymmetry of the financial returns).

Many works have extended the Black–Scholes model: in the stochastic volatility extensions, for example, prices are modeled by the two following equations: $dS_t = S_t[\mu\, dt + \sigma_t\, dB_t]$ and $d\sigma_t = \sigma_t[\nu dt + \zeta dW_t]$, where B and W are two correlated Brownian motions having a constant correlation coefficient ρ. Both parametric and nonparametric estimators are available for the parameters μ, ν, ζ, ρ, and σ_0.

Challenging research topics now concern the problem of pricing sophisticated derivative products (American options, Asian or Bermudian options, swaptions, etc.). Longstaff and Schwartz, for example, gave an interesting pricing algorithm for American options, where they combined Monte Carlo simulations with ►least squares to estimate the conditional expected payoff of the optionholder. Monte Carlo simulation is now widely used in financial engineering; for example, Broadie and Glasserman 1996 used simulations to estimate security price derivatives within a modeling framework much more realistically than the simple Black–Scholes paradigm. Monte Carlo simulations are also used in stress testing (which identifies potential losses under simulated extreme market conditions) and in the estimation of nonlinear stochastic volatility models.

EVT and Financial Risks

The Extreme Value theory (EVT) gives interesting tools for modeling and estimating extreme financial risk (see Embrecht et al. 1997 for a general survey). One common use of EVT concerns the estimation of Value-at-Risk (an extreme quantile of the loss distribution). If at day t, $\mathrm{VaR}_t(\alpha)$ denotes the Value-at-Risk of a single asset at confidence level $1 - \alpha$ with a prediction horizon of one day, then VaR writes: $\Pr(R_{t+1} \leq -VaR_t(\alpha)|\mathcal{H}_t) = \alpha$, where R_{t+1} is the return at $t + 1$ and $\mathcal{H}_t$ denotes the $\sigma-$algebra modeling all the information available at time t. Many statistical methods were used to estimate the extreme quantile $VaR_t(\alpha)$. McNeil and Frey (2000), for example, combined ARCH and EVT to take into account volatility clustering and leptokurtosis. They used an AR(1) model for the average returns μ_t and a GARCH(1,1) with pseudo-maximum-likelihood estimation for the stochastic volatility dynamics σ_t. McNeil and Frey used the previous AR-GARCH for estimating the parameters of the model $R_t = \mu_t + \sigma_t Z_t$ where $\{Z_t\}_t$ is a strict white noise process. EVT peaks-over-threshold approach is then used on the AR-GARCH-residuals $z_1, \ldots, z_k$ in order to estimate their extreme quantiles. These estimates are plugged in the estimator of the $VaR_t(\alpha)$. The idea behind this method is the elimination of data dependence by the use of time series models and then the use of EVT tools to estimate extreme quantiles of the i.i.d. residuals.

When VaR of a multi-asset portfolio is considered, multivariate statistical tools should be used: variance–covariance, multivariate GARCH, simulation approach, Multivariate Extreme Theory, dynamic copula approach, etc. In the variance–covariance approach, for example, the portfolio returns are modeled as a linear combination of slected market factors. The copula approach gives generally more efficient portfolio VaR estimations since it improves the modeling of the dependence structure between the studied assets and the risk factors.

Conclusions

Statistical science has provided essential tools for market finance. These important contributions concern the problems of portfolio selection and performance analysis, the pricing and hedging of derivative securities, the assessment of financial risks (market risk, operational risk, credit risk), the modeling of crises contagion, etc. Many challenging research topics concern both statistics and finance: the huge amount of data (called high-frequency data) need new statistical modeling approaches. The high complexity of the new financial products and the management of portfolios with high number of assets need more tractable multivariate statistical models. New research challenges are also given by the multivariate extreme value theory where copula functions gave promising results when used to model extreme comovements of asset prices or stock indices. Copula modeling has become an increasingly popular tool in finance, especially for modeling dependency between different assets. However many statistical questions remain open: copula parameter estimations, statistical comparison of competitive copula, etc. Another use of copula functions in market finance concerns the modeling of crises contagion (see, e.g., Rodriguez 2007). Many empirical works proved that dependence structure between international markets during crises is generally nonlinear and therefore better modeled by copula functions.

Cross References

►Banking, Statistics in
►Brownian Motion and Diffusions
►Copulas
►Copulas in Finance
►Financial Return Distributions
►Heavy-Tailed Distributions
►Heteroscedastic Time Series
►Lévy Processes
►Monte Carlo Methods in Statistics
►Nonlinear Time Series Analysis
►Optimal Statistical Inference in Financial Engineering
►Portfolio Theory
►Quantitative Risk Management
►Random Walk
►Statistical Modelling in Market Research

References and Further Reading

Bachelier L (1900) Théorie de la spéculation. Ann Sci École Norm S 81(3):21–86. Available at www.numdam.org
Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–654
Broadie M, Glasserman P (1996) Estimating security price derivatives using simulation. Manag Sci 42(2):269–285
Embrecht P, Kluppelberg C, Mikosch T (1997) Modeling extremal events for insurance and finance. Springer, Berlin
Engel RF, Granger CWJ (1987) Co-integration and error correction: representation, estimation and testing. Econometrica 55(2):251–276
Longstaff FA, Schwartz ES (2001) Valuing American options by simulation: a simple least-squares approach. Rev Financ Stud 14(1):113–147
Markowitz H (1952) Portfolio selection. J Financ 7:77–91
McNeil A, Frey R (2000) Estimation of tail related risk measures for heteroscedastic financial time series: an extreme value approach. J Empirical Financ 7:271–300
Rodriguez JC (2007) Measuring financial contagion: a copula approach. J Empirical Financ 14:401–423
Sharpe W (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Financ 19:425–442

Statistical Modelling in Market Research

Anatoly Zhigljavsky
Professor, Chair in Statistics
Cardiff University, Cardiff, UK

Mathematical modelling is a key element of quantitative marketing and helps companies around the globe in making important marketing decisions about launching new products and managing existing ones. Most mathematical models used in marketing research are either purely statistical or include elements of statistical models.

An extensive discussion (by the top market research academics) of the state-of-art in the field of marketing modelling and its prospects for the future is contained in Steemkamp (2000), a special issue of the International Journal of Research in Marketing. One can consult Steemkamp (2000) for many references related to the subject; see also recent books (Wierenga 2008; Wittink et al. 2000; Mort 2001; Zikmund and Babin 2009).

We look at the field of market modelling from a viewpoint of a professional statistician with twenty years of experience on designing and using statistical models in market research. We start with distinguishing the following types of statistical models used in market research:

1. Direct simulation models
2. Standard statistical models
3. Models of consumer purchase behaviour
4. Dynamic models for modelling competition, pricing and advertising strategies
5. Statistical components of inventory and other management science models

Let us briefly consider these types of models separately.

1. Direct *simulation models*. These are specialized models based on attempts to directly imitate the market (e.g., via the behaviour of individual customers) using a synergy of stochastic and deterministic rules. These models were popular 20–30 years ago but are less popular now. The reasons are the lack of predictive power, huge number of parameters in the models and impossibility of their validation.

2. *Standard statistical models*. All standard statistical models and methods can be used in market research, see Mort (2001); Zikmund and Babin (2009); Rossi et al. (2005); Hanssens et al. (2003). Most commonly, the following statistical models are used:

- Various types of regression
- ARIMA and other time series models
- Bayesian models
- Models and methods of multivariate statistics; especially, structural equation and multinomial response models, conjoint, factor, and principal component analyses

3. *Models of consumer purchase behaviour*. Several types of statistical models are used for modelling consumer purchase behaviour including brand choice. The following three basic models (and some of their extensions) have

proved to be the most useful: Mixed ►Poisson processes, the Dirichlet model, and Markovian models.

The mixed Poisson process model assumes that a customer makes his/her purchase according to a Poisson process with some intensity λ where λ is random across the population. In the most popular model, called Gamma-Poisson, λ has Gamma distribution (with two unknown parameters); this yields that the number of purchases for a given period is the Negative Binomial Distribution. Typical questions, which the Poisson process model answers, is the forecasting of the behaviour of the market research measures (like penetration, purchase frequency and repeat buying measures) in the form of the so-called growth curves. Extensions of the mixed Poisson models cover the issues like the zero-buyer problem (some zero-buyers do have a positive propensity to buy but some other don't), seasonality of the market and the panel flow-through.

The Dirichlet model is a brand-choice model. It assumes that customers make their brand choice independently with certain propensities; these propensities are different for all customers and are independent realizations from the Dirichlet distribution which parameters are determined by the market shares of the brands. In Markovian brand-choice models, the propensity to buy a given brand for a random customer may vary depending on either the previous purchase or other market variables. These models are more complicated than the mixed Poisson process and Dirichlet models but in some circumstances are easily applicable and sometimes are able to accurately describe some features of the market.

Of course, the models above are unrealistic on the individual level (e.g., few people have the Poisson process pattern as their purchase sequence). However, these models (and especially the mixed Poisson model) often fit data extremely accurately on the aggregated level (when the time period considered and the number of customers are sufficiently large). These models can be classified as descriptive (rather than "prescriptive") and help in explaining different aspects of market research dynamics and some phenomena related to the brand-choice.

4. *Dynamic models for modelling competition, pricing and advertising strategies.* There is extensive literature on this subject, see, e.g., Erickson (1991). The majority of the models are so-called differential games or simpler models still written in terms of differential equations. The models are deterministic and the statistical aspect only arrives through the assumption that the data contain random errors. Statistical modelling part is therefore negligible in these models. Alternatively, in some Markovian brand-choice models mentioned above, there is an option of including the market variables (e.g., promotion) into the updating rule for the buying propensities. These models are proper stochastic models but they are often too complicated (have too many parameters) and therefore difficult to validate.

5. *Statistical components of inventory and other management science models.* Inventory and other management science models applied in market research are typically standard models of Operations Research, see Ingene and Parry (2004) for a recent review of these models. Despite these models often have a large stochastic component, they do not represent anything special from the statistics view-point.

Statistical models are used for the following purposes: (a) forecasting the market behaviour of a new brand to prepare its launch and (b) managing existing brands. In case (a), the models are usually based solely on standard statistical models, type 2 above. Sometimes, other types of models (especially, large simulation models, type 1) are used too. A lot of specific market research data are often collected to feed these models. These data includes market surveys, various types of questionnaires and focus group research in direct contact with customers. All available market data, for example economic trends and specific industry sector reports, is used too. In case (b), the models are used for making decisions about pricing, promotion and advertising strategies, production and inventory management etc. All available statistical models and methods are used to help managers to make their decisions.

While reading academic papers and books on marketing research, one can get an impression that mathematical and statistical modelling in marketing is a mature subject with many models developed and used constantly for helping market research managers in working out their decisions. Indeed, there are many models available (some of them are quite sophisticated). However, only a small number of them are really used in practice: the majority of practical models can be reduced either to a simple regression or sometimes to another standard model among those mentioned above. One of the reasons for this gloomy observation is the fact that managers rarely want a description of the market. Instead, they want 'a prescription'; that is, a number (with a hope that no confidence interval is attached to this number) which would lead them to a right decision. Another reason is the fact that only a very few models used in market research satisfy the following natural requirements for a good statistical model: (a) simplicity, (b) robustness to the deviations from the model assumptions, (c) clear range of applicability, and (d) empirical character, which means that the models have to be built with the data (and data analysis) in view and with the purpose of explaining/fitting/forecasting relevant data.

Despite huge amounts of market data is available to analysts, these data are typically messy, not reliable, badly structured and become outdated very quickly. Development of reliable statistical models dealing with such data is hard. The progress in understanding all these issues and tackling them by means of the development of appropriate models and making them correctly applicable is visible but it is justifiably slow.

Cross References

►Box–Jenkins Time Series Models
►Gamma Distribution
►Model Selection
►Multivariate Statistical Distributions
►Poisson Processes
►Statistical Modeling of Financial Markets

References and Further Reading

Erickson GM (1991) Dynamic models of advertising competition: open- and closed-loop extensions. Springer
Hanssens DM, Parsons LJ, Schultz RL (2003) Market response models: econometric and time series analysis. Springer, Berlin
Ingene CA, Parry ME (2004) Mathematical models of distribution channels. Springer, New York
Mort D (2001) Understanding statistics and market research data. Europa publications, London
Rossi PE, Allenby GM, McCullog R (2005) Bayesian statistics and marketing. Wiley/Blackwell, New York
Steenkamp, J-BEM (ed) (2000) Marketing modeling on the threshold of the 21st century. Int J Res Mark 17(2–3):99–253
Wierenga B (ed) (2008) Handbook of marketing decision models. Springer, New York
Wittink DR, Leeflang PSH, Wedel M, Naert PA (2000) Building models for marketing decisions. Kluwer Academic, Boston, MA
Zikmund WG, Babin BJ (2009) Exploring marketing research. South-western Educational Publishing, Florence, KY

Statistical Natural Language Processing

Florentina T. Hristea
Associate Professor, Faculty of Mathematics and Computer Science
University of Bucharest, Bucharest, Romania

Natural language processing (*NLP*) is a field of artificial intelligence concerned with the interactions between computers and human (natural) languages. It refers to a technology that creates and implements ways of executing various tasks concerning natural language (such as designing natural language based interfaces with databases, machine translation, etc.). NLP applications belong to three main categories:

1. Text-based applications (such as knowledge acquisition, information retrieval, information extraction, text summarization, machine translation, etc.)
2. Dialog-based applications (such as learning systems, question answering systems, etc.)
3. Speech processing (although NLP may refer to both text and speech, work on speech processing has gradually evolved into a separate field)

Natural language engineering deals with the implementation of large-scale natural language–based systems. It refers to the related field of *Human Language Technology* (*HLT*).

NLP represents a difficult and largely unsolved task. This is mainly due to the interdisciplinary nature of the problem that requires interaction between many sciences and fields: linguistics, psycholinguistics, computational linguistics, philosophy, statistics, computer science in general, and artificial intelligence in particular.

Statistical NLP has been the most widely used term to refer to nonsymbolic and nonlogical work on NLP over the past decade. Statistical NLP comprises all *quantitative approaches* to automated language processing, including probabilistic modeling, information theory, and linear algebra (Manning and Schütze 1999).

As computational problems, many problems posed by NLP (such as WSD – word sense disambiguation) were often described as AI-complete, that is, problems whose solutions presuppose a solution to complete natural language understanding or common-sense reasoning. This view originated from the fact that possible statistical approaches to such problems were almost completely ignored in the past. As it is well known, starting with the early 1990s, the artificial intelligence community witnessed a great revival of empirical methods, especially statistical ones. This is due to the success of statistical approaches, as well as of machine learning, in solving problems such as speech recognition or part-of-speech tagging. It was mainly research into speech recognition that inspired the revival of statistical methods within NLP, and many of the techniques used nowadays were developed first for speech and then spread over into NLP (Manning and Schütze 1999). Nowadays statistical methods and machine learning algorithms are used for solving a great number of problems posed by artificial intelligence in general and by NLP in particular. Furthermore, the availability of large

text corpora has changed the scientific approach to language in linguistics and cognitive science, with language and cognition being viewed as probabilistic phenomena.

From the point of view of NLP, the two main components of statistics are:

1. *Descriptive statistics*: methods for summarizing (large) datasets
2. *Inferential statistics*: methods for drawing inferences from (large) datasets

The use of statistics in NLP falls mainly into three categories (Nivre 2002):

1. *Processing*: We may use probabilistic models or algorithms to process natural language input or output.
2. *Learning*: We may use inferential statistics to learn from examples (corpus data). In particular, we may estimate the parameters of probabilistic models that can be used in processing.
3. *Evaluation*: We may use statistics to assess the performance of language processing systems.

As pointed out in Manning and Schütze (1999), "complex probabilistic models can be as explanatory as complex non-probabilistic models – but with the added advantage that they can explain phenomena that involve the type of *uncertainty* and *incompleteness* that is so pervasive in cognition in general and in language in particular."

A practical NLP system must be good at making *disambiguation decisions* of word sense, word category, syntactic structure, and semantic scope. One could say that disambiguation abilities, together with robustness, represent the two main hallmarks of statistical natural language processing models. Again as underlined in Manning and Schütze (1999), "a statistical NLP approach seeks to solve these problems by automatically learning lexical and structural preferences from corpora... The use of statistical models offers a good solution to the ambiguity problem: statistical models are robust, generalize well, and behave gracefully in the presence of errors and new data. Thus statistical NLP methods have led the way in providing successful disambiguation in large scale systems using naturally occurring text. Moreover, the parameters of Statistical NLP models can often be estimated automatically from text corpora, and this possibility of automatic learning not only reduces the human effort in producing NLP systems, but raises interesting scientific issues regarding human language acquisition."

Cross References

►Data Mining
►Distance Measures
►Estimation
►Expert Systems
►Information Theory and Statistics
►Statistical Inference

References and Further Reading

Manning C, Schütze H (1999) Foundations of statistical natural language processing. The MIT Press, Cambridge, MA

Nivre J (2002) On statistical methods in natural language processing. In: Berbenko J, Wangler B (eds) Promote IT. Second Conference for the Promotion of Research in IT at New Universities and University Colleges in Sweden. University of Skovde, Billingehus, Skövde, pp 684–694

Statistical Pattern Recognition Principles

Nicholas A. Nechval[1], Konstantin N. Nechval[2], Maris Purgailis[3]
[1]Professor, Head of the Mathematical Statistics Department
University of Latvia, Riga, Latvia
[2]Assistant Professor
Transport and Telecommunication Institute, Riga, Latvia
[3]Professor, Dean of the Faculty of Economics and Management
University of Latvia , Riga, Latvia

Problem Description

Mathematically, pattern recognition is a classification problem. Consider the recognition of characters. We wish to design a system such that a handwritten symbol will be recognized as an "A," a "B," etc. In other words, the machine we design must classify the observed handwritten character into one of 26 classes. The handwritten characters are often ambiguous, and there will be misclassified characters. The major goal in designing a pattern recognition machine is to have a low probability of misclassification.

There are many problems that can be formulated as pattern classification problems. For example, the weather may be divided into three classes, fair, rain, and possible rain, and the problem is to classify tomorrow's weather into one of these three classes. In the recognition of electrocardiograms, the classes are disease categories plus the class of normal subjects. In binary data transmission, a "one" and a "zero" are represented by signals of amplitudes A_1 and A_0, respectively. The signals are distorted or corrupted by noise when transmitted over communication channels, and the

receiver must classify the received signal into "ones" and "zeros." Hence, many of the ideas and principles in pattern recognition may be applied to the design of communication systems and vice versa (Nechval 1997; Nechval and Nechval 1999).

Pattern recognition theory deals with the mathematical aspects common to all pattern recognition problems. Application of the theory to a specific problem, however, requires a thorough understanding of the problem, including its peculiarities and special difficulties (Bishop 2006).

The input to a pattern recognition machine is a set of p measurements, and the output is the classification. It is convenient to represent the input by a p-dimensional vector $\mathbf{x}$, called a *pattern vector*, with its components being the p measurements. The classification at the output depends on the input vector $\mathbf{x}$, hence we write

$$C = d(x). \tag{1}$$

In other words, the machine must make a decision as to the class to which $\mathbf{x}$ belongs, and $d(\mathbf{x})$ is called a *decision function*.

A pattern recognition machine may be divided into two parts, a feature extractor and a classifier. The classifier performs the classification, while the feature extractor reduces the dimensionality of input vectors to the classifier. Thus, feature extraction is a linear or nonlinear transformation

$$y = Y(x), \tag{2}$$

which transforms a pattern vector $\mathbf{x}$ (in the pattern space Ω_x) into a *feature vector* $\mathbf{y}$ (in a *feature space* Ω_y). The classifier then classifies $\mathbf{x}$ based on $\mathbf{y}$. Since Ω_y is of lower dimensionality than Ω_x, the transformation is singular and some information is lost. The feature extractor should reduce the dimensionality but at the same time maintain a high level of machine performance. A special case of feature extraction is feature selection, which selects as features a subset of the given measurements.

The division of a pattern recognition machine into feature extractor and classifier is done out of convenience rather than necessity. It is conceivable that the two could be designed in an unified manner using a single performance criterion. When the structure of the machine is very complex and the dimensionality p of the pattern space is high, it is more convenient to design the feature extractor and the classifier separately.

The problem of pattern classification may be discussed in the framework of hypothesis testing. Let us consider a simple example. Suppose that we wish to predict a student's success or failure in graduate study based on his GRE (Graduate Record Examination) score. We have two hypotheses – the null hypothesis H_0, that he or she will be successful, and the alternative hypothesis H_1, that he or she will fail. Let x be the GRE score, $f_0(x)$ be the conditional probability density of x, given that the student will be successful, and $f_1(x)$ be the conditional density of x, given that he or she will fail. The density functions $f_0(x)$ and $f_1(x)$ are assumed known from our past experience on this problem. This is a hypothesis testing problem and an obvious decision rule is to retain H_0 and reject H_1 if x is greater than a certain threshold value h, and accept H_1 and reject H_0 if $x \leq h$. A typical example of multiple hypothesis testing is the recognition of English alphabets where we have 26 hypotheses.

Illustrative Examples

Applicant Recognition for Project Realization with Good Contract Risk

One of the most important activities that an employer has to perform is recognition of applicant for realization of project with good contract risk. The employer is defined as a firm or an institution or an individual who is investing in a development. The above problem is a typical example of a pattern classification problem. An applicant for contract can be represented by a random $p \times 1$ vector $\mathbf{X} = (X_1, \ldots, X_p)'$ of features or characteristics. We call this $p \times 1$ vector the applicant's pattern vector. Using historical data and the applicant's pattern vector, a decision-maker must decide whether to accept or reject the contract request. The historical data are summarized in a collection of pattern vectors. There are pattern vectors of former applicants who received contract and proved to be good risks, and there are patterns of former applicants who were accepted and proved to be poor risks. The historical data should include the pattern vectors and eventual contract status of applicants who were rejected. The eventual contract status of rejected applicants is difficult to determine objectively, but without this information, the historical data will contain the basis of former decision rules. The historical data consist of the pattern vectors and eventual contract status of n applicants; $n = n_1 + n_2$: n_1 of the n applicants proved to be good contract risks, and n_2 proved to be poor contract risks. Given this situation and a new applicant's pattern vector, the decision-maker deals with the problem of how to form his or her decision rule in order to accept or reject new applicants. In this entry, we shall restrict attention to the case when $p(\mathbf{X}; H_i)$, $i = 1, 2$, are multivariate normal with unknown parameters. All statistical information is contained in the historical data. In this case, the procedure based on a generalized likelihood ratio test is proposed. This procedure is relatively simple to carry out and can be

recommended in those situations when we deal with small samples of the historical data (Nechval and Nechval 1998).

Generalized Likelihood Ratio Test for Applicant Recognition. Let $\mathbf{X}$ be a random $p\times 1$ vector that is distributed in the population Π_i $(i = 0,1,2)$ according to the p-variate nonsingular normal distribution $N(\mathbf{a}_i, \mathbf{Q}_i)$ $(i = 0,1,2)$. Let $\mathbf{x}_0$ be an observation on $\mathbf{X}$ in Π_0. The n_i independent observations from Π_i will be denoted by $\{\mathbf{x}_{ij}, j = 1,2,\ldots,n_i\}$ distributed with the density $p(\mathbf{x}_{ij};\mathbf{a}_i,\mathbf{Q}_i)$ for $i = 1,2$ and the density of the unidentified observation $\mathbf{x}_0$ will be taken as $p(\mathbf{x}_0;\mathbf{a}_0,\mathbf{Q}_0)$. The $\mathbf{a}_i$s and $\mathbf{Q}_i$s are unknown and it is assumed that either $(\mathbf{a}_0,\mathbf{Q}_0) = (\mathbf{a}_1,\mathbf{Q}_1)$, or $(\mathbf{a}_0,\mathbf{Q}_0) = (\mathbf{a}_2,\mathbf{Q}_2)$, and $\mathbf{a}_1 \neq \mathbf{a}_2$, $\mathbf{Q}_1 \neq \mathbf{Q}_2$. Assume for the moment that there are prior odds of $\xi/(1-\xi)$ in favor of type 1 for $\mathbf{x}_0$. Then the likelihood ratio statistic for testing the null hypothesis $H_1 : (\mathbf{a}_0 = \mathbf{a}_1,\ \mathbf{Q}_0 = \mathbf{Q}_1)$ versus the alternative hypothesis $H_2 : (\mathbf{a}_0 = \mathbf{a}_2,\ \mathbf{Q}_0 = \mathbf{Q}_2)$ is given by

$$LR = \frac{\xi \max\limits_{H_1} p(\mathbf{x}_0;\mathbf{a}_1,\mathbf{Q}_1)\prod\limits_{i=1}^{2}\prod\limits_{j=1}^{n_i} p(\mathbf{x}_{ij};\mathbf{a}_i,\mathbf{Q}_i)}{(1-\xi)\max\limits_{H_2} p(\mathbf{x}_0;\mathbf{a}_2,\mathbf{Q}_2)\prod\limits_{i=1}^{2}\prod\limits_{j=1}^{n_i} p(\mathbf{x}_{ij};\mathbf{a}_i,\mathbf{Q}_i)}, \quad (3)$$

where

$$p(\mathbf{x}_0;\mathbf{a}_0,\mathbf{Q}_0) = (2\pi)^{-P/2}|\mathbf{Q}_0|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_0-\mathbf{a}_0)'\mathbf{Q}_0^{-1}(\mathbf{x}_0-\mathbf{a}_0)\right\}, \quad (4)$$

$$p(\mathbf{x}_{ij};\mathbf{a}_i,\mathbf{Q}_i) = (2\pi)^{-P/2}|\mathbf{Q}_i|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}_{ij}-\mathbf{a}_i)'\mathbf{Q}_i^{-1}(\mathbf{x}_{ij}-\mathbf{a}_i)\right\}. \quad (5)$$

The maximum likelihood estimators of the unknown parameters under H_1 are

$$\widehat{\mathbf{a}}_1 = \frac{n_1\bar{\mathbf{x}}_1 + \mathbf{x}_0}{n_1+1}, \quad (6)$$

$$\widehat{\mathbf{a}}_2 = \bar{\mathbf{x}}_2, \quad (7)$$

$$\widehat{\mathbf{Q}}_1 = \frac{1}{n_1+1}\left[(n_1-1)\mathbf{S}_1 + \frac{n_1}{n_1+1}(\mathbf{x}_0-\bar{\mathbf{x}}_1)(\mathbf{x}_0-\bar{\mathbf{x}}_1)'\right], \quad (8)$$

$$\widehat{\mathbf{Q}}_2 = \frac{n_2-1}{n_2}\mathbf{S}_2, \quad (9)$$

where

$$\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i}\mathbf{x}_{ij}/n_i, \quad (10)$$

$$S_i = \sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)(\mathbf{x}_{ij}-\bar{\mathbf{x}}_i)'/(n_i-1), \quad i = 1,2, \quad (11)$$

with obvious changes for the corresponding estimators under H_2. Substitution of the estimators in (3) gives, after some simplification,

$$LR = \left[\frac{(n_1+1)(n_2-1)}{(n_2+1)(n_1-1)}\right]^{p/2} = \left[\frac{(n_2/(n_2+1))^{pn_2/2}}{(n_1/(n_1+1))^{pn_1/2}}\left(\frac{|S_2|}{|S_1|}\right)^{1/2} \times \frac{(1+n_2v_2(x_0)/(n_2^2-1))^{(n_2+1)/2}}{(1+n_1v_1(x_0)/(n_1^2-1))^{(n_1+1)/2}}\right]\left(\frac{\xi}{1-\xi}\right), \quad (12)$$

where

$$v_i(\mathbf{x}_0) = (\mathbf{x}_0-\bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}_0-\bar{\mathbf{x}}_i), \quad i = 1,2. \quad (13)$$

For $\mathbf{Q}_1 = \mathbf{Q}_2$, the likelihood ratio statistic simplifies to

$$\mathrm{LR} = \left[\frac{1+\dfrac{n_2v_2(x_0)}{(n_2+1)(n_1+n_2-2)}}{1+\dfrac{n_1v_1(x_0)}{(n_1+1)(n_1+n_2-2)}}\right]^{(n_1+n_2+1)/2}\left(\frac{\xi}{1-\xi}\right), \quad (14)$$

and hypothesis H_1 or H_2 is favoured according to whether LR is greater or less than 1, that is,

$$LR\begin{cases} > 1, & \text{then} \quad H_1 \\ \leq 1, & \text{then} \quad H_2 \end{cases}. \quad (15)$$

Signal Detection in Clutter

The problem of detecting the unknown deterministic signal $\mathbf{s}$ in the presence of a clutter process, which is incompletely specified, can be viewed as a binary hypothesis-testing problem (Nechval 1992; Nechval et al. 2004). The decision is based on a sample of observation vectors $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})'$, $i = 1(1)n$, each of which is composed of clutter $\mathbf{w}_i = (w_{i1},\ldots,w_{ip})'$ under the null hypothesis H_0 and a signal $\mathbf{s} = (s_1,\ldots,s_p)'$ added to clutter $\mathbf{w}_i$ under the alternative H_1, where $n > p$. The two hypotheses that the detector must distinguish are given by

$$H_0 : \mathbf{X} = \mathbf{W} \quad \text{(clutter alone)}, \quad (16)$$

$$H_1 : \mathbf{X} = \mathbf{W} + \mathbf{c}\mathbf{s}' \quad \text{(signal present)}, \quad (17)$$

where

$$\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)', \quad (18)$$

$$\mathbf{W} = (\mathbf{w}_1,\ldots,\mathbf{w}_n)', \quad (19)$$

are $n\times p$ random matrices, and

$$\mathbf{c} = (1,\ldots,1)' \quad (20)$$

is a column vector of n units. It is assumed that $\mathbf{w}_i$, $i = 1(1)n$, are independent and normally distributed with common mean 0 and covariance matrix (positive definite) $\mathbf{Q}$, i.e.,

$$\mathbf{w}_i \sim N_p(0,\mathbf{Q}), \quad \forall i = 1(1)n. \quad (21)$$

Thus, for fixed n, the problem is to construct a test, which consists of testing the null hypothesis

$$H_0 : \mathbf{x}_i \sim N_p(0, \mathbf{Q}), \quad \forall i = 1(1)n, \tag{22}$$

versus the alternative

$$H_1 : \mathbf{x}_i \sim N_p(\mathbf{s}, \mathbf{Q}), \quad \forall i = 1(1)n, \tag{23}$$

where the parameters $\mathbf{Q}$ and $\mathbf{s}$ are unknown.

One of the possible statistics for testing H_0 versus H_1 is given by the generalized maximum likelihood ratio (GMLR)

$$GMLR = \max_{\theta \in \Theta_1} L_{H_1}(\mathbf{X}; \boldsymbol{\theta}) \Big/ \max_{\theta \in \Theta_0} L_{H_0}(\mathbf{X}; \boldsymbol{\theta}), \tag{24}$$

where $\boldsymbol{\theta} = (\mathbf{s}, \mathbf{Q}), \Theta_0 = \{(\mathbf{s}, \mathbf{Q}) : \mathbf{s} = 0, \mathbf{Q} \in Q_p\}$, $\Theta_1 = \Theta - \Theta_0$, $\Theta = \{(\mathbf{s}, \mathbf{Q}) : \mathbf{s} \in \mathbb{R}^p, \mathbf{Q} \in Q_p\}$, Q_p denotes the set of $p \times p$ positive definite matrices. Under H_0, the joint likelihood for $\mathbf{X}$ based on (22) is

$$L_{H_0}(\mathbf{X}; \boldsymbol{\theta}) = (2\pi)^{-np/2} |\mathbf{Q}|^{-n/2} \exp\left(-\sum_{i=1}^{n} \mathbf{x}_i' \mathbf{Q}^{-1} \mathbf{x}_i / 2 \right). \tag{25}$$

Under H_1, the joint likelihood for $\mathbf{X}$ based on (23) is

$$L_{H_1}(\mathbf{X}; \boldsymbol{\theta}) = (2\pi)^{-np/2} |\mathbf{Q}|^{-n/2} \exp\left(-\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{s})' Q^{-1} (\mathbf{x}_i - \mathbf{s}) / 2 \right). \tag{26}$$

It can be shown that

$$GMLR = \left| \widehat{\mathbf{Q}}_0 \right|^{n/2} \left| \widehat{\mathbf{Q}}_1 \right|^{-n/2}, \tag{27}$$

and

$$\widehat{\mathbf{Q}}_0 = \mathbf{X}'\mathbf{X}/n, \tag{28}$$

$$\widehat{\mathbf{Q}}_1 = (\mathbf{X}' - \hat{\mathbf{s}}\mathbf{c}')(\mathbf{X}' - \hat{\mathbf{s}}\mathbf{c}')'/n, \tag{29}$$

and

$$\hat{\mathbf{s}} = \mathbf{X}'\mathbf{c}/n \tag{30}$$

are the well-known maximum likelihood estimators of the unknown parameters $\mathbf{Q}$ and $\mathbf{s}$ under the hypotheses H_0 and H_1, respectively. It can be shown, after some algebra, that (27) is equivalent finally to the statistic

$$y = \mathbf{T}_1' \mathbf{T}_2^{-1} \mathbf{T}_1 / n, \tag{31}$$

where $\mathbf{T}_1 = \mathbf{X}'\mathbf{c}$, $\mathbf{T}_2 = \mathbf{X}'\mathbf{X}$. It is known that $(\mathbf{T}_1, \mathbf{T}_2)$ is a complete sufficient statistic for the parameter $\boldsymbol{\theta} = (\mathbf{s}, \mathbf{Q})$. Thus, the problem has been reduced to consideration of the sufficient statistic $(\mathbf{T}_1, \mathbf{T}_2)$. It can be shown that under H_0, the result (31) is a $\mathbf{Q}$-free statistic y, which has the property that its distribution does not depend on the actual covariance matrix $\mathbf{Q}$. It is clear that the statistic y is equivalent to the statistic

$$v = [(n-p)/p]\, y/(1-y) = [n(n-p)/p] \left(\hat{\mathbf{s}}' \left[\widehat{\mathbf{G}}_1 \right]^{-1} \hat{\mathbf{s}} \right), \tag{32}$$

where

$$\widehat{\mathbf{G}}_1 = n\widehat{\mathbf{Q}}_1 = (\mathbf{X}' - \hat{\mathbf{s}}\mathbf{c}')(\mathbf{X}' - \hat{\mathbf{s}}\mathbf{c}')' = \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mathbf{s}})(\mathbf{x}_i - \hat{\mathbf{s}})'. \tag{33}$$

Under H_1, the statistic v is subject to a noncentral F-distribution with p and $n-p$ degrees of freedom, the probability density function of which is (Nechval 1992; Nechval et al. 2004)

$$f_{H_1}(v; n, q) = \left[B\left(\frac{p}{2}, \frac{n-p}{2} \right) \right]^{-1} \frac{\left(\frac{p}{n-p} \right)^{p/2} v^{p/2-1}}{\left(1 + \frac{p}{n-p} v \right)^{n/2}} \times e^{-q/2} {}_1F_1\left(\frac{n}{2}; \frac{p}{2}; \frac{q}{2} \left(\frac{p}{n-p} v \left(1 + \frac{p}{n-p} v \right)^{-1} \right) \right), \quad 0 < v < \infty, \tag{34}$$

where ${}_1F_1(a; b; x)$ is the confluent hypergeometric function (Abramowitz and Stegun 1964),

$$q = n\left(\mathbf{s}' \mathbf{Q}^{-1} \mathbf{s} \right) \tag{35}$$

is a noncentrality parameter representing the generalized signal-to-noise ratio (GSNR). Under H_0, when $q = 0$, (34) reduces to a standard F-distribution with p and $n-p$ degrees of freedom,

$$f_{H_0}(v; n) = \left[B\left(\frac{p}{2}, \frac{n-p}{2} \right) \right]^{-1} \frac{\left(\frac{p}{n-p} \right)^{p/2} v^{p/2-1}}{\left(1 + \frac{p}{n-p} v \right)^{n/2}}, \quad 0 < v < \infty. \tag{36}$$

The test of H_0 versus H_1, based on the GMLR statistic v, is given by

$$v \begin{cases} > h, & \text{then} \quad H_1 \quad \text{(signal present)}, \\ \le h, & \text{then} \quad H_0 \quad \text{(clutter alone)}, \end{cases} \tag{37}$$

and can be written in the form of a decision rule $u(v)$ over $\{v : v \in (0, \infty)\}$,

$$u(v) = \begin{cases} 1, & v > h \quad (H_1), \\ 0, & v \le h \quad (H_0), \end{cases} \tag{38}$$

where $h > 0$ is a threshold of the test that is uniquely determined for a prescribed level of significance so that

$$\sup_{\theta \in \Theta_0} E_\theta \{u(v)\} = \alpha. \tag{39}$$

For fixed n, in terms of the probability density function (36), tables of the central F-distribution permit one to choose h to achieve the desired test size (false alarm probability P_{FA}),

$$P_{FA} = \alpha = \int_h^\infty f_{H_0}(v; n) dv. \tag{40}$$

Furthermore, once h is chosen, tables of the noncentral F-distribution permit one to evaluate, in terms of the probability density function (34), the power (detection probability P_D) of the test,

$$P_D = \gamma = \int_h^\infty f_{H_1}(v; n, q) dv. \tag{41}$$

The probability of a miss is given by

$$\beta = 1 - \gamma. \tag{42}$$

It follows from (36) and (40) that the GMLR test is invariant to intensity changes in the clutter background and achieves a fixed probability of a false alarm, that is, the resulting analyses indicate that the test has the property of a constant false alarm rate (CFAR). Also, no learning process is necessary in order to achieve the CFAR. Thus, operating in accordance to the local clutter situation, the test is adaptive.

About the Authors

Dr. Nicholas A. Nechval is a Professor and Head, Department of Mathematical Statistics, EVF Research Institute, University of Latvia, Riga, Latvia. He is also a Principal Investigator in the Institute of Mathematics and Computer Science at the University of Latvia. Dr. Nechval was a Professor of Mathematics and Computer Science and the Head of the Research Laboratory at the Riga Aviation University (1993–1999). In 1992, Dr. Nechval was awarded a Silver Medal of the Exhibition Committee (Moscow, Russia) for his research on the problem of Prevention of Collisions between Aircraft and Birds. He is a Member of the Russian Academy of Science. Professor Nechval has authored and coauthored more than 350 papers and 9 books, including the book *Aircraft Protection from Birds* (Moscow: Russian Academy of Science, 2007) coauthored with V.D. Illyichev (Academician of the Russian Academy of Science), and the book *Improved Decisions in Statistics* (Riga: SIA "Izglitibas soli", 2004) coauthored with E.K. Vasermanis. This book was awarded the "2004 Best Publication Award" by the Baltic Operations Research Society. Dr. Nechval is also an Associate editor of the following international journals: *Scientific Inquiry* (2005–), *An International Journal of Computing Anticipatory Systems* (2002–), et al.

Dr. Konstantin N. Nechval is an Assistant Professor, Applied Mathematics Department, Transport and Telecommunication Institute, Riga, Latvia. He has authored and co-authored more than 50 papers. Dr. Konstantin N. Nechval was awarded the "CASYS'07 Best Paper Award" for his paper: "Dual Control of Education Process" presented at the Eight International Conference on Computing Anticipatory Systems (Liege, Belgium, August 6-11, 2007) and the "MM2009 Best Paper Award" for his paper: "Optimal Statistical Decisions in a New Product Lifetime Testing" presented at the Fourth International Conference on Maintenance and Facility Management (Rome, Italy, April 22-24, 2009).

Dr. Maris Purgailis is a Professor and Dean, Faculty of Economics and Management, University of Latvia, Riga, Latvia. Professor Purgailis has authored and co-authored more than 120 papers and 6 books.

Cross References

►Data Analysis
►Fuzzy Sets: An Introduction
►Pattern Recognition, Aspects of
►Statistical Signal Processing

References and Further Reading

Abramowitz M, Stegun IA (1964) Handbook of mathematical functions. National Bureau of Standards, New York

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Nechval NA (1992) Radar CFAR thresholding in clutter under detection of airborne birds. In: Proceedings of the 21st meeting of bird strike committee Europe. BSCE, Jerusalem, pp 127–140

Nechval NA (1997) Adaptive CFAR tests for detection of a signal in noise and deflection criterion. In: Wysocki T, Razavi H, Honary B (eds) Digital signal processing for communication systems. Kluwer, Boston, pp 177–186

Nechval NA, Nechval KN (1998) Recognition of applicant for project realization with good contract risk. In: Pranevicius H, Rapp B (eds) Organisational structures, management, simulation of business sectors and systems. Kaunas University of Technology, Lithuania, pp 70–72

Nechval NA, Nechval KN (1999) CFAR test for moving window detection of a signal in noise. In: Proceedings of the 5th international symposium on DSP for communication systems, Curtin University of Technology, Perth-Scarborough, pp 134–141

Nechval NA, Nechval KN, Srelchonok VF, Vasermanis EK (2004) Adaptive CFAR tests for detection and recognition of targets signals in radar clutter. In: Berger-Vachon C, Gil Lafuente AM (eds) The 2004 conferences best of. AMSE Periodicals, Barcelona, pp 62–80

Statistical Publications, History of

Vassiliy Simchera
Director of Rosstat's Statistical Research Institute
Moscow, Russia

Statistical publications are editions that contain summarized numerical data about socio-economic phenomena, usually presented in the form of statistical tables, charts, diagrams, graphs, etc. These statistical publications are an inseparable part of common numerical information concerning the state and development of healthcare, education, science, and culture provided by the statistical authorities.

Depending on the common purpose, one may distinguish their various types. These include *Statistical Yearbook, Annual Statistics; Statistical Abstract; Manual Guide, Handbook of Statistics*; overview of census, and other major surveys. By order of coverage, statistical publications can be common (*National Accounts*), industrial (*Industrial Indicators*), or may deal with other activities of an economy (for example, *Financial Statistics*). By the level of details they can be complete (*Yearbooks, Almanacs*, etc.) or short (*Pocket Book and Statistisches Handbook* are the most common types).

There are also differences by the domain of coverage among one or another statistical publication: an entire country, an administrative territorial part of the country (for example, state, region, land, county, etc.); in international statistical publications, this could be several countries, an entire continent, or the whole world (for example, *UN Statistical Publications*).

The outcomes of large surveys are presented in nonrecurrent statistical publications; among the recurrent statistical publications, the most significant are periodical statistical publications (published annually, quarterly, or monthly), the least significant are non-periodical statistical publications (containing demographic figures, birth and death rates, marriage status, etc.).

The statistical publications cover current and previous years (retrospective statistical publications) with the scope of decades and centuries (*Historical Statistics of the US from* 1789, *Colonial Times to* 1957, 1960, 1975, *and* 2008; *USSR's Economics* 60 *years*, 1987; *Russia*: 100 *Years of Economic Growth* 1900–2008 *Historical Series; Annuaire Statistique de la France, vols*. 1–106, 1878–2003).

Statistical publications have various forms of editions: yearbooks, reports, series of books (for example, a census of the population), bulletins, and journals, "notebooks", which contain statistical reviews (quarterly, monthly, *Bulletin of Statistics, Journal of Statistics, Survey of Statistics and Review of Statistics*), summaries, and reports.

The form and content of statistical publications have been changing along with history.

The first statistical publications (similar to modern ones) appeared in 15^{th} century in Venice and then later on in Holland (a series of 60 small volumes under a common name "Elsevier republics," from 1624). In England numerical statistical figures appeared in the 17^{th} century in works by the founders of "political arithmetic," William Petty and John Graunt, and in the 18^{th} century in the works by Gregory King. In Germany ("The Holy Roman Empire of the German Nation"), the second half of the 17^{th} and 18^{th} centuries were predominated by "descriptive government statistics" (H. Conring, G. Achenwall, A. L. Schlözer); only in the last quarter of the 18^{th} century did a new type of statistical publications appeared, i.e., the works of "linear arithmeticians" tending to represent numerical data about one or several countries in the shape of statistical graphs—diagrams and cartograms (the founder of these statistical publications is August Friedrich Crome, who published "*Producten-Karte von Europa*" (1782) and *Über die Größe und Bevölkerung der sämtlichen europäischen Staaten* (1785)). In Russia, the first statistical publications date back to 1831 (historical, ethnographic, and economic atlases with a statistical description of Russia by I. K. Kirilov). The classified yearbooks (with the scope of data for a period of 100 years and more by various types of figures describing territories, natural resources, population, GDP, standard of living etc.) of the USA have been published in the United States since 1878 (125 yearbooks), in Great Britain since 1850 (150 yearbooks), in France since 1860 (85 yearbooks of old series and 23 of new series), in Germany since 1872, in Canada since 1818, in Sweden since 1915, and in Japan since 1818.

Apart from yearbooks there are also many other specialized statistical publications, the most important among them being "Census of Population," "Census of Manufacturers," etc., annual surveys on separate industries "Annual Survey on Manufacturers," enterprises "Moody's manual" in the U.S. "Compas" in Germany, France, and Belgium, and also personal references such as "Who's Who," "Who's Who in the world," "Poor's Register of Corporations Directors and Executives," "Great Minds of the 21^{st} Century," etc.

The first international statistical dictionary was by Michael G. Mulhall, "The Dictionary of Statistics," which ran into several editions (1884, 1892, 1899, 1909) included figures on 30–50 countries for a period from 1800 to 1900. Augustus D. Webb's "The New Dictionary of Statistics"x covered 1896–1905. From 1916 to 1926, the International

Statistical Institute (ISI) published the "International Statistical Yearbook" (from 1853–1876 there were editions from the International Statistical Congresses). With the establishment of the League of Nations (1919) the number of statistical publications increased. The significant statistical publications by the League of Nations were "Statistical Yearbook of the League of Nations" (11 yearbooks for a period from 1932 to 1945), "Monthly Bulletin of Statistics," "World Economic Surveys" (1933–1945, 11 issues), "World Production and Prices" (1925–1939, 7 issues), "Review of World Trade" (1932–1939, 8 issues), etc. In 1919, the International Labour Organization began publication of the "Yearbook of Labour Statistics," and in 1921 the International Institute of Agriculture started publication of the "International Yearbook of Agriculture Statistics".

In 1949, the United Nations Organization (UN) and its specialized institutions started a new stage of statistical publications subdivided into nine series - A, B, C, D, J, K, M, P, F. The most important of them are: "Statistical Yearbook," "Demographic Yearbook," "Yearbook of National Accounts Statistics," "Yearbook of International Trade Statistics," "Balance of Payments Yearbook," "Annual Epidemiological and Vital Statistics," "United Nations Juridical Yearbook," and "Yearbook of the United Nations."

The Food and Agriculture Organization publishes "Yearbook of Food and Agricultural Statistics," "Yearbook of Fishery Statistics," and "Yearbook of Forest Products."

UNESCO publishes "International Yearbook of Education," "Yearbook of Youth Organizations," and "UNESCO Statistical Yearbook."

EU, OECD, WHO, EuroStat, IMF, and World Bank have their own statistical publications. The most important statistical publications are world economic reviews (published separately by the UN and its commissions for Europe, Asia, Africa and Latin America, on annual basis) and various statistical editions. There are also statistical journals, for example, the UN's "Monthly Bulletin of Statistics" and the UN's reference books, "World Weight and Measures," "Nomenclature of Geographic Areas for Statistical Purposes," "Name's of Countries and Adjectives of Nationality," etc. The international bibliographies, indexes, dictionaries, and encyclopedias are also considered to be statistical publications.

The specialized editions and international statistical classifiers, questionnaires, systems, methods, and standards (there are over 120,000 of titles including 175 standard classifiers in the world) regulate the procedures of the international comparisons, the most recognized standards of which are UN's System of National Accounts, trade, banking and monetary transactions, and standards of EuroStat and IMF on the statistical ethics and assessment of data quality.

About the Author

For Biography *see* the entry ►Actuarial Methods.

Cross References

►Census
►Eurostat
►Statistics, History of

References and Further Reading

Nixon JW (1960) A history of the International Statistical Institute 1855–1960. International Statistical Institute, Hague
Simchera VM, Sokolin VL (2001) Encyclopedia of statistical publications X–XX centuries. Financy i Statistika, Moscow
Simchera VM (2006) Russia: 100 years of economic growth: 1900–2000: historical series, trends of centuries, institutional cycles. Nauka, Moscow

Statistical Quality Control

M. Ivette Gomes
Professor
Universidade de Lisboa, DEIO and CEAUL, Lisboa, Portugal

Quality: A Brief Introduction

The main objective of *statistical quality control* (SQC) is to achieve *quality* in production and service organizations, through the use of adequate statistical techniques. The following survey relates to manufacturing rather than to the service industry, but the principles of SQC can be successfully applied to either. For an example of how SQC applies to a service environment, see Roberts (2005). *Quality* of a product can be defined as its adequacy to be used (Montgomery 2009), which is evaluated by the so-called *quality characteristics*. Those are random variables in a probability language, and are usually classified as: *physical*, like length and weight; *sensorial*, like flavor and color; *temporally oriented*, like the maintenance of a system.

Quality Control (QC) has been an activity of engineers and managers, who have felt the need to work jointly with statisticians. Different quality characteristics are measured and compared with pre-determined specifications, the *quality norms*. QC began a long time ago, when manufacturing began and competition accompanied it,

with consumers comparing and choosing the most attractive product. The *Industrial Revolution*, with a clear distinction between producer and consumer, led producers to the need of developing methods for the control of their manufactured products. On the other hand, SQC is comparatively new, and its greatest developments have taken place during the twentieth century. In 1924, at the Bell Laboratories, Shewhart developed the concept of *control chart* and, more generally, *statistical process control* (SPC), shifting the attention from the product to the production process (Shewhart 1931). Dodge and Romig (1959), also in the Bell Laboratories, developed *sampling inspection*, as an alternative to the 100% inspection.

Among the pioneers in SPC we also distinguish W.E. Deming, J.M. Juran, P.B. Crosby and K. Ishikawa (see other references in Juran and Gryna 1993). But it was during the *Second World War* that there was a generalized use and acceptance of SQC, largely used in USA and considered as primordial for the defeat of Japan. In 1946, the *American Society for Quality Control* was founded, and this enabled a huge push to the generalization and improvement of SQC methods.

After the II World War, Japan was confronted with rare food and lodging, and the factories were in ruin. They evaluated and corrected the causes of such a defeat. The quality of the products was an area where USA had definitely over passed Japan, and this was one of the items they tried to correct, becoming rapidly masters in inspection sampling and SQC, and leaders of quality around 1970. Recently, the quality developments have also been devoted to the motivation of workers, a key element in the expansion of the Japanese industry and economy.

Quality is more and more the prime decision factor in the consumer preferences, and quality is often pointed out as the key factor for the success of organizations. The implementation of a *production* QC clearly leads to a reduction in the manufacturing costs, and the money spent with control is almost irrelevant. At the moment, the quality improvement in all areas of an organization, a philosophy known as *Total Quality Management* (TQM) is considered crucial (see Vardeman and Jobe 1999). The challenges are obviously difficult. But the modern SQC methods surely provide a basis for a positive answer to these challenges. SQC is at this moment much more than a set of *statistical instruments*. It is a global way of thinking of workers in an organization, with the objective of making things *right in the first place*. This is mainly achieved through the systematic *reduction of the variance* of relevant quality characteristics.

Usual Statistical Techniques in SQC

The statistical techniques useful in SQC are quite diverse. In this survey, we shall briefly mention SPC, an on-line control technique of a process production with the use of ►*control charts*. ►*Acceptance sampling*, performed out of the line production (before it, for sentencing incoming batches, and after it, for evaluating the final product), is another important topic in SQC (see Duncan [1986] and Pandey [2007], among others). A similar comment applies to *reliability theory* and *reliability engineering*, off-line techniques performed when the product is complete, in order to detect the resistance to failure of a device or system (see Pandey [2007], also among others).

It is however sensible to mention that, additionally to these techniques, there exist other statistical topics useful in the *improvement* of a process. We mention a few examples: in a line of production, we have the *input variables*, the *manufacturing process* and the *final product* (output). It is thus necessary to model the relationship between input and output. Among the statistical techniques useful in the building of these models, we mention *Regression* and *Time Series Analysis*. The area of *Experimental Design* (see Taguchi et al. 1989) has also proved to be powerful in the detection of the most relevant input variables. Its adequate use enables a reduction of variance and the identification of the controllable variables that enable the optimization of the production process.

Statistical Process Control (SPC). Key monitoring and investigating tools in SPC include *histograms*, *Pareto charts*, *cause and effect diagrams*, *scatter diagrams* and *control charts*. We shall here focus on control chart methodology.

A *control chart* is a popular statistical tool for monitoring and improving quality, and its success is based on the idea that no matter how well the process is designed, there exists a certain amount of nature variability in output measurements. When the variation in process quality is due to random causes alone, the process is said to be *in-control*. If the process variation includes both random and special causes of variation, the process is said to be *out-of-control*. The control chart is supposed to detect the presence of special causes of variation.

Generally speaking, the main steps in the construction of a control chart, performed at a *stable* stage of the process, are the following: determine the process parameter you want to monitor, choose a convenient statistic, say W, and create a *central line* (CL), a *lower control limit* (LCL) and an *upper control limit* (UCL). Then, sample the

production process along time, and group the process measurements into *rational subgroups* of size n, by time period t. For each rational subgroup, compute w_t, the observed value of W_t, and plot it against time t. The majority of measurements should fall in the so-called *continuation interval* $C = [LCL, UCL]$. Data can be collected at *fixed sampling intervals* (FSI), with a size equal to d, or alternatively, at *variable sampling intervals* (VSI), usually with sampling intervals of sizes d_1, d_2 $(0 < d_1 < d_2)$. The region C is then split in two disjoint regions C_1 and C_2, with C_2 around CL. The sampling interval d_1 is used as soon as a measurement falls in C_1; otherwise, it is used the largest sampling interval d_2. If the measurements fall within LCL and UCL no action is taken and the process is considered to be *in-control*. A point w_t that exceeds the control limits signals an alarm, i.e., it indicates that the process is *out of control*, and some action should be taken, ranging from taking a re-check sample to the tracing and elimination of these causes. Of course, there is a slight chance that is a *false alarm*, the so-called α-risk. The design of control charts is a compromise between the risks of not detecting real changes (β-risks) and of α-risks. Other relevant *primary characteristics* of a chart are the *run length* (RL) or *number of samples to signal* (NSS) and the associated mean value, the *average run length*, ARL=$\mathbb{E}(RL) = 1/(1-\beta)$, as well as the *capability indices*, C_k and C_{pk} (see Pearn and Kotz 2006). Essentially, a control chart is a test, performed along time t, of the hypothesis H_0 : the process is in-control *versus* H_1 : the process is out-of-control.

Stated differently, we use historical data to compute the initial control limits. Then the data are compared against these initial limits. Points that fall outside of the limits are investigated and, perhaps, some will later be discarded. If so, the limits need to be recomputed and the process repeated. This is referred to as *Phase* I. Real-time process monitoring, using the limits from the end of Phase I, is *Phase* II. There thus exists a strong link between control charts and hypothesis testing performed along time.

Note that a *preliminary statistical data analysis* (usually *histograms* and *Q-Q plots*) should be performed on the prior collected data. A common assumption in SPC is that quality characteristics are distributed according to a *normal* distribution. However, this is not always the case, and in practice, if data seem very far from meeting this assumption, it is common to transform them through a ▶Box–Cox transformation (Box and Cox 1964). But much more could be said about the case of nonnormal data, like the use of robust control charts (see Figueiredo and Gomes [2004], among others).

With its emphasis on early detection and prevention of problems, SPC has a distinct advantage over quality methods such as inspection, that apply resources to detecting and correcting problems in the final product or service. In addition to reducing waste, SPC can lead to a reduction in the time required to produce the final products. SPC is recognized as a valuable tool from both a cost reduction and a customer satisfaction standpoint. SPC indicates when an action should be taken in a process, but it also indicates when no action should be taken.

Classical Shewhart Control Charts: A Simple Example. In this type of charts, measurements are assumed to be independent and distributed according to a normal distribution. Moreover, the statistics W_t built upon those measurements are also assumed to be independent. The main idea underlying these charts is to find a simple and convenient statistic, W, with a sampling distribution easy to find under the validity of the *in-control* state, so that we can easily construct a confidence interval for a location or spread measure of that statistic. For continuous quality characteristics, the most common Shewhart-charts are the average chart ($\overline{X}$-chart) and the range chart (R-chart), as an alternative to the standard-deviation chart (S-chart). For discrete quality characteristics, the most usual charts are the p-charts and np-charts in a $Binomial(n, p)$ background, and the so-called c-charts and u-charts for $Poisson(c)$ backgrounds.

Example 1 ($\overline{X}$-chart). *Imagine a breakfast cereal packaging line, designed to fill each cereal box with* 500 *grams of product. The production manager wants to monitor on-line the mean weight of the boxes, and it is known that, for a single pack, an estimate of the weight standard-deviation* σ *is* 10 *g. Daily samples of* $n = 5$ *packs are taken during a stable period of the process, the weights* $x_i, 1 \le i \le n$, *are recorded, and their average,* $\overline{x} = \sum_{i=1}^{n} x_i/n$, *is computed. These averages are estimates of the process mean value* μ, *the parameter to be monitored. The center line is* $CL = 500$ *g (the target). If we assume that data are normally distributed, i.e.,* $X \sim N(\mu = 500, \sigma = 10)$, *the control limits can be determined on the basis that* $\overline{X} \sim N(\mu = 500, \sigma/\sqrt{n} = 10/\sqrt{5} = 4.472)$. *In-control, it thus expected that* $100(1-\alpha)\%$ *of the average weights are between* $500 + 4.472\ \xi_{\alpha/2}$ *and* $500 - 4.472\ \xi_{\alpha/2}$ *where* $\xi_{\alpha/2}$ *is the* $(\alpha/2)$*-quantile of a standard normal distribution. For a* α*-risk equal to* 0.002 *(a common value in English literature),* $\xi_{\alpha/2} = -3.09$. *The American Standard is based on "*3 *– sigma" control limits (corresponding to* 0.27% *of false alarms), while the British Standard uses*

"3.09–sigma" limits (corresponding to 0.2% of false alarms). In this case, the 3-sigma *control limits are* $LCL = 500 - 3 \times 10/\sqrt{5} = 486.584$ *and* $UCL = 500 + 3 \times 10/\sqrt{5} = 513.416$.

Other Control Charts. Shewhart-type charts are efficient in detecting medium to large shifts, but are insensitive to small shifts. One attempt to increase the power of these charts is by adding supplementary stopping rules based on runs. The most popular stopping rules, supplementing the ordinary rule, "one point exceeds the control limits," are: two out of three consecutive points fall outside warning (2-sigma) limits; four out of five consecutive points fall beyond 1-sigma limits; eight consecutive points fall on one side of the centerline.

Another possible attempt is to consider some kind of dependency between the statistics computed at the different sampling points. To control the mean value of a process at a target μ_0, one of the most common control charts of this type is the *cumulative sum* (CUSUM) chart, with an associated control statistic given by $S_t := \sum_{j=1}^{t}(x_j - \mu_0) = S_{t-1} + (\overline{x}_t - \mu_0)$, $t = 1, 2, \cdots \quad (S_0 = 0)$. Under the validity of $H_0 : X \frown N(\mu_0, \sigma)$, we thus have a *random walk* with null mean value (see ►Random Walk). It is also common to use the *exponentially weighted moving average* (EWMA) statistic, given by $Z_t := \lambda\overline{x}_t + (1-\lambda)Z_{t-1} = \lambda\sum_{j=0}^{t-1}(1-\lambda)^j\,\overline{x}_{t-j} + (1-\lambda)^t Z_0$, $t = 1, 2, \ldots, \quad Z_0 = \overline{\overline{x}}, \quad 0 < \lambda < 1$, where $\overline{\overline{x}}$ denotes the overall average of a small number of averages collected *a priori*, when the process is considered stable and in-control. Note that it is also possible to replace averages by individual observations (for details, see Montgomery 2009).

ISO 9000, Management and Quality

The main objective of this survey was to speak about statistical instruments useful in the improvement of quality. But these instruments are a small part of the total effort needed to achieve quality. Nowadays, essentially due to an initiative of the International Organization for Standardization (ISO), founded in 1946, all organizations are pushed towards quality. In 1987, ISO published the ISO 9000 series, with general norms for quality management and quality guarantee, and additional norms were established later on diversified topics. The ISO 9000 norms provide a guide for producers, who want to implement efficient quality. They can also be used by consumers, in order to evaluate the producers' quality. In the past, the producers were motivated to the establishment of quality through the increasing satisfaction of consumers. Nowadays, most of the them are motivated by the ISO 9000 certification – if they do not have it, they will lose potential clients.

Regarding management and quality: as managers have a final control of all organization resources, management has a ultimate responsibility in the quality of all products. Management should thus establish a quality policy, making it perfectly clear to all workers (see Burrill and Ledolter 1999, for details).

Acknowledgment

Research partially supported by FCT/OE, POCI 2010 and PTDC/FEDER.

About the Author

Dr. Gomes is Professor of Statistics at the Department of Statistics and Operations Research (DEIO), Faculty of Science, University of Lisbon. She is Past President of Portuguese Statistical Society (1989–1993). She is Founding Editor, *Revstat* (2003–), Associate Editor of *Extremes* (2007–) and Associate Editor of *J. Statistical Planning and Inference* (2007–).

Cross References

►Acceptance Sampling
►Box–Cox Transformation
►Control Charts
►Random Walk
►Rao–Blackwell Theorem
►Relationship Between Statistical and Engineering Process Control
►Statistical Design of Experiments (DOE)
►Statistical Quality Control: Recent Advances

References and Further Reading

Burril CW, Ledolter J (1999) Achieving quality through continual improvement. Wiley, New York

Box GEP, Cox DR (1964) An analisys of transformations. J R Stat Soc B26:211–256

Dodge HF, Romig HG (1959) Sampling inspection tables, single and double sampling, 2nd edn. Wiley

Duncan AJ (1986) Quality control and industrial statistics, 5th edn. Irwin, Homehood

Figueiredo F, Gomes MI (2004) The total median in statistical quality control. Appl Stoch Model Bus 20(4):339–353

Juran JM, Gryna FM (1993) Quality planning and analysis. MacGraw-Hill, New York

Montgomery DC (2009) Statistical quality control: a modern introduction, 6th edn. Wiley, Hoboken, NJ

Pandey BN (2007) Statistical techniques in life-testing, reliability, sampling theory and quality control. Narosa, New Delhi

Pearn WL, Kotz S (2006) Encyclopedia and handbook of process capability indices: a comprehensive exposition of quality control measures. World Scientific, Singapore

Roberts L (2005) SPC for right-brain thinkers: process control for non-statisticians. Quality, Milwaukee
Shewhart WA (1931) Economic control of quality of manufactured product. Van Nostrand, New York
Taguchi G, Elsayed E, Hsiang T (1989) Quality engineering in production systems. Mc-Graw-Hill, New York
Vardeman S, Jobe JM (1999) Statistical quality assurance methods for engineers. Wiley, New York

Statistical Quality Control: Recent Advances

Fugee Tsung[1], Yanfen Shang[2], Xianghui Ning[2]
[1]Professor and Head
Hong Kong University of Science and Technology, Hong Kong, China
[2]Hong Kong University of Science and Technology, Hong Kong, China

Statistical quality control aims to achieve the product or process quality by utilizing statistical techniques, in which statistical process control (SPC) has been demonstrated to be one primary tool for monitoring the process or product quality. Since 1920s, the control chart, as one of the most important SPC techniques, has been widely studied.

Univariate Control Charts Versus Multivariate Control Charts

In terms of the number of variables, ▶control charts can be classified into two types, that is, univariate control charts and multivariate control charts.

The performance of the conventional univariate control charts, including Shewhart control charts, cumulative sum (CUSUM) control charts and exponentially weighted moving average (EWMA) control charts have been extensively reviewed. The research demonstrates that the Shewhart chart is more sensitive to large shifts than the EWMA and CUSUM chart and vice versa. These traditional control charts usually assume that the observations are independent and identically follow the normal distribution. In some practical situations, however, these assumptions are not valid. Therefore, other control charts that are different or extended from the traditional charts are developed for some special cases, such as monitoring autocorrelated processes and/or processes with huge sample data, detecting dynamic mean change and/or a range of mean shifts. See Han and Tsung (2005, 2006, 2007, 2009), Han et al. (2007a, b), Wang and Tsung (2005), Zhao et al. (2005) and Zou et al. (2008c) for detailed discussion.

Although the aforementioned univariate charts perform well in monitoring some process or product qualities, their performance is not satisfactory when the quality of a product or process is characterized by several correlated variables. Therefore, multivariate statistical process control (MSPC) techniques were developed and widely applied. Hotelling's T^2 chart, the traditional multivariate control chart, was proposed in 1947 (Hotelling 1947) to deal with the multivariate monitoring case, which assumed that several variables follow the multivariate normal distribution (see ▶Multivariate Normal Distributions). Following that, a variety of studies extended this research further. Among others, see Tracy et al. (1992), Mason et al. (1995), and Sullivan and Woodall (1996) for discussion concerning the property and performance of the T^2 chart.

Besides the Hotelling's T^2 chart, the other traditional multivariate control charts include the Multivariate cumulative sum (MCUSUM) chart presented by Crosier (1988) and Pignatiello and Runger (1990) and the multivariate exponentially weighted moving average (MEWMA) chart proposed by Lowry et al. (1992). Similarly to Hotelling's T^2, these two charts are sensitive to moderate and small mean shifts. Other extensions of traditional MSPC techniques, i.e., adaptive T^2 chart for dynamic processes (see Wang and Tsung (2007, 2008)), have been analyzed. Besides the multivariate charts for mean shifts, the multivariate charts for monitoring the process variation were also presented recently, such as the multivariate exponentially weighted mean squared deviation (MEWMS) chart and a multivariate exponentially weighted moving variance (MEWMV) chart (Huwang et al. (2007)). The extensive literature reviews were provided by Kourti and MacGregor (1996) and Bersimis et al. (2007), in which other statistical methods applied in MSPC, i.e., ▶principal component analysis (PCA) and partial least square (PLS), are also reviewed.

Most of the mentioned charts have a common assumption that process variables follow normal distributions. When there is no distribution assumption, nonparametric methods, like the depth function (Zuo and Serfling (2000)), can be used, the advantages of which are examined by Chakraborti et al. (2001). However, with the development of technology, a more complicate situation occurs. Numerical process variables may be mixed up with the categorical process variables to represent the real condition of a process. Direct application of the aforementioned methods may lead to inappropriate ARL and unsatisfactory false alarms. An alternative way to solve this problem is to use some distribution-free methods, like the K-chart proposed by Sun and Tsung (2003). More research is needed in this area.

SPC for Profile Monitoring

In most SPC applications, either in the univariate or multivariate cases, it is assumed that the quality of a process or product can be adequately represented by the distribution of a single quality characteristic or by the general multivariate distribution of the several correlated quality characteristics. In some practical situations, however, the quality of a process or product is better characterized and summarized by a relationship between a response variable and one or more explanatory variables (Woodall et al. 2004). Therefore, studies on profile monitoring have been steadily increasing.

The early research on profile monitoring usually assumes that the relationship can be represented by the linear model. There has been extensive existing research on linear profile monitoring in the literature. For example, as early as 2000, Kang and Albin presented two methods in order to monitor the linear profiles. One approach is to monitor the intercept and slope of the linear model by constructing the multivariate chart (T^2 chart). The other is to monitor the average residuals by using the exponential weighted moving average (EWMA) chart and rang (R) chart simultaneously. It can be noted that some different control schemes were also developed for solving different linear profile monitoring problems, i.e., the self-staring control chart for linear profiles with unknown parameters (Zou et al. (2007a)). In addition, Zou et al. (2007b) proposed a multivariate EWMA (MEWMA) scheme for monitoring the general linear profile. Furthermore, recent studies on the nonlinear profile monitoring can be sourced in the relevant literature. Among others, the nonparametric methods are commonly used in monitoring the nonlinear profiles (see Zou et al. 2008b, Jensen et al. 2009). Besides, Woodall et al. (2004) provided an extensive review on profile monitoring. Recent research focused on the control scheme for monitoring profiles with categorical data rather than continuous data (Yeh et al. 2009)), in which a Phase I monitoring scheme for profiles with binary output variables was proposed.

SPC for Processes with Multiple Stages

In modern manufacturing and service environments, it is very common that most manufacturing and/or service processes involve a large number of operating stages rather than one single stage. Many examples of such multistage processes can be found in semiconductor manufacturing, automobile assembly lines and bank services, etc. For instance, the print circuit board (PCB) manufacturing process includes several stages, that is, exposure to black oxide, lay-up, hot press, cutting, drilling, and inspection. However, most of the abovementioned conventional SPC methods focus on single-stage processes without considering the multistage scenario, which do not consider the relationship among different stages. Therefore, the recent research on multistage processes has been widely conducted.

The existing popular SPC methods for multistage processes usually involve three types of approaches, which are the regression adjustment method, the cause-selecting method and methods based on linear state space models. The regression adjustment method was developed by Hawkins (1991, 1993), while Zhang (1984, 1985, 1989, 1992) proposed the cause-selecting method. A review of the cause-selecting method can be found in Wade and Woodall (1993). Recent research on the use of cause-selecting charts for multistage processes can be found in Shu et al. (2003), Shu and Tsung (2003), Shu et al. (2004) and Shu et al. (2005). A variety of current studies on multistage processes also adopt engineering models with a linear state space model structure. This model incorporates physical laws and engineering knowledge in order to describe the quality linkage among multiple stages in a process. Latest works on multistage process monitoring and diagnosis can be referred to Xiang and Tsung (2008), Zou et al. (2008a), Jin and Tsung (2009), and Li and Tsung (2009). With respect to multistage processes with categorical variables, some monitoring schemes were developed recently. For example, Skinner et al. (2003, 2004) proposed the generalized linear model (GLM)-based control chart for the Poisson data obtained from multiple stages.

An extensive review on the quality control of multistage systems including monitoring and diagnosing schemes was presented by Shi and Zhou (2009).

SPC Applications in Service Industries

SPC techniques can be applied in different industries such as manufacturing or service industries, although most of these techniques are originally developed for manufacturing industries, i.e., machining processes, assembly processes, semiconductor processes etc. Because the SPC techniques have been demonstrated to be efficient for manufacturing processes, the application of these techniques in service processes was argued in some papers (see Wyckoff (1984), Palm et al. (1997) and Sulek (2004)). In the existing literature, several control charts have been applied in service processes, i.e., quick service restaurant, the auto loan process that provides better service from the loan company to car dealers and buyers, and invoicing processes. See Apte and Reynolds (1995), Mehring (1995), Cartwright and Hogg (1996) for detailed discussion. In addition, the control charts were also widely applied in health-care and public-health fields (see Wardell and Candia (1999),

Green (1999)). Recently, Woodall (2006) discussed in great detail different control charts that have been proposed in health-care and public-health fields. Both the manufacturing process and the service operation process involve multiple operating stages rather than a single stage. Therefore, Sulek et al. (2005) proposed to use the cause selecting control chart for monitoring the service process with multiple stages in the grocery store and showed that it outperformed the Shewhart chart in monitoring the multistage service process. More recent studies on the application of SPC techniques, especially in service industries, were reviewed by Maccarthy and Wasusri (2002) and Tsung et al. (2008). All these applications showed that SPC techniques were efficient in monitoring and identifying service processes.

Statistical Process Control as one primary tool for quality control is very efficient and important in monitoring the process/product quality. SPC techniques will be applied in more industries with different characteristics. Therefore, more advanced studies on SPC schemes will be widely conducted in order to achieve the quality required for products or processes.

About the Author

Dr. Fugee Tsung is Professor and Head of the Department of Industrial Engineering and Logistics Management (IELM), Director of the Quality Lab, at the Hong Kong University of Science & Technology (HKUST). He is a Fellow of the Institute of Industrial Engineers (IIE), Fellow of the American Society for Quality (ASQ) and Fellow of the Hong Kong Institution of Engineers (HKIE). He received both his MSc and PhD from the University of Michigan, Ann Arbor and his BSc from National Taiwan University. He is currently Department Editor of the IIE Transactions, Associate Editor of *Technometrics*, *Naval Research Logistics*, and on the Editorial Boards for *Quality and Reliability Engineering International* (QREI). He is an ASQ Certified Six Sigma Black Belt, ASQ authorized Six Sigma Master Black Belt Trainer, Co-funder and Chair of the Service Science Section at INFORMS, Regional Vice President (Asia) of IIE. He has authored over 70 refereed journal publications, and is also the winner of the Best Paper Award for the IIE Transactions in 2003 and 2009. His research interests include quality engineering and management to manufacturing and service industries, statistical process control, monitoring and diagnosis.

Cross References

►Control Charts
►Moving Averages
►Multivariate Normal Distributions
►Multivariate Statistical Process Control
►Relationship Between Statistical and Engineering Process Control
►Statistical Quality Control

References and Further Reading

Apte UM, Reynolds CC (1995) Quality management at Kentucky fried chicken. Interfaces 25:6–21
Bersimis S, Psarakis S, Panaretos J (2007) Multivariate statistical process control charts: an overview. Quality Reliab Eng Int 23(5):517–543
Cartwright G, Hogg B (1996) Measuring processes for profit. The TQM Magazine 8(1):26–30
Chakraborti S, Van der Laan P, Bakir ST (2001) Nonparametric control charts: an overview and some results. J Qual Technol 33:304–315
Crosier RB (1988) Multivariate generalizations of cumulative sum quality-control schemes. Technometrics 30(3):291–303
Green RS (1999) The application of statistical process control to manage global client outcomes in behavioral healthcare. Eval Program Plann 22:199–210
Han D, Tsung F (2005) Comparison of the Cuscore. GLRT and CUSUM control charts for detecting a dynamic mean change. Ann Inst Stat Math 57:531–552
Han D, Tsung F (2006) A reference-free cuscore chart for dynamic mean change detection and a unified framework for charting performance comparison. J Am Stat Assoc 101:368–386
Han D, Tsung F (2007) Detection and diagnosis of unknown abrupt changes using CUSUM multi-chart schemes. Sequential Anal 26:225–249
Han D, Tsung F (2009) Run length properties of the CUSUM and EWMA control schemes for stationary autocorrelated processes. Statistica Sinica 19:473–490
Han D, Tsung F, Li Y (2007a) A CUSUM chart with local signal amplification for detecting a range of unknown shifts. Int J Reliab Qual Saf Eng 14:81–97
Han D, Tsung F, Hu X, Wang K (2007b) CUSUM and EWMA multi-charts for detecting a range of mean shifts. Statistica Sinica 17:1139–1164
Hawkins DM (1991) Multivariate quality control based on regression-adjusted variables. Technometrics 33:61–75
Hawkins DM (1993) Regression adjustment for variables in multivariate quality control. J Qual Technol 25:170–182
Hotelling H (1947) Multivariate quality control. In: Eisenhart C, Hastay M, Wallis WA (eds) Techniques of statistical analysis. McGraw-Hill, New York
Huwang L, Yeh AB, Wu C (2007) Monitoring multivariate process variability for individual observations. J Qual Technol 39(3):258–278
Jensen WA, Birch JB, Woodall WH (2009) Profile monitoring via nonlinear mixed models. J Qual Technol 41:18–34
Jin M, Tsung F (2009) A chart allocation strategy for multistage processes. IIE Trans 41(9):790–803
Kang L, Albin SL (2000) On-line monitoring when the process yields a linear profile. J Qual Technol 32:418–426
Kourti T, MacGregor JF (1996) Multivariate SPC methods for process and product monitoring. J Qual Technol 28(4):409–428
Li Y, Tsung F (2009) False discovery rate-adjusted charting schemes for multistage process monitoring and fault identification. Technometrics 51:186–205

Lowry A, Woodall WH, Champ CW, Rigdon SE (1992) A multivariate exponentially weighted moving average control chart. Technometrics 34(1):46–53
Maccarthy BL, Wasusri T (2002) A review of non-standard applications of statistical process control (SPC) charts. Int J Qual Reliab Manage 19(3):295–320
Mason RL, Tracy ND, Young JC (1995) Decomposition of T2 for multivariate control chart interpretation. J Qual Technol 27(2): 99–108
Mehring JS (1995) Achieving multiple timeliness goals for auto loans: a case for process control. Interfaces 25:81–91
Palm AC, Rodriguez RN, Spiring FA, Wheeler DJ (1997) Some perspectives and challenges for control chart methods. J Qual Technol 29:122–127
Pignatiello J, Runger GC (1990) Comparison of multivariate CUSUM charts. J Qual Technol 22:173–186
Shi J, Zhou S (2009) Quality control and improvement for multistage systems: a survey. IIE Trans 41:744–753
Shu LJ, Tsung F (2003) On multistage statistical process control. J Chin Inst Ind Eng 20:1–8
Shu LJ, Apley DW, Tsung F (2003) Autocorrelated process monitoring using triggered CUSCORE charts. Qual Reliab Eng Int 18:411–421
Shu LJ, Tsung F, Kapur KC (2004) Design of multiple cause-selecting charts for multistage processes with model uncertainty. Qual Eng 16:437–450
Shu LJ, Tsung F, Tsui KL (2005) Effects of estimation errors on cause-selecting charts. IIE Trans 37(6):559–567
Skinner KR, Montgomery DC, Runger GC (2003) Process monitoring for multiple count data using generalized linear model-based control charts. Int J Prod Res 41(6):1167–1180
Skinner KR, Montgomery DC, Runger GC (2004) Generalized linear model-based control charts for discrete semiconductor process data. Qual Reliab Eng Int 20:777–786
Sulek J (2004) Statistical quality control in services. Int J Serv Tech Manag 5:522–531
Sulek JM, Marucheck A, Lind MR (2005) Measuring performance in multi-stage service operations: an application of cause selecting control charts. J Oper Manag 24:711–727
Sullivan JH, Woodall WH (1996) A comparison of multivariate control charts for individual observations. J Qual Technol 28(4):398–408
Sun R, Tsung F (2003) A kernel-distance-based multivariate control chart using support vector methods. Int J Prod Res 41:2975–2989
Tracy ND, Young JC, Mason RL (1992) Multivariate control Charts for Individual Observations. J Qual Technol 24(2):88–95
Tsung F, Li Y, Jin M (2008) Statistical process control for multistage manufacturing and service operations: a review and some extensions. Int J Serv Oper Inform 3:191–204
Wade MR, Woodall WH (1993) A review and analysis of cause-selecting control charts. J Qual Technol 25(3):161–169
Wang K, Tsung F (2005) Using profile monitoring techniques for a data-rich environment with huge sample size. Qual Reliab Eng Int 21(7):677–688
Wang K, Tsung F (2007) Monitoring feedback-controlled processes using adaptive T2 schemes. Int J Prod Res 45:5601–5619
Wang K, Tsung F (2008) An adaptive T2 chart for monitoring dynamic systems. J Qual Technol 40:109–123
Wardell DG, Candia MR (1999) Statistical process monitoring of customer satisfactior survey data. Qual Manag J 3(4):36–50
Woodall WH (2006) The use of control charts in health-care and public-health surveillance. J Qual Technol 38(2):89–104
Woodall WH, Spitzner DJ, Montgomery DC, Gupta S (2004) Using control charts to monitor process and product quality profiles. J Qual Technol 36:309–320
Wyckoff DD (1984) New tools for achieving service quality. Cornell Hotel Rest Admin Quartr 25:78–91
Xiang L, Tsung F (2008) Statistical monitoring of multistage processes based on engineering models. IIE Trans 40(10):957–970
Yeh AB, Huwang L, Li YM (2009) Profile monitoring for binary response. IIE Trans 41(11):931–941
Zhang GX (1984) A new type of control charts and a theory of diagnosis with control charts. World Qual Congr Trans: 175–185
Zhang GX (1985) Cause-selecting control charts - a new type of quality control charts. The QR Journal 12:221–225
Zhang GX (1989) A new diagnosis theory with two kinds of quality. world quality congress transactions. Am Soc Qual Control 00:594–599
Zhang GX (1992) Cause-selecting control chart and diagnosis. Theory and practice. Aarhus School of Business. Department of Total Quality Management. Aarhus, Denmark
Zhao Y, Tsung F, Wang Z (2005) Dual CUSUM control schemes for detecting a range of mean shifts. IIE Trans 37:1047–1057
Zou C, Tsung F, Wang Z (2007a) Monitoring general linear profiles using multivariate EWMA schemes. Technometrics 49: 395–408
Zou C, Zhou C, Wang Z, Tsung F (2007b) A self-starting control chart for linear profiles. J Qual Technol 39:364–375
Zou C, Tsung F, Liu Y (2008a) A change point approach for phase I analysis in multistage processes. Technometrics 50(3): 344–356
Zou C, Tsung F, Wang Z (2008b) Monitoring profiles based on nonparametric regression methods. Technometrics 50: 512–526
Zou C, Wang Z, Tsung F (2008c) Monitoring an autocorrelated processes using variable sampling schemes at fixed-times. Qual Reliab Eng Int 24:55–69
Zuo Y, Serfling R (2000) General notions of statistical depth function. Annal Stat 28(2):461–482

Statistical Signal Processing

Debasis Kundu
Chair Professor
Indian Institute of Technology Kanpur, Kanpur, India

Signal processing may broadly be considered to involve the recovery of information from physical observations. The received signals is usually disturbed by thermal, electrical, atmospheric or intentional interferences. Due to the random nature of the signal, statistical techniques play an important role in signal processing. Statistics is used in the formulation of appropriate models to describe the behavior of the system, the development of appropriate techniques

for estimation of model parameters, and the assessment of model performances. Statistical Signal Processing basically refers to the analysis of random signals using appropriate statistical techniques. The main purpose of this article is to introduce different signal processing models and different statistical and computational issues involved in solving them.

The Multiple Sinusoids Model

The multiple sinusoids model may be expressed as

$$y(t) = \sum_{k=1}^{M} \{A_k \cos(\omega_k t) + B_k \sin \omega_k t)\} + n(t); \quad t = 1, \ldots, N. \tag{1}$$

Here A_k's and B_k's represent the amplitudes of the signal, ω_k's represent the real radian frequencies of the signals, $n(t)$'s are error random variables with mean zero and finite variance. The assumption of independence of the error random variables is not that critical to the development of the inferential procedures. The problem of interest is to estimate the unknown parameters $\{A_k, B_k, \omega_k\}$ for $k = 1, \ldots, M$, given a sample of size N. In practical applications often M is also unknown. Usually, when M is unknown, first estimate M using some model selection criterion, and then it is assumed that M is known, and estimate the amplitudes and frequencies.

The sum of sinusoidal model (1) plays the most important role in the Statistical Signal Processing literature. Most of the periodic signals can be well approximated by the model (1) with the proper choice of M and with the amplitudes and frequencies. For several applications of this model in different fields see Brillinger (1987).

The problem is an extremely challenging problem both from the theoretical and computational points of view. As a statistician Fisher (1929) first considered this problem. It seems that the standard least squares estimators will be the natural choice in this case, but finding the least squares estimators, and establishing their properties are far from trivial issues. Although, the model (1) is a non-linear regression model, but the standard sufficient conditions needed for the least squares estimators to be consistent and asymptotically normal do not hold true in this case. Special care is needed in establishing the consistency and ►asymptotic normality properties of the least squares estimators, see for example Hannan (1973) and Kundu (1997) in this respect. Moreover, for computing the least squares estimators, most of the standard techniques like Newton–Raphson or its variants do not often converge even from good starting values. Even if it converges, it may converge to a local minimum rather than the global minimum due to highly non-linear nature of the least squares surface. Special purpose algorithms have been developed to solve this problem.

Several approximate solutions have been suggested in the literature. Among several approximate estimators, Forward Backward Linear Prediction (FBLP) and modified EquiVariance Linear Prediction (EVLP) work very well. But it should be mentioned that none of these methods behaves uniformly better than the other. More than 200 references on this topic can be found in Stoica (1993), and see also Quinn and Hannan (2001), the only monograph written by statisticians in this topic.

Two-Dimensional Sinusoidal Model

Two dimensional periodic signals are often being analyzed by the two-dimensional sinusoidal model, which can be written as follows:

$$y(s,t) = \sum_{k=1}^{M} \{A_k \cos(\omega_k s + \mu_k t) + B_k \cos(\omega_k s + \mu_k t)\} + n(s,t), \quad s = 1, \ldots S, \; t = \ldots, T. \tag{2}$$

Here A_k's and B_k's are amplitudes and ω_k's and μ_k's are frequencies. The problem once again involves the estimation of the signal parameters namely A_k's, B_k's, ω_k's and μ_k's from the data $\{y(s,t)\}$.

The model (2) has been used very successfully for analyzing two dimensional gray texture data, see for example Zhang and Mandrekar (2001). A three dimensional version of it can be used for analyzing color texture data also, see Prasad (2009) and Prasad and Kundu (2009). Some of the estimation procedures available for the one-dimensional problem may be extended quite easily to two or three dimensions. However, several difficulties arise when dealing with high dimensional data. There are several open problems in multidimensional frequency estimation, and this continues to be an active area of research.

Array Model

The area of array processing has received a considerable attention in the past several decades. The signals recorded at the sensors contain information about the structure of the generating signals including the frequency and amplitude of the underlying sources. Consider an array of P sensors receiving signals from M sources ($P > M$). The array geometry is specified by the applications of interest. In array processing, the signals received at the $i-th$ censor is given by

$$y_i(t) = \sum_{j=1}^{M} a_i(\theta_j) x_j(t) + n_i(t), \quad i = 1, \ldots, P. \tag{3}$$

Here $x_j(t)$ represents the signal emitted by the $j-th$ source, and $n_i(t)$ represents additive noise. The model (3) may be written in the matrix form as;

$$y(t) = [a(\theta_1) : \ldots : a(\theta_M)]\, x(t) + n(t)$$
$$= A(\theta)x(t) + n(t), \quad t = 1, \ldots, N. \qquad (4)$$

The matrix $A(\theta)$ has a Vandermonde structure if the underlying array is assumed to be uniform linear array. The signal vector $x(t)$ and the noise vector $n(t)$ are assumed to be independent and zero mean random processes with covariance matrices Γ and $\sigma^2 I$ respectively. The main problem here is to estimate the signal vector θ, based on the sample $y(1), \ldots, y(N)$, when the structure of A is known.

Interestingly, instead of using the traditional maximum likelihood method, different subspace fitting methods, like MUltipe SIgnal Classification (MUSIC) and Estimation of Signal Parameters via Rotational Invariance Technique (ESPRIT) and their variants are being used more successfully, see for example the text by Pillai (1989) for detailed descriptions of the different methods.

For basic introduction of the subject the readers are referred to Kay (1987) and Srinath et al. (1996) and for advanced materials see Bose and Rao (1993) and Quinn and Hannan (2001).

Acknowledgments

Part of this work has been supported by a grant from the Department of Science and Technology, Government of India.

About the Author

Professor Debasis Kundu received his Ph.D in Statistics in 1989, Pennsylvania State University. He has (co-)authored about 165 papers. He is an Associate Editor of *Communications in Statistics, Theory and Methods* and an Associate Editor of *Communications in Statistics, Simulation and Computations*. He is an invited member of New York Academy of Sciences and a Fellow of the National Academy of Sciences, India.

Cross References

►Least Squares
►Median Filters and Extensions
►Nonlinear Models
►ROC Curves
►Singular Spectrum Analysis for Time Series
►Statistical Pattern Recognition Principles

References and Further Reading

Bose NK, Rao CR (1993) Signal processing and its applications. Handbook of Statistics, 10, North-Holland, Amsterdam

Brillinger D (1987) Fitting cosines: some procedures and some physical examples. In MacNeill IB, Umphrey GJ (eds) Applied statistics, stochastic processes and sampling theory. Reidel, Dordrecht

Fisher RA (1929) Tests of significance in Harmonic analysis. Proc R Soc London A 125:54–59

Hannan EJ (1973) The estimation of frequencies. J Appl Probab 10:510–519

Kay SM (1987) Modern spectral estimation. Prentice Hall, New York, NY

Kundu D (1997) Asymptotic theory of the least squares estimators of sinusoidal signal. Statistics 30:221–238

Pillai SU (1989) Array signal processing. Springer, New York, NY

Prasad A (2009) Some non-linear regression models and their applications in statistical signal processing. PhD thesis, Indian Institute of Technology Kanpur, India

Prasad A, Kundu D (2009) Modeling and estimation of symmetric color textures. Sankhya Ser B 71(1):30–54

Quinn BG, Hannan EJ (2001) The estimation and tracking of frequency. Cambridge University Press, Cambridge, UK

Srinath MD, Rajasekaran PK, Viswanathan R (1996) Introduction to statistical processing with applications. Prentice-Hall, Englewood Cliffs, NJ

Stoica P (1993) List of references on spectral estimation. Signal Process 31:329–340

Zhang H, Mandrekar V (2001) Estimation of hidden frequencies for 2-D stationary processes. J Time Ser Anal 22:613–629

Statistical Significance

Jan M. Hoem
Professor, Director Emeritus
Max Planck Institute for Demographic Research, Rostock, Germany

Statistical thinking pervades the empirical sciences. It is used to provide principles of initial description, concept formation, model development, observational design, theory development and theory testing, and much more. Some of these activities consist in computing significance tests for statistical hypotheses. Such a hypothesis typically is a statement about a regression coefficient in a linear regression or a relative risk for a chosen life-course event, such as marriage formation or death. The hypothesis can state that the regression coefficient equals zero (or that the relative risk equals 1), implying that the corresponding covariate has no impact on the transition in question and thus does not affect the behavior it represents, or that for all practical purposes the analyst may act as if this

were the case. Alternatively the hypothesis may predict the sign of the coefficient, for example that higher education leads to lower marriage rates, *ceteris paribus*, as argued by some economists. The converse (namely that the sign is zero or positive) would be called the *null hypothesis*. Other hypotheses concern the form of the statistical model for the behavior in question. In such a case the null hypothesis would be that the model specified is correct; this leads to questions of goodness of fit. In any case the statistician's task is to state whether the data at hand justify rejecting whatever null hypothesis has been formulated.

The null hypothesis is typically rejected when a suitable *test statistic* has a value that is unlikely when the null hypothesis is correct; usually the criterion is that the test statistic lies in (say) the upper tail of the probability distribution it has when the hypothesis is correct. An upper bound on the probability of rejecting the null hypothesis when it actually is correct is called *the level of significance* of the test method. It is an important task for the investigator to keep control of this upper bound. A test of significance is supposed to prevent that a conclusion is drawn (about a regression coefficient, say) when the data set is so small that a pattern "detected" can be caused by random variation. Operationally an investigator will often compute the probability (when the null hypothesis is correct) that in a new data set, say, the test statistic would exceed the value actually observed and reject the null hypothesis when this so-called *p-value* is very small, since a small *p*-value is equivalent to a large value of the test statistic.

Ideally, hypotheses should be developed on the basis of pre-existing theory and common sense as well as of empirical features known from the existing literature. Strict protocols should be followed that require any hypothesis experimentation to be made on one part of the current data set, with testing subsequently to be carried out on a virgin part of the same data, or on a new data set. Unfortunately, most empirical scientists in the economic, social, biological, and medical disciplines, say, find such a procedure too confining (assuming that they even know about it). It is common practice to use all available data to develop a model, formulate scientific hypotheses, and to compute test statistics or ►*p*-values from the same data, perhaps using canned computer programs that provide values of test statistics as if scientific statistical protocol could be ignored (Ziliak and McCloskey 2008). The danger of such practices is that the investigator loses control over any significance levels, a fact which has been of concern to professional statisticians for a good while (For some contributions from recent decades see Guttman (1985), Cox (1986), Schweder (1988), and Hurvich and Tsai (1990). Such concerns also extend to many others. For instance, Chow (1996) describes a litany of criticism appearing in the psychological literature in Chapter 1 of a book actually written to *defend* the null-hypothesis significance-test procedure. [See Hoem (2008) for a discussion of further problems connected to common practices of significance testing, namely the need to embed an investigation into a genuine theory of behavior rather than to rely on mechanical significance testing, the avoidance of grouped *p*-values (often using a system of asterisks), the selection of substantively interesting contrasts rather than those thrown up mechanically by standard software, and other issues]). For twenty years and more, remedies have been available to overcome the weaknesses of the procedures just described, including rigorous methods for model development and data snooping. Such methods prevent the usual loss of control over the significance level and also allow the user to handle model misspecification (The latter feature is important because a model invariably is an imperfect representation of reality.). Users of event-history analysis may want to consult Hjort (1988, 1992), Sverdrup (1990), and previous contributions from these authors and their predecessors.

Unfortunately such contributions seem to be little known outside a circle of professional statisticians, a fact which for example led Rothman (1998) to attempt to eradicate significance tests from his own journal (*Epidemiology*). He underlined the need to see the interpretation of a study based not on statistical significance, or lack of it, for one or more study variables, but rather on careful quantitative consideration of the data in light of competing explanations for the findings. For example, he would prefer a researcher to consider whether the magnitude of an estimated effect could be readily explained by uncontrolled confounding or selection biases, rather than simply to offer the uninspired interpretation that the estimated effect is significant, as if neither chance nor bias could then account for the findings.

About the Author

For biography *see* the entry ►Demography.

Cross References

►Event History Analysis
►Frequentist Hypothesis Testing: A Defense
►Misuse of Statistics
►Null-Hypothesis Significance Testing: Misconceptions
►Power Analysis
►Presentation of Statistical Testimony
►Psychology, Statistics in
►P-Values
►Significance Testing: An Overview

►Significance Tests, History and Logic of
►Significance Tests: A Critique
►Statistics: Nelder's view

References and Further Reading

Chow SL (1996) Statistical significance: rationale validity and utility. Sage Publications, London
Cox DR (1986) Some general aspects of the theory of statistics. J Am Stat Assoc 49:559–575
Guttman L (1985) The illogic of statistical inference for cumulative science. Appl Stoch Model Data Anal 1:3–10
Hjort NL (1988) On large-sample multiple comparison methods. Scand J Stat 15(4):259–271
Hjort NL (1992) On inference in parametric survival data models. Int Stat Rev 60(3):355–387
Hoem JM (2008) The reporting of statistical significance in scientific journals: A reflection. Demographic Res 18(15):437–442
Hurvich CM, Tsai C-L (1990) The impact of model selection on inference in linear regression. The American Statistician 44(3):214–217
Rothman KJ (1998) Special article: writing for epidemiology. Epidemiology 9(3):333–337
Schweder T (1988) A significance version of the basic Neyman–Pearson theory for scientific hypothesis testing. Scand J Stat 15(4):225–235 (with a discussion by Ragnar Norberg, pp 235–242)
Sverdrup E (1990) The delta multiple comparison method: performance and usefulness. Scand J Stat 17(2):115–134
Ziliak ST, McCloskey DN (2008) The cult of statistical significance; how the standard error costs us jobs, justice, and lives. The University of Michigan Press, Ann Arbor

Statistical Software: An Overview

Jan de Leeuw
Distinguished Professor and Chair
University of California-Los Angeles, Los Angeles, CA, USA

Introduction

It is generally acknowledged that the most important changes in statistics in the last 50 years are driven by technology. More specifically, by the development and universal availability of fast computers and of devices to collect and store ever-increasing amounts of data. Satellite remote sensing, large-scale sensor networks, continuous environmental monitoring, medical imaging, micro-arrays, the various genomes, and computerized surveys have not just created a need for new statistical techniques. These new forms of massive data collection also require efficient implementation of these new techniques in software. Thus development of statistical software has become more and more important in the last decades.

Large data sets also create new problems of their own. In the early days, in which the *t*-test reigned, including the data in a published article was easy, and reproducing the results of the analysis did not take much effort. In fact, it was usually enough to provide the values of a small number of sufficient statistics. This is clearly no longer the case. Large data sets require a great deal of manipulation before they are ready for analysis, and the more complicated data analysis techniques often use special-purpose software and some tuning. This makes *reproducibility* a very significant problem. There is no science without replication, and the weakest form of replication is that two scientists analyzing the same data should arrive at the same results.

It is not possible to give a complete overview of all available statistical software. There are older publications, such as Francis (1979), in which detailed feature matrices for the various packages and libraries are given. This does not seem to be a useful approach any more, there simply are too many programs and packages. In fact many statisticians develop ad-hoc software packages for their own projects.

We will give a short historical overview, mentioning the main general purpose packages, and emphasizing the present state of the art. Niche players and special purpose software will be largely ignored. There is a well-known quote from Brian Ripley (2002): "Let's not kid ourselves: the most widely used piece of software for statistics is Excel." This is surely true, but it is equally true that only a tiny minority of statisticians have a degree in statistics. We have to distinguish between "statistical software" and the much wider terrain of "software for statistics." Only the first type is of interest to us here – we will go on kidding ourselves.

BMDP, SAS, SPSS

The original statistical software packages were written for IBM mainframes. BMDP was the first. Its development started in 1957, at the UCLA Health Computing Facility. SPSS arrived second, developed by social scientists at the University of Chicago, starting around 1968. SAS was almost simultaneous with SPSS, developed since 1968 by computational statisticians at North Carolina State University. The three competitors differed mainly in the type of clients they were targeting. And of course health scientists, social scientists, and business clients all needed the standard repertoire of statistical techniques, but in addition some more specialized methods important in their field. Thus the packages diverged somewhat, although their basic components were very much the same.

Around 1985 all three packages added a version for personal computers, eventually developing WIMP (window, icon, menu, pointer) interfaces. Somewhat later they also added matrix languages, thus introducing at least some form of extensibility and code sharing.

As in other branches of industry, there has been some consolidation. In 1996 SPSS bought BMDP, and basically killed it, although BMDP-2009 is still sold in Europe by Statistical Solutions. It is now, however, no longer a serious contender. In 2009 SPSS itself was bought by IBM, where it now continues as PASW (Predictive Analytics Software). As the name change indicates, the emphasis in SPSS has shifted from social science data analysis to business analytics. The same development is going on at SAS, which was originally the Statistical Analysis System. Currently SAS is not an acronym any more. Its main products are SAS Analytics and SAS Business Intelligence, indicating that the main client base is now in the corporate and business community. Both SPSS (now PASW) and SAS continue to have their statistics modules, but the keywords have definitely shifted to analytics, forecasting, decision, and marketing.

Data Desk, JMP, Stata

The second generation of statistics packages started appearing in the 1980's, with the breakthrough of the personal computer. Both Data Desk (1985) and JMP (1989) were, from the start, written for Macintosh, i.e., for the WIMP interface. They had no mainframe heritage and baggage. As a consequence they had a much stronger emphasis on graphics, visualization, and exploratory data analysis.

Data Desk was developed by Paul Velleman, a former student of John Tukey. JMP was the brain child of John Sall, one of the co-founders and owners of SAS, although it existed and developed largely independent of the main SAS products. Both packages featured dynamic graphics, and used graphical widgets to portray and interactively manipulate data sets. There was much emphasis on brushing, zooming, and spinning. Both Data Desk and JMP have their users and admirers, but both packages never became dominant in either statistical research or statistical applications. They were important, precisely because they emphasized graphics and interaction, but they were still too rigid and too difficult to extend.

Stata, another second generation package for the personal computer, was an interesting hybrid of a different kind. It was developed since 1985, like BMDP starting in Los Angeles, near UCLA. Stata had a CLI (command line interface), and did not get a GUI until 2003. It emphasized, from the start, extensibility and user-contributed code. Stata did not get its own matrix language Mata until Stata-9, in 2007.

Much of Stata's popularity is due to its huge archive of contributed code, and a delivery mechanism that uses the Internet to allow for automatic downloads of updates and new submissions. Stata is very popular in the social sciences, where it attracts those users that need to develop and customize techniques, instead of using the more inflexible procedures of SPSS or SAS. For such users a CLI is often preferable to a GUI.

Until Stata developed its contributed code techniques, the main repository had been CMU's statlib, modeled on netlib, which was based on the older network interfaces provided by ftp and email. There were no clear organizing principles, and the code generally was FORTRAN or C, which had to be compiled to be useful. We will see that the graphics from Data Desk and JMP, and the command line and code delivery methods from Stata, were carried over into the next generation.

S, LISP-STAT, R

Work had on the next generation of statistical computing systems had already started before 1980, but it mostly took place in research labs. Bell Laboratories in Murray Hill, N.J., as was to be expected, was the main center for these developments.

At Bell John Chambers and his group started developing the S language in the late seventies. S can be thought of as a statistical version of MATLAB, as a language and an interpreter wrapped around compiled code for numerical analysis and probability. It went through various major upgrades and implementations in the eighties, moving from mainframes to VAX'es and then to PC's. S developed into a general purpose language, with a strong compiled library of linear algebra, probability and optimization, and with implementations of both classical and modern statistical procedures. The first 15 years of S history are ably reviewed by Becker (1994), and there is a 30 year history of the S language in Chambers (2008, Appendix A) . The statistical techniques that were implemented, for example in the *White Book* (Chambers and Hastie 1992), were considerably more up-to-date than techniques typically found in SPSS or SAS. Moreover the S system was build on a rich language, unlike Stata, which until recently just had a fairly large number of isolated data manipulation and analysis commands. Statlib started a valuable code exchange of public domain S programs.

For a long time S was freely available to academic institutions, but it remained a product used only in the higher reaches of academia. AT&T, later Lucent, sold S to

the Insightful corporation, which marketed the product as S-plus, initially quite successfully. Books such as Venables and Ripley; Venables and Ripley (1994; 2000) effectively promoted its use in both applied and theoretical statistics. Its popularity was increasing rapidly, even before the advent of R in the late nineties. S-plus has been quite completely overtaken by R. Insightful was recently acquired by TIBCO, and S-plus is now TIBCO Spotfire S+. We need not longer consider it as a serious contender.

There were two truly exciting developments in the early nineties. Luke Tierney (1990) developed LISP-STAT, a statistics environment embedded in a Lisp interpreter. It provided a good alternative to S, because it was more readily available, more friendly to personal computers, and completely open source. It could, like S, easily be extended with code written in either Lisp or C. This made it suitable as a research tool, because statisticians could rapidly prototype their new techniques, and distribute them along with their articles. LISP-STAT, like Data Desk and JMP, also had interesting dynamic graphics capabilities, but now the graphics could be programmed and extended quite easily. Around 2000 active development of LISP-STAT stopped, and R became available as an alternative (Valero-Mora and Udina 2004).

R was written as an alternative implementation of the S language, using some ideas from the world of Lisp and Scheme (Ihaka and Gentleman 1996). The short history of R is a quite unbelievable success story. It has rapidly taken over the academic world of statistical computation and computational statistics, and to an ever-increasing extend the world of statistics teaching, publishing, and real-world application. SAS and SPSS, which initially tended to ignore and in some cases belittle R, have been forced to include interfaces to R, or even complete R interpreters, in their main products. SPSS has a Python extension, which can run R since SPSS-16. The SAS matrix language SAS/IML, starting at version 3.2. has an interface to an R interpreter.

R is many things to many people: a rapid prototyping environment for statistical techniques, a vehicle for computational statistics, an environment for routine statistical analysis, and a basis for teaching statistics at all levels. Or, going back to the origins of S, a convenient interpreter to wrap existing compiled code. R, like S, was never designed for this all-encompassing role, and the basic engine is straining to support the rate of change in the size and nature of data, and the developments in hardware.

The success of R is both dynamic and liberating. But it remains an open source project, and nobody is really in charge. One can continue to tag on packages extending the basic functionality of R to incorporate XML, multicore processing, cluster and grid computing, web scraping, and so on. But the resulting system is in danger of bursting at the seams. There are now four ways to do (or pretend to do) object-oriented programming, four different systems to do graphics, and four different ways to link in compiled C code. There are thousands of add-on packages, with enormous redundancies, and often with code that is not very good and documentation that is poor. Many statisticians, and many future statisticians, learn R as their first programming language, instead of learning real programming languages such as Python, Lisp, or even C and FORTRAN. It seems realistic to worry at least somewhat about the future, and to anticipate the possibility that all of those thousands of flowers that are now blooming may wilt rather quickly.

Open Source and Reproducibility

One of the consequences of the computer and Internet revolution is that more and more scientists promote open source software and reproducible research. Science should be, per definition, both open and reproducible. In the context of statistics (Gentleman and Temple-Lang 2004) this means that the published article or report is not the complete scientific result. In order for the results to be reproducible, we should also have access to the data and to a copy of the computational environment in which the calculations were made.

Publishing is becoming more open, with e-journals, preprint servers, and open access. Electronic publishing makes both open source and reproducibility more easy to realize. The Journal of Statistical Software, at http://www.jstatsoft.org, the only journal that publishes and reviews statistical software, insists on complete code and completely reproducible examples. Literate Programming systems such as Sweave, at http://www.stat.uni-muenchen.de/~leisch/Sweave/, are becoming more popular ways to integrate text and computations in statistical publications.

We started this overview of statistical software by indicating that the computer revolution has driven much of the recent development of statistics, by increasing the size and availability of data. Replacement of mainframes by minis, and eventually by powerful personal computers, has determined the directions in the development of statistical software. In more recent times the Internet revolution has accelerated these trends, and is changing the way scientific knowledge, of which statistical software is just one example, is disseminated.

About the Author

Dr. Jan de Leeuw is Distinguished Professor and Chair, Department of Statistics, UCLA. He has a 1973 Ph.D. in

Social Sciences from the University of Leiden, Netherlands. He came to UCLA in 1987, after leading the Department of Data Theory at the University of Leiden for about 10 years. He is Elected Fellow, Royal Statistical Society (1984), Elected Member, International Statistical Institute (1986), Corresponding Member, Royal Netherlands Academy of Sciences (1987), Elected Fellow, Institute of Mathematical Statistics (2001) and American Statistical Association (2001). Dr. de Leeuw is Editor-in-Chief, and Founding Editor of *Journal of Statistical Software*, and Editor-in-Chief, *Journal of Multivariate Analysis* (1997–). He is a Former President of the *Psychometric Society* (1987). Professor de Leeuw has (co-)authored over 550 papers, book chapters and reviews, including *Introducing Multilevel Modeling* (with Ita Kreft, Sage, 1998), and *Handbook of Multilevel Analysis* (edited with Erik Meijer, Springer, New York, 2007).

Cross References

- ►Analysis of Variance
- ►Behrens–Fisher Problem
- ►Chi-Square Tests
- ►Computational Statistics
- ►Multiple Imputation
- ►R Language
- ►Selection of Appropriate Statistical Methods in Developing Countries
- ►Spreadsheets in Statistics
- ►Statistical Analysis of Longitudinal and Correlated Data
- ►Statistical Consulting
- ►Statistical Fallacies: Misconceptions, and Myths
- ►Statistics and Climate Change

References and Further Reading

Becker RA (1994) A brief history of S. Technical report, AT&T Bell Laboratories, Murray Hill, N.J. URL http://www2.research.att.com/areas/stat/doc/94.11.ps

Chambers JM (2008) Software for data analysis: programming with R. Statistics and computing. Springer, New York, NY

Chambers JM, Hastie TJ (eds) (1992) Statistical models in S. Wadsworth, California

Francis I (1979) A comparitive review of statistical software. International Association for Statistical Computing, Voorburg, The Netherlands

Gentleman R, Temple-Lang D (2004) Statistical analyses and reproducible research. Bioconductor Project Working Papers 2. URL http://www.bepress.com/cgi/viewcontent.cgi?article=1001&context=bioconductor

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Gr Stat 5:299–314

Ripley BD (2002) Statistical methods *need* software: a view of statistical computing. Presentation RSS Meeting, September. URL http://www.stats.ox.ac.uk/~ripley/RSS2002.pdf

Tierney L (1990) LISP-STAT. An object-oriented environment for statistical computing and dynamic graphics. Wiley, New York

Valero-Mora PM, Udina F (2004) Special issue: lisp-stat: Past, present and future. J Stat Software 13, URL http://www.jstatsoft.org/v13

Venables WN, Ripley BD (1994) Modern applied statistics with S, 1st edn. Springer, New York

Venables WN, Ripley BD (2000) S Programming. Statistics and Computing. Springer, New York, NY

Statistical View of Information Theory

Adnan M. Awad
Professor
University of Jordan, Amman, Jordan

Information Theory has origins and applications in several fields such as: thermodynamics, communication theory, computer science, economics, biology, mathematics, probability and statistics. Due to this diversity, there are numerous information measures in the literature. Kullback (1978), Sakamoto et al. (1986), and Pardo (2006) have applied several of these measures to almost all statistical inference problems.

According to The Likelihood Principle, all experimental information relevant to a parameter θ is mainly contained in the likelihood function $L(\theta)$ of the underlying distribution. Bartlett's information measure is given by $-\log(L(\theta))$. Entropy measures (see ►Entropy) are expectations of functions of the likelihood. Divergence measures are also expectations of functions of likelihood ratios. In addition, Fisher-like information measures are expectations of functions of derivatives of the log-likelihood. DasGupta (2008, Chap. 2) reported several relations among members of these information measures. In sequential analysis, Wald (1947, p. 53) showed earlier that the average sample number depends on a divergence measure of the form

$$E_\theta\left[\log\frac{f(X,\theta_1)}{f(X,\theta_0)}\right]$$

where θ_0 and θ_1 are the assumed values of the parameter θ of the density function f of the random variable X under the null and the alternative hypothesis, respectively.

It is worth noting that, and from the point of view of decision making, the expected change in utility can be

used as a quantitative measure of the worth of an experiment. In this regard Bayes' rule can be viewed as a mechanism that processes information contained in data to update the prior distribution into the posterior probability distribution.

Furthermore, according to Jaynes' Principle of Maximum Entropy (1957), information in a probabilistic model is the available moment constraints on this model. This principle is in fact a generalization of Laplace's Principle of Insufficient Reason.

From a statistical point of view, one should concentrate on the statistical interpretation of properties of entropy-information measures with regard to the extent of their agreement with statistical theorems and to their degree of success in statistical applications.

The following provides a discussion of preceding issues with particular concentration on Shannon's entropy. For more details, the reader can consult the list of references.

1. Consider a discrete random variable X taking a finite number of values $\overrightarrow{X} = (x_1, \ldots, x_n)$ with probability vector $P = (p_1, \ldots, p_n)$. Shannon's entropy (information) of P or of X (1948) is given by

$$H(X) = H(P) = -\sum_{i=1}^{n} p_i \log(p_i).$$

   The most common bases of the logarithm are 2 and e. With base 2, H is measured in bits whereas, in base e, the units of H are nats. In coding theory the base is 2 whereas, in statistics the base is e.
2. It is quite clear that $H(P)$ is symmetric in the components of the vector P. This implies that components of P can be rearranged to get different density functions which are either: symmetric, negatively skewed, positively skewed, unimodal or bimodal. Such distributions carry different information even though they all have same value of $H(P)$. Therefore, $H(P)$ is unable to reflect the information implied by the shape of the underlying distribution.
3. ▶Entropy of a discrete distribution is always positive while the differential entropy $H(f) = -\int_{-\infty}^{\infty} f(x) \log(f(x))\, dx$ of a continuous variable X with pdf f may take any value on the extended real line. This is due to the fact that the density $f(x)$ need not be less than one as in the discrete case. Thus, Shannon's entropy lacks the ability to give a proper assessment of information when the random variable is continuous. To overcome this problem, Awad (1987) introduced sup-entropy as $-E[\log(f(X)/s)]$, where s is the supremum of $f(x)$.
4. Based on a random sample $O_n = (X_1, \ldots, X_n)$ of size n from a distribution and according to Fisher (1925), a sufficient statistic T carries all information in the sample while any other statistic carries less information than T. The question that arises here is that: "Does Shannon's entropy agree with Fisher's definition of a sufficient statistic?". Let us consider the following two examples.

   First, let $Y : N(\theta, \sigma^2)$ denote a normal random variable with mean θ and variance σ^2. It can be shown that $H(Y) = \log(2\pi e\sigma^2)/2$ which is free of θ. Let O_n be a random sample of size n from $X : N(\theta, 1)$ then by the additivity property of Shannon's entropy, $H(O_n) = nH(X) = n\log(2\pi e)/2$. On the other hand, Shannon's entropy of the sufficient statistic $\overline{X}_n$ is $H(\overline{X}_n) = \log(2\pi e/n)/2 = H(X) - \log(n)/2$. Since $H(X)$ is positive, $H(O_n) \geq H(\overline{X}_n)$ with equality if $n = 1$, i.e., Shannon's entropy of sufficient statistic is less than that of the sample.

   Second, consider a random sample O_n of size n from a continuous uniform distribution on the interval $[0, \theta]$. Let $X_{1:n}$ and $X_{n:n}$ denote the minimum and the maximum ▶order statistics in O_n. It can be shown that $H(X_{1:n}) = H(X_{n:n})$, i.e., Shannon's entropy of sufficient statistic $X_{n:n}$ equals Shannon's entropy of a non-sufficient statistic $X_{1:n}$. These examples illustrate that Shannon's entropy does not agree with Fisher's definition of a sufficient statistic.
5. If $Y = \alpha + \beta X$, $\beta \neq 0$, then $H(Y) = H(X)$ when X is a discrete random variable. However, if X is continuous, $H(Y) = H(X) + \log(|\beta|)$. So, this result implies that two sufficient statistics T_1 and $T_2 = \beta T_1$ will carry (according to Shannon's entropy) unequal amounts of information, which contradicts the sufficiency concept.
6. Referring to the first example in (4), it is clear that Shannon's information in the sample mean is a decreasing function of the sample size n. This is in direct conflict with the usual contention that the larger the sample size is the more information one has. It is also interesting to recall in this regard Basu's example (1975), where a sample of size 2 is more informative (about an unknown parameter) than a sample of size 25. In fact, a rewording of Basu's conclusion is that some observations in the sample are more influential than others.

Acknowledgment

The author wish to thank Prof. Tarald O. Kvalseth and Prof. Miodrag Lovric for their careful reading and valuable suggestions that improved the presentation of the article.

About the Author

Adnan Awad graduated from Yale University, USA, with a PhD. in Statistics, 1978. He chaired the Department of Statistics, Yarmouk University, (1982–1985). He was past chair of the Mathematics Department, University of Jordan. He served as past Vice Dean of Faculty of Graduate studies, Jordan University (1998), and past Vice Dean of the Faculty of Research, Al-albayet University, Jordan (1999). He has authored and co-authored about 80 research papers and more than 20 text books. He supervised seven Ph.D and 28 M.Sc. theses. Moreover, he was a member of the UNESCO team, (1992–1996), of improving teaching mathematics in the Arab World. Professor Awad has been awarded the medal of High Research Evaluation at the Faculty of Science, Yarmouk University (1984), and Abdul-Hammed Shooman Prize for Young Arab Scientists in Mathematics and Computer Science, Jordan (1987), for his contributions to both Prediction Analysis and Information Theory.

Cross References

►Diversity
►Entropy
►Entropy and Cross Entropy as Diversity and Distance Measures
►Information Theory and Statistics
►Measurement of Uncertainty
►Sufficient Statistical Information
►Sufficient Statistics

References and Further Reading

Awad AM (1987) A statistical information measure. Dirasat (Science) 14(12):7–20
Bartlett MS (1936) Statistical information and properties of sufficiency. Proc R Soc London A 154:124–137
Basu D (1975) Statistical information and likelihood. Sankhya A 37(1):1–71
DasGupta A (2008) Asymptotic theory of statistics and probability. Springer Science Media, LLC
Fisher RA (1925) Theory of statistical estimation. Proc Cambridge Philos Soc 22:700–725
Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:620–630; 180:171–197
Kullback S (1978) Information theory and statistics. Gloucester, Peter Smith, MA
Lindley DV (1956) On the measure of information provided by an experiment. Ann Stat 27:986–1005
Pardo L (2006) Statistical inference based on divergence measures. Chapman and Hall, New York
Sakamoto Y, Ishiguro M, Kitagawa G (1986) Akaike information criterion statistics. KTK
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423 and 623–656
Wald A (1947) Sequential analysis. Dover, New York

Statistics and Climate Change

Implication of Statisticians and Statistics Education

Parinbanu Kurji
Head of Biometry, Faculty of Agriculture
University of Nairobi, Nairobi, Kenya

What Does Climate Change Hold for the Future?

There is general agreement among experts that we can expect a rise in temperatures and an increase in the number of extreme events, but for other climate variables such as rainfall there is no clear prediction. However there does not seem to be any doubt that communities coping with poverty will be particularly vulnerable – this means developing countries like Africa will be the hardest hit (Cooper et al. 2006; Washington et al. 2006; Climate Proofing Africa, DFID 2005; Burton and van Aaist 2004). The climate change dialogue brings with it an enormous need for more and better climate data and greater rigor in its analysis. To understand both risks and opportunities associated with the season-to-season variability that is characteristic of current climates as well as changes in the nature of that variability due to climate change, there is need for all stakeholders, including the statistical community, policy makers, and scientists, to work together to propose appropriate strategies to counteract one and enhance the other. Such strategies must be based on scientific studies of climate risk and trend analyses and not fashionable perceptions or anecdotal evidence. Statisticians have a vital role to play here.

What Is Needed?

One of the ways of approaching this issue of climate change as it affects the people in the developing countries is through a better understanding of the season-to-season variability in weather that is a defined characteristic of current climate (*Climate: The statistical description in terms of means and variability of key weather parameters for a given area over a period of time – usually at least 30 years*) and using this to address future change. Managing current climate-induced risk is already an issue for farmers who practice rain-fed agriculture. Helping them to cope better with this risk while preparing for future change seems to be the best way of supporting the needy both for the present and for the future. Agriculture is one field where

the vagaries of climate have an impact but other fields such as health, construction, and transport among others would benefit equally from this approach.

Why Do Statisticians Need to Be Involved?

Meteorology departments are the custodians of climate data and, especially in many developing countries, data quality and management, rather than analysis, have been priority issues and the institutions have limited themselves mainly to providing data to users. There is now a move to shift from providing basic data and services to meeting increasingly challenging user needs.

Effective use of climatic summaries and especially applications require an understanding of statistical concepts underlying these summaries as well as proficiency in using and interpreting the advanced statistical techniques and models that are being suggested to understand climate change.

Statistics is the glue that brings the different disciplines together and statisticians need to form an integral part of multidisciplinary teams to understand, extend, and share knowledge of existing and upcoming technologies and statistical methods for development purposes.

Where Should Changes Occur?

The three areas where statisticians can be proactive in addressing the climate change issue are:

1. Working actively with researchers in various disciplines in guiding research to develop and test adaptation strategies.

   For example, if, as is expected, temperatures are going to rise, and this affects crop growth, it is now that research agendas must be set if we are to meet the new challenges. There needs to be a clear understanding about the implications of such conditions.
2. Being aggressively involved in building capacities of data producers and data users. At present the capacity in many developing countries for modeling and interpreting data is highly inadequate

   For example, creating awareness of the need for quantity, quality, and timeliness of climate data required for use in modeling climate processes and for using and extending these models in collaboration with agriculture scientists and extension workers.
3. Promoting changes in statistics training at all levels to meet the expanding needs.

   For example, innovative statistics curriculum at universities & colleges that mainstream climate data analysis and that emphasize understanding and application of concepts using a data-based approach.

Some Available Resources

Given the availability and affordability of computers today, they should now form an integral part of good statistics training. Among the many resources available to enhance statistics training in general, and training in climatic statistics in particular are:

- *CAST for Africa* (www.cast.massey.ac.nz), an electronic statistics textbook that provides an interesting interactive way of understanding statistical concepts with a number of real-life data sets from different disciplines. Climate CAST, which is an offshoot of this, provides the slant for exploring climatic data. The textbook goes from the very basic to reasonably complex topics.
- *Instat* (www.reading.ac.uk/ssc), a simple software package with a special climate menu and a number of useful guides in the help section to facilitate training as well as self study.
- *GenStat* (www.vsni.co.uk), a major statistical package, is an all-embracing data analysis tool, offering ease of use through comprehensive menu system reinforced with the flexibility of a sophisticated programming language. It has many useful facilities including analysis of extremes. The discovery version is provided free for nonprofit organizations while the latest version is available at very reasonable rates to training and research institutions. Here again there is wealth of information for the user in terms of guides, including a guide to climatic analyses, and tutorials and examples from diverse fields.
- *DSSAT* (www.icasa.net/dssat) and *ApSim* (www.apsim.info/apsim), crop simulation models, driven by long-term daily climatic data, which can be used to simulate realistic long-term field experiments. These are probably more useful at postgraduate or faculty levels but have great potential for statisticians working with agriculture scientists to explore possible scenarios without actually undertaking long costly field experiments.

Some Working Initiatives

- *Statistics Curriculum, at Faculty of Agriculture, University of Nairobi, Kenya*

  An innovative data-based problem-solving approach to service teaching for the Agriculture Faculty uses building blocks approach – from descriptive to modeling to application – to broaden and deepen

the students' understanding of how statistics is used in practice. The curriculum includes computer proficiency and soft skills as an integral part of the curriculum and exposes students to all types of data, including climatic data, which is not only important in its own right but also an important example of monitoring data. Examples of how climatic analyses have been incorporated into the service teaching of statistics are given by Kurji and Stern (2005).

- *Masters in Climate Data Analysis, at Science Faculty, Maseno University, Kenya*

  Currently there are a number of students who are working on their postgraduate degree with specific climate-related projects, both advancing the science, encouraging statisticians to embrace the new challenges of development, and building capacity in the field of climate analysis.
- *Statistics for Applied Climatology (SIAC) at IMTR (Institute of Meteorological Training & Research), Kenya*

  This is a regional program run by the Institute for groups comprising officers from National Met services and Agriculture Research Scientists to develop statistical skills and build networks for further collaborative work. The course has two components, a 6-week e-learning course followed by a 4-week face-to-face course, which culminates in a project that can be continued after the participants return to their bases.

Cross References

►Agriculture, Statistics in
►Mathematical and Statistical Modeling of Global Warming
►Role of Statistics
►Statistical Aspects of Hurricane Modeling and Forecasting
►Statistics Education

References and Further Reading

Burton I, van Aaist M (2004) Look before you leap: a risk management approach for incorporating climate change adaptation into World Bank Operations. World Bank Monograph, Washington (DC), DEV/GEN/37 E. 10

Cooper PJM, Dimes J, Rao KPC, Shapiro B, Shiferaw B, Twomlow S (2006) Coping better with current climatic variability in the rain-fed farming systems of sub-Saharan Africa: a dress rehearsal for adapting to future climate change? Global theme on agro-ecosystems Report no. 27. International Crops Research Institute for the Semi-Arid Tropics, PO Box 29063-00623, Nairobi, Kenya, 24pp

DFID (2005) Climate proofing Africa: climate and Africa's development challenge. Department for International Development, London

Kurji P, Stern RD (2005) Teaching statistics using climatic data. http://www.ssc.rdg.ac.uk/bucs/MannafromHeaven.pdf

Washington R, Harrison M, Conway D, Black E, Challinor A, Grimes D, Jones R, Morse A, Kay G, Todd M (2006) African climate change: taking the shorter route. Bull Am Meteorol Soc 87: 1355–1366

Statistics and Gambling

Kyle Siegrist
Professor
University of Alabama in Huntsville, Huntsville, AL, USA

Introduction

Statistics can broadly be defined as the science of decision-making in the face of (random) uncertainty. Gambling has the same definition, except in the narrower domain of a gambler making decisions that affect his fortune in games of chance. It is hardly surprising, then, that the two subjects are closely related. Indeed, if the definitions of "game," "decision," and "fortune" in the context of gambling are sufficiently broadened, the two subjects become almost indistinguishable.

Let's review a bit of the history of the influence of gambling on the development of probability and statistics. First, of course, gambling is one of the oldest of human activities. The use of a certain type of animal heel bone (called the *astragalus*) as a crude die dates to about 3500 BCE (and possibly much earlier). The modern six-sided die dates to about 2000 BCE.

The early development of probability as a mathematical theory is intimately related to gambling. Indeed, the first probability problems to be analyzed mathematically were gambling problems:

1. *De Mere's problem* (1654), named for Chevalier De Mere and analyzed by Blaise Pascal and Pierre de Fermat, asks whether it is more likely to get at least one six with 4 throws of a fair die or at least one double six in 24 throws of two fair dice.
2. The *problem of points* (1654), also posed by De Mere and analyzed by Pascal and Fermat, asks for the fair division of stakes when a sequence of games between two players (Bernoulli trials in modern parlance) is interrupted before its conclusion.
3. *Pepys' Problem* (1693), named for Samuel Pepys and analyzed by Isaac Newton, asks whether it is more likely to get at least one six in six rolls of a fair die or at least two sixes in 12 rolls of the die.

4. *The matching problem* (1708), analyzed by Pierre-Redmond de Montmort, is to find the probability that in a sequence of card draws, the value of a card is the same as the draw number.
5. *St. Petersburg Paradox* (1713), analyzed by Daniel Bernoulli, deals with a gambler betting on a sequence of coin tosses who doubles his bet each time he loses (and leads to a random variable with infinite expected value).

Similarly, the first books on probability were written by mathematician-gamblers to analyze games of chance: *Liber de Ludo Aleae* written sometime in the 1500s by the colorful Girolamo Cardano and published posthumously in 1663, and *Essay d'Analyse sur les Jeux de Hazard* by Montmort, published in 1708. See David 1998 and Epstein 1977 for more on the influence of gambling on the early development of probability and statistics.

In more modern times, the interplay between statistics and game theory has been enormously fruitful. Hypothesis testing, developed by Ronald Fisher and Karl Pearson and formalized by Jerzy Neyman and Egon Pearson is one of the cornerstones of modern statistics, and has a game-theory flavor. The basic problem is choosing between a presumed null hypothesis and a conjectured alternative hypothesis, with the decision based on the data at hand and the probability of a type 1 error (rejecting the null hypothesis when it's true). Influenced by the seminal work of John von Neumann and Oscar Morgenstern on game theory and economics (von Neumann 1944), the Neyman-Pearson hypothesis-testing framework was extended by Abraham Wald in the 1940s to *statistical decision theory* (Wald 1950). In this completely game-theoretic framework, the statistician (much like the gambler) chooses among a set of possible decisions, based on the data at hand according to some sort of value function. Statistical decision theory remains one of the fundamental paradigms of statistical inference to this day.

Bold Play in Red and Black

Gambling continue to be a source of interesting and deep problems in probability and statistics. In this section, we briefly describe a particularly beautiful problem analyzed by Dubins and Savage (1976). A gambler bets, at even stakes, on a sequence of Bernoulli trials (independent, identically distributed trials) with success parameter $p \in (0,1)$. The gambler starts with an initial fortune and must continue playing until he is ruined or reaches a fixed target fortune. (The last two sentences form the mathematical definition of *red and black*.) On each trial, the gambler can bet any proportion of his current fortune, so it's convenient to normalize the target fortune to 1; thus the space of fortunes is the interval $[0,1]$.

The gambler's goal is to maximize the probability $F(x)$ of reaching the target fortune 1, starting with an initial fortune x (thus, F is the value function in the context of statistical decision theory). The gambler's strategy consists of decisions on how much to bet on each trial. Since the trials are independent, the only information of use to the gambler on a given trial is his current fortune. Thus, we need only consider *stationary, deterministic strategies*. Such a strategy is defined by a *betting function* $S(x)$ that gives the amount bet on a trial as a function of the current fortune x.

Dubins and Savage showed that in the sub-fair case $\left(p \le \frac{1}{2}\right)$, an optimal strategy is *bold play*, whereby the gambler, on each trial, bets his entire fortune or the amount needed to reach the target (whichever is smaller). That is, the betting function for bold play is

$$S(x) = \begin{cases} x, & 0 \le x \le \frac{1}{2} \\ 1 - x, & \frac{1}{2} \le x \le 1 \end{cases}$$

Conditioning on the first trial shows that the value function F for bold play satisfies the functional equation

$$F(x) = \begin{cases} pF(2x), & x \in \left[0, \frac{1}{2}\right] \\ p + (1-p)F(2x-1), & x \in \left[\frac{1}{2}, 1\right] \end{cases} \tag{1}$$

with boundary conditions $F(0) = 0$, $F(1) = 1$. Moreover, F is the unique bounded solution of (1) satisfying the boundary conditions. This functional equation is one of the keys in the analysis of bold play. In particular, the proof of optimality involves showing that if the gambler starts with some other strategy on the first trial, and then plays boldly thereafter, the new value function is no better than the value function with bold play.

Interestingly, as Dubins and Savage also showed, bold play is not the unique optimal strategy. Consider the following strategy: Starting with fortune $x \in \left[0, \frac{1}{2}\right)$, the gambler plays boldly, but with the goal of reaching $\frac{1}{2}$. Starting with fortune $x \in \left(\frac{1}{2}, 1\right]$, the gambler plays boldly, but with the goal of not falling below $\frac{1}{2}$. In either case, if the gambler's fortune reaches $\frac{1}{2}$, he plays boldly and bets $\frac{1}{2}$. Thus, the betting function S_2 for this new strategy is related to the betting function S of bold play by

$$S_2(x) = \begin{cases} \frac{1}{2}S(2x), & 0 \le x < \frac{1}{2} \\ \frac{1}{2}S(2x-1), & \frac{1}{2} < x \le 1 \\ \frac{1}{2}, & x = \frac{1}{2} \end{cases}$$

By taking the three cases $x \in \left[0, \frac{1}{2}\right)$, $x = \frac{1}{2}$, and $x \in \left(\frac{1}{2}, 1\right]$, it's easy to see that that the value function F_2 for strategy S_2 satisfies the functional equation (1). Trivially the boundary conditions are also satisfied, so by uniqueness, $F_2 = F$ and thus S_2 is also optimal.

Once one sees that this new strategy is also optimal, it's easy to construct an entire sequence of optimal strategies. Specifically, let $S_1 = S$ denote the betting function for ordinary bold play and then define S_n recursively by

$$S_{n+1}(x) = \begin{cases} \frac{1}{2} S_n(2x), & 0 \le x < \frac{1}{2} \\ \frac{1}{2} S_n(2x-1), & \frac{1}{2} < x \le 1 \\ \frac{1}{2}, & x = \frac{1}{2} \end{cases}$$

Then S_n has the same value function F as bold play and so is optimal for each n. Moreover, if $x \in (0,1)$ is not a binary rational (that is, does not have the form $\frac{k}{2^n}$ for some k and n), then there exist optimal strategies that place arbitrarily small bets when the fortune is x. This is a surprising result that seems to run counter to a naive interpretation of the law of large numbers.

Bold play in red and black leads to some exotic functions of the type that are not usually associated with a simple, applied problem. The value function F can be interpreted as the distribution function of a random variable X (the variable whose binary digits are the complements of the trial outcomes). Thus F is continuous, but has derivative 0 almost everywhere if $p \neq \frac{1}{2}$ (singular continuous). If $p = \frac{1}{2}$, X is uniformly distributed on $[0,1]$ and $F(x) = x$. If $G(x)$ denotes the expected number of trials under bold play, starting with fortune x, then G is discontinuous at the binary rationals and continuous at the binary irrationals.

Finally, note that when the gambler plays boldly, his fortune process follows the deterministic map $x \mapsto 2x \bmod 1$, until the trial that ends the game (with fortune 0 or 1). Thus, bold play is intimately connected with a discrete dynamical system. This connection leads to other interesting avenues of research (see Pendergrass and Siegrist 2001).

About the Author

Kyle Siegrist is Professor of Mathematics at the University of Alabama in Huntsville, USA. He was Chair of the Department of Mathematical Sciences from 2001 to 2005 and was Editor of the *Journal of Online Mathematics at Its Applications* from 2005 to 2009. He is the author of over 30 journal articles and one book. He is the principle developer of Virtual Laboratories in Probability and Statistics, a web project that has twice received support from the US National Science Foundation.

Cross References

►Actuarial Methods
►Components of Statistics
►Decision Theory: An Introduction
►Decision Theory: An Overview
►Martingales
►Monty Hall Problem: Solution
►Probability Theory: An Outline
►Probability, History of
►Significance Testing: An Overview
►St. Petersburg Paradox
►Uniform Random Number Generators

References and Further Reading

Blackwell D, Girshick MA (1979) Theory of games and statistical decisions. Dover, New York

David FN (1998) Games, gods and gambling, a history of probability and statistical ideas. Dover, New York

Dubins LE, Savage LJ (1976) Inequalities for stochastic processes (how to gamble if you must). Dover, New York

Epstein RA (1977) The theory of gambling and statistical logic. Academic, New York

Pendergrass M, Siegrist K (2001) Generalizations of bold play in red and black. Stoch Proc Appl 92

Savage LJ (1972) The foundations of statistics. Dover, New York

von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton

Wald A (1950) Statistical decision functions. Wiley, New York

Statistics and the Law

Mary W. Gray
Professor
American University, Washington DC, USA

The role of the statistician in litigation has much in common with that of a consultant in any field. To be an effective expert witness, we should be certain that we know what questions must be answered and what data will be required in order to answer them. Other guidelines include

- Promoting and preserving the confidence of the client and the public without exaggerating the accuracy or explanatory power of the data
- Avoiding unrealistic expectations and not promising more than you can deliver
- Being responsible and accountable, guarding your reputation

- Providing adequate information to permit methods, procedures, techniques, and findings to be assessed
- Addressing rather than minimizing uncertainty

However, the statistician must understand that litigation is an adversarial process; one must consider the strategy of the other side and be prepared for what is likely to be presented. The keys to effective statistical evidence are

- Early involvement by the statistician (as is the case in any situation)
- Adequate data
- Clarity of presentation
- Effective supplemental anecdotal evidence (not the task of the statistician, but an important complement to it)
- Understanding of the statistics by the litigator
- Recognizing that the statistician cannot reach legal conclusions nor can s/he be an advocate (for anything other than statistics!)

In the United States statistical evidence has been used in cases involving

- Race, sex, and age discrimination in employment and education
- Evidence-based medicine
- Environmental effects of business practices
- DNA, ear prints, bullet composition
- Death penalty
- Product liability
- Intellectual property and many other issues
- On the international scene, statistical evidence was used in the war crimes trial of Milosevic and in other human rights cases.

The techniques used span the range of statistical methodology from descriptive statistics to *t*-tests to regression (nearly ubiquitous), non-parametric tests, capture-recapture, urn models, change point analysis, multiple systems analysis, Mantel-Hanszel tests to Bayesian techniques (not generally popular with the courts) and a variety of other sophisticated methods. Courts have a great deal of difficulty with the concept of sampling, especially when the sample is very small in comparison with a population. They also often have difficulty in seeing the applicability of statistics to an individual case. For example, evidence that, all else being equal, the death penalty was far more likely to be imposed when the victim was white than when the victim was black, has not kept individuals whose victims were white from being sentenced to death.

An important observation to keep in mind is that an expert with a newly-developed technique may not fare well in court. The usual standard for admission of statistical or other scientific evidence is that

1. The testimony is based upon sufficient facts or data,
2. The testimony is the product of reliable principles and methods, and
3. The witness has applied the principles and methods reliably to the facts of the case

Peer-reviewed publication usually meets the second requirement.

The classic example of the ▶misuse of statistics is in *People v. Collins* (1968), where the following analysis sent Malcolm Collins to prison. Witnesses reported various characteristics, characteristics that Malcolm and Janet Collins had, and the prosecutor got the expert to agree to certain hypothetical probabilities as follows (expert witnesses can testify about their opinions based on hypotheses).

Characteristic	Probability
Partly yellow automobile	1/10
Man with mustache	1/4
Woman with ponytail	1/10
Blond woman	1/3
Black man with beard	1/10
Interracial couple in a car	1/1000

Then the prosecutor said: the probability of having all of these characteristics is 1/12,000,000, overriding the expert's objection about their lack of independence. He continued: since there are 12,000,000 people in metropolitan Los Angeles Malcolm and Janet Collins must be the only couple with these characteristics and thus the perpetrators of the mugging in question. In addition to the problem with independence, of course, the probability of "more than one given at least one" in a Poisson distribution turns out to be .43, hardly the "beyond a reasonable doubt" required for a criminal conviction. The unfortunate Malcolm spent some time in prison before his conviction was overturned on appeal, as did the Garrett Wilson of

Maryland v. Wilson (2002), where not only was the probability of two children dying of Sudden Infant Death Syndrome similarly miscalculated, but the prosecutor argued not only that there was a low probability that two deaths would occur in one family but that there was a low probability that the defendant was innocent (This is called the "prosecutor's fallacy."). Analogous bad statistics in the UK led to the physician who testified about statistics being stricken from the registry and 250 prior convictions being reviewed. Unfortunately one of the victims of the erroneous testimony, faced with a ruined career as a solicitor, committed suicide when eventually released from prison.

But there are better results: statistics in cases I have worked on helped convince the courts that similarly situated women and men should receive equal pensions and that women's sports teams should be supported in colleges and universities as well as are men's. In the former case a man who had the same accumulation of pension funds in a defined contribution plan as a woman, was getting 15% more in monthly benefits on the stated grounds that (statistically speaking!) women live longer than men. The U.S. law clearly stated that discrimination on the basis of sex in employment-related matters such as pensions was forbidden, but the pension fund administrators insisted that the discrimination was on the basis of longevity, admitting of course that no individual woman could be expected to live long than any individual man. We showed that of a cohort of 1000 men and women at age 65, 7% of the population would be women could be expected to live longer than men with whom they could be matched and 7% of the population would be men who would die young, unmatched by women's early deaths. Hence 86% of the population could be paired up as to age at death – i.e., 86% of the men and women "died at the same age" (for statistical purposes). Thus for 86% of the population, those "similarly situated with respect to longevity," men and women were being treated differently. This together with the fact that, at least at the time (more than 20 years ago) men indulged in more voluntary life-shortening behavior like smoking and drinking to excess and the – what seemed to many – clear statutory mandate of equal treatment, convinced the courts.

In the sports case it was simply that 51% of the undergraduate students at Brown University were women, while only 39% of the student athletes were. The probability of such a disparity were it due to chance was about 1 in a million. Thus the courts found that the distribution of athletes by sex was not "substantially proportionate" to the distribution of students by sex. Statistical significance isn't everything, but in this case it prevented the cancellation of university support for some of the women's teams.

My late husband used to say that mathematics and the law both have axiom systems – it is just that the law's is inconsistent. Sometimes we all feel that way, but statistics can sometimes help bring justice.

About the Author

Dr. Mary W. Gray is a Professor and Chair, Department of Mathematics and Statistics, American University, Washington DC. She is the founding President of the Association for Women in Mathematics and past President of the Caucus for Women in Statistics. She was the Chair of the Department of Mathematics and Statistics (1977–1981, 1983–1985, 2001–2003). In 1976 Dr Gray was elected the second female Vice President of the American Mathematical Society (70 years after Charlotte Scott became the first female Vice President). In 1993 she became Chair of the USA Board of Directors of Amnesty International. She is an Elected member of the International Statistical Institute and a Fellow of the American Statistical Association, the American Association for the Advancement of Science, and the Association for Women in Science. She has authored and co-authored more than 100 papers and 2 books. Professor Gray has received the (U.S.) Presidential Award for Excellence in Science, Technology, Engineering and Mathematics Mentoring, the Lifetime Mentoring Award of the American Association for the Advancement of Science, and three honorary doctorates. Professor Gray has mentored twenty-three students through successful dissertations in mathematics, including fourteen women and eight African-American students. She has lectured throughout the United States, Europe, Latin America and the Middle East. She is a member of the District of Columbia and U.S. Supreme Court Bars. Currently, she is an Associate editor for the *International Journal of Surgery*.

Cross References

►Forensic DNA: Statistics in
►Misuse of Statistics
►Presentation of Statistical Testimony
►Statistical Evidence
►Statistical Significance

References and Further Reading

Asher J, Banks D, Scheuren F (eds) (2008) Statistical methods for human rights. Springer-Verlag, New York

Ball P, Asher J (2002) Statistics and Slobodan: using data analysis and statistics in the war crimes trial of former president Milosevic. Chance 15:17–24

Fienberg S, Kadane JB (1983) The presentation of Bayesian statistical analyses in legal proceedings. The Statistician 32:88–108
Fienberg S (ed) (1989) The evolving role of statistical assessments in the courts. Springer, New York
Fienberg SE, Krislov SH, Straf ML (1995) Understanding and evaluating statistical evidence in litigation. Jurimetrics Journal 36:1–32
Finkelstein MO, Levin B (2001) Statistics for lawyers, 2nd edn. Springer-Verlag, New York
Gastwirth JL (ed) (2000) Statistical science in the courtroom. Springer-Verlag, New York
Gray MW (1993) Can statistics tell us what we do not want to hear? The case of complex salary structures. Stat Sci 8:144–179
Gray MW (1996) The concept of "substantial proportionality" in Title IX athletics cases. Duke J Gender Soc Policy 3:165–185

Statistics Education

Richard L. Scheaffer
Professor Emeritus
University of Florida, Gainesville, FL, USA

Overview

Statistics education at all levels, school, undergraduate, graduate, and in the workplace, has been the subject of much debate over most of the 20th century and into the 21st. Proposals to make statistics a part of everyone's basic education surfaced in the 1930s and 1940s, but gained little traction. World War II forced a renewed emphasis on scientific thinking and statistics gained attention as an essential component of applied science and industrial management. This led to the few existing graduate programs in statistics being expanded and new ones being developed at various universities around the world, a trend that went on for about the subsequent forty years. Some of these programs emphasized application and some theory, but as the need for statistics in many different fields (business, engineering, health sciences, social sciences, to name a few) became essential and the advent of electronic computing made it possible to meet those needs, graduate programs in statistics tended to merge toward a combination of application and theory, a very healthy trend indeed.

During that same period, introductory undergraduate courses were developed, but these courses stayed on the theory track perhaps too long and only since about 1980 have been giving more attention to applications emphasizing data analysis, again with the assistance of ubiquitous computing. Work beyond the introductory course has not kept pace with the need; even today most colleges and universities offer little in the way of undergraduate statistics beyond the basic course.

Although overtures to making statistics a part of the school curriculum were advanced prior to the 1940s, nothing in that arena really took root until the early 1980s as well. Today, there is great debate on the place of statistics in the school curriculum, but most educators agree that it should be included in the broader picture of mathematical sciences to which all school students should be exposed before moving on to college or the workplace.

An enlightened 21st century view of the role of statistics in society was presented quite clearly in a recent article by Hal Varian of Google:

> The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it. (The McKinsey Quarterly, January 2009)

This view of the importance of statistics is becoming the predominant one among those affecting education in the mathematical sciences, and it appears that statistics education is on an upward swing as the information age continues.

University Education in Statistics

The American Statistical Association (http://www.amstat.org/) has links to lists that contain information on over 300 college and university programs in statistics around the world. This is a relatively small number, compared to, say, mathematics, and many of the programs are small or highly specialized (►biostatistics, for example). In the United States, the nearly one hundred graduate programs in statistics produced about 410 doctoral degrees and 408 master's degrees in the 2006–2007 academic year. A much smaller number of bachelors degree programs produced about 445 degrees in that same year. These numbers are underestimates, especially at the master's level, as they come from a survey of mathematical science departments conducted by the American Mathematical Society (http://www.ams.org/), but they do give a perspective on the relatively small numbers of degrees awarded in statistics at all levels. Yet, the number of job opportunities in statistics remains large even in times of economic downturn, especially for those with at least a master's degree

in the subject, and the number of degrees awarded lags behind demand.

Enrollments and other details on the undergraduate teaching of statistics in the United States can be found at in the CBMS 2005 Survey: *Statistical Abstract of Undergraduate Programs in the Mathematical Sciences in the United States* (http://www.cbmsweb.org/). Details on current thinking in the teaching of statistics at the college level can be found in one of two journals, the *Journal of Statistics Education* (http://www.amstat.org/PUBLICATIONS/JSE/) and the *Statistics Education Research Journal* of the International Association for Statistics Education (IASE) (http://www.stat.auckland.ac.nz/~iase/). The former is directed toward experiences with teaching practices in the classroom, often including useful data sets, while the latter is directed toward research on effective teaching and learning of statistics. A good resource on all aspects of undergraduate statistics education can be found at the Consortium for Advancing Undergraduate Statistics Education (CAUSE) (http://www.causeweb.org/).

School Education in Statistics

The modern era of statistics education at the school level dates from the late 1970s, when the United Kingdom, Australia, New Zealand and Sweden led the way in developing educational programs and materials that were effective in enlisting the interest of school children (as well as their teachers) in data analysis. The journal *Teaching Statistics* (http://ts.rsscse.org.uk/), now a product of the Royal Statistical Society's Center for Statistics Education, was an outcome of those efforts in the UK and still remains a premier source of information on effective teaching of statistics in the schools. These efforts influenced work in the United States that led the National Council of Teachers of Mathematics (NCTM) (http://www.nctm.org/) to place an emphasis on data analysis in their *Principles and Standards for School Mathematics*, first published in 1989 and revised in 2000.

Over the years, national and international assessments of school mathematics have included increasingly larger emphases on data analysis, statistics and probability. In its 2006 framework, the OECD Program for International Student Assessment (PISA) (http://www.pisa.oecd.org/) lists Uncertainty as one of the four main areas of mathematics, along with Space and shape, Change and relationships, and Quantity. There description of this area is enlightening:

> As an overarching idea, *uncertainty* suggests two related topics: data and chance. These phenomena are respectively the subject of mathematical study in statistics and probability. Relatively recent recommendations concerning school curricula are unanimous in suggesting that statistics and probability should occupy a much more prominent place than has been the case in the past. Specific mathematical concepts and activities that are important in this area are collecting data, data analysis and display/visualization, probability and inference.

For the United States, the 2009 framework of the National Assessment of Educational Progress (NAEP) (http://www.nagb.org/publications/frameworks/math-framework09.pdf) gives data analysis, statistics and probability 25% of the weight of questions at the high school level, in connection with number properties (10%), measurement and geometry (30%) and algebra (35%).

As to content emphases, *the Guidelines for Assessment and Instruction in Statistics Education* (GAISE) (http://www.amstat.org/education/gaise/) report of the American Statistical Association has been instrumental in shaping the revision of mathematics standards for many states and some other countries. GAISE views statistics as a problem-solving process built around the steps of:

- Formulate questions
- Collect data
- Analyze data
- Interpret results

Its guiding principles for teaching statistics are:

- Conceptual understanding takes precedence over procedural skill.
- Active learning is key to the development of conceptual understanding.
- Real-world data must be used wherever possible in statistics education.
- Appropriate technology is essential in order to emphasize concepts over calculations.
- All four steps of the investigative process should be encountered at each grade level.
- The illustrative investigations should show situations in which the statistics is essential to the answering of a question, not just an add-on.
- Such investigations should be tied to the mathematics that they illustrate, motivate and emphasize.

Statistics in the Workplace

As Hal Varian expressed it in the article cited above, "I keep saying the sexy job in the next ten years will be statisticians." There seems to be no end of the demand for

statisticians, or those trained in statistics, so long as they can combine theoretical knowledge and problem-solving skills with the ability to do practical work with data and computers. Another manifestation of the huge need for statistical knowledge lies in the area of productivity and product improvement in industry, as reflected by the interest and excitement that surrounds the Six Sigma program. (See the American Society for Quality, Six Sigma program at http://www.asq.org/learn-about-quality/six-sigma/overview/overview.html.)

Statistics has a bright future, and statistics education must expand and adapt to meet the increasing needs of a world economy that runs on data.

About the Author

Dr. Richard Scheaffer is Professor Emeritus in statistics at Department of Statistics, Florida State University. He was Chairman of the Department for a period of 12 years. He has published numerous papers in the statistical literature and is co-author of five textbooks covering aspects of sampling, probability and mathematical statistics. In recent years, he focused on statistics education throughout the school and college curriculum. He was one of the developers of the Quantitative Literacy Project in the United States that formed the basis of the data analysis emphasis in the mathematics curriculum standards recommended by the National Council of Teachers of Mathematics. He continues to work on educational projects at the elementary, secondary and college levels, and served as the Chief Faculty Consultant for the Advanced Placement Statistics Program in the United States during its first two years (1997–1998). Dr. Scheaffer is a Fellow and Past President of the American Statistical Association (2001), from whom he has received a Founder's Award.

Cross References

►Business Statistics
►Careers in Statistics
►Data Analysis
►Decision Trees for the Teaching of Statistical Estimation
►Learning Statistics in a Foreign Language
►Online Statistics Education
►Promoting, Fostering and Development of Statistics in Developing Countries
►Rise of Statistics in the Twenty First Century
►Role of Statistics in Advancing Quantitative Education
►Statistical Literacy, Reasoning, and Thinking
►Statistics and Climate Change
►Statistics: Nelder's view

References and Further Reading

American Mathematical Society: http://www.ams.org/
American Society for Quality, Six Sigma
American Statistical Association, Guidelines for Assessment and Instruction in Statistics Education (GAISE): http://www.amstat.org/education/gaise/
Consortium for advancing undergraduate statistics education (CAUSE): http://www.causeweb.org/
http://www.asq.org/learn-about-quality/six-sigma/overview/overview.html
International Association for Statistics Education (IASE), Statistics Education Research Journal: http://www.stat.auckland.ac.nz/~iase/
Journal of Statistics Education: http://www.amstat.org/PUBLICATIONS/JSE/
National Council of Teachers of Mathematics (NCTM): http://www.nctm.org/
Statistical abstract of undergraduate programs in the mathematical sciences in the United States: http://www.cbmsweb.org/
Teaching statistics: http://ts.rsscse.org.uk/
Conference Board of the Mathematical Sciences (CBMS) 2005 Survey
National Assessment of Educational Progress (NAEP), Mathematics framework for 2009: http://www.nagb.org/publications/frameworks/math-framework09.pdf
OECD Programme for International Student Assessment (PISA), A framework for PISA 2006: http://www.pisa.oecd.org/

Statistics of Extremes

Anthony C. Davison
Professor
Ecole Polytechnique Fédérale de Lausanne,
EPFL-FSB-IMA-STAT, Lausanne, Switzerland

Introduction

Statistics of extremes concerns the occurrence of rare events: catastrophic flooding due to very high tides or landslides following unusually heavy rain, structural failure of dams and bridges, massive earthquakes, stock market crashes, and so forth. It has applications in many domains of engineering, in meteorology, hydrology and other earth sciences, in telecommunications, in finance and insurance – indeed, in any domain in which major risks arise due to unusual events or combinations thereof. In applications the available data are often very limited in relation to the event of interest, so a key issue is the validity of extrapolation far into the tail of a distribution, based on data that are less extreme. This is usually formulated mathematically in terms of stability properties that reasonable models ought to possess, and these properties place strong restrictions on the families of distributions on

which extrapolation should be based. The relevance of such properties to an application must be carefully considered, and any relevant subject-matter knowledge incorporated, if wholly inappropriate extrapolation is to be avoided.

Maxima

Consider the maximum $M_k = \max(X_1, \ldots, X_k)$ of independent identically distributed continuous random variables $X_1, \ldots, X_k$ from a distribution F whose upper support point is $x_{\max} = \sup\{x : F(x) < 1\}$. In analogy with the central limit theorem (see ►Central Limit Theorems), we seek a useful limiting distribution for M_k as $m \to \infty$. The distribution function of M_k is $F^k(x)$, but this converges to a degenerate distribution putting unit mass at $x_{\max}$, so instead we consider the sequence of linearly rescaled maxima $Y_k = (M_k - b_k)/a_k$ for $b_k \in \mathbb{R}$ and $a_k > 0$, and ask whether the sequences $\{a_k\}, \{b_k\}$ can be chosen so that a non-degenerate limiting distribution exists. Remarkably it can be shown that if such a limit exists, it must lie in the generalized extreme-value family

$$H(y) = \exp\left\{-\left[1 + \xi\left(\frac{y-\eta}{\tau}\right)\right]_+^{-1/\xi}\right\}, \quad -\infty < \eta, \xi < \infty, \tau > 0, \tag{1}$$

where $x_+ = \max(x, 0)$. This result, known as the extremal types theorem, provides strong motivation for the use of (1) when modeling maxima, in analogy with the use of the Gaussian distribution for averages. Note however the conditional nature of the theorem: there is no guarantee that such a limiting distribution will exist in practice. The connection with the stability properties mentioned above is that (1) is the entire class of so-called max-stable distributions, i.e., those satisfying the natural functional stability relation $H(y)^m = H(b_m + a_m y)$ for suitable sequences $\{a_m\}, \{b_m\}$ for all $m \in \mathbb{N}$.

The parameters η and τ in (1) are location and scale parameters. The shape parameter ξ plays a central role, as it controls the behavior of the upper tail of the distribution H. Taking $\xi > 0$ gives distributions with heavy upper tails and taking $\xi < 0$ gives distributions with a finite upper endpoint, while the Gumbel distribution function $\exp\{-\exp[-(y-\eta)/\tau]\}$ valid for $-\infty < y < \infty$ emerges as $\xi \to 0$. Fisher and Tippett (1928) derived these three classes of distributions, which are known as the Gumbel or Type I class when $\xi = 0$, the Fréchet or Type II class when $\xi > 0$, and the (negative or reversed) Weibull or Type III class when $\xi < 0$. The appearance of the ►Weibull distribution signals that there is a close link with reliability and with survival analysis, though in those contexts the behavior of minima is typically the focus of interest. Since $\min(X_1, \ldots, X_k) = -\max(-X_1, \ldots, -X_k)$, results for maxima may readily be converted into results for minima; for example, the extremal types theorem implies that if a limiting distribution for linearly rescaled minima exists, it be of form $1 - H(-y)$. Below we describe the analysis of maxima, but the ideas apply equally to minima.

Application

A typical situation in environmental science is that n years of daily observations are available, and then it is usual to fit the generalized extreme-value distribution (1) to the n annual maxima, effectively taking $k = 365$ and ignoring any seasonality or dependence in the series. The fitting is typically performed by maximum likelihood estimation or by Bayesian techniques. The method of moments is generally quite inefficient relative to maximum likelihood because (1) has a finite rth moment only if $r\xi < 1$. Often in environmental applications it is found that $|\xi| < 1/2$, but in financial applications second and even first moments may not exist. Probability weighted moments fitting of (1) is quite widely performed by hydrologists, but unlike likelihood estimation, this method is too inflexible to deal easily more complex settings, for example trend in location or censored observations.

The parameters of (1) are rarely the final goal of the analysis, which usually focuses on quantities such as the $1/p$-year return level, i.e., the level exceeded once on average every $1/p$ years; here $0 < p < 1$. The quantity $1/p$ is known as the return period and is important in engineering design. The usual return level estimate is the $1-p$ quantile of (1),

$$y_{1-p} = \eta + \frac{\tau}{\xi}\left\{[-\ln(1-p)]^{-\xi} - 1\right\},$$

with parameters replaced by estimates. Analogous quantities, the value at risk and expected shortfall, play a central role in the regulation of modern financial markets. Two major concerns in practice are that inference is often required for a return period much longer than the amount of data available, i.e., $np \ll 1$, and that the fitted distribution is very sensitive to the values of the most extreme observations; these difficulties are inherent in the subject.

Threshold Exceedances

The use of annual maxima alone seems to be wasteful of data: much sample information is ignored. A potentially more efficient approach may be based on the following characterization. Let $X_1, \ldots, X_{nk}$ be a set of nk independent identically distributed random variables, and consider the planar point pattern with points at (x, y) coordinates $(j/(nk+1), a_k(X_j - b_k))$, $j = 1, \ldots, nk$.

Then provided a_k and b_k are chosen so that the limiting distribution for $(M_k - b_k)/a_k$ as $k \to \infty$ is given by expression (1), the empirical point pattern above a high threshold t will converge to a nonhomogeneous Poisson process (see ►Poisson Processes) with measure

$$\begin{aligned}&\Lambda\{(x_1, x_2) \times (u, \infty)\} \\ &\quad = \exp\left[-n(x_2 - x_1)\left(1 + \xi\frac{u-\eta}{\tau}\right)_+^{-1/\xi}\right], \\ &\quad 0 < x_1 < x_2 < 1, u > t. \end{aligned} \tag{2}$$

A variety of results follow. For example, on noting that the rescaled maximum of k observations, M_k, is less than $y > t$ only if there are no points in the set $(0, 1/n) \times (y, \infty)$, (2) immediately gives (1). The model (2) shows that if N observations, $y_1, \ldots, y_N$, exceed a threshold $u > t$ over a period of n years, their joint probability density function is

$$\exp\left[-n\left(1 + \xi\frac{u-\eta}{\tau}\right)_+^{-1/\xi}\right]\prod_{j=1}^{N}\frac{1}{\tau}\left(1 + \xi\frac{y_j - \eta}{\tau}\right)_+^{-1/\xi-1},$$

which can be used as a likelihood for η, τ, and ξ. Maximum likelihood inference can be performed numerically for this point process model (see ►Point Processes) and regression models based on it. A popular and closely related approach is the fitting of the generalized Pareto distribution

$$\Pr(X \le t + y \mid X > t) = G(y) = 1 - (1 + \xi y/\tau)_+^{-1/\xi}, \quad y > 0; \tag{3}$$

to the exceedances over the threshold t. As $\xi \to 0$ expression (3) becomes the exponential distribution with mean τ, which here occupies the same central role as the Gumbel distribution for maxima. The distribution (3) has the stability property that if $X \sim G$, then conditional on $X > u$, $X - u$ also has distribution G, but with parameters ξ and $\tau_u = \tau + u\xi$. The conditioning in (3) appears to remove dependence on the location parameter η, but this is illusory because the probability of an exceedance of t must be modeled in this setting.

One important practical matter is the choice of threshold t. Too high a value for t will result in loss of information about the process of extremes, while too low a value will lead to bias because the point process model applies only asymptotically for high thresholds. The value of t is usually chosen empirically, by calculating parameter estimates and other quantities of interest for a number of thresholds and choosing the lowest above which the results appear to be stable. In practice the threshold exceedances are typically dependent owing to clustering of rare events, and this is usually dealt with by identifying clusters of exceedances, and fitting (3) to the cluster maxima, a procedure that may be justified using the asymptotic theory.

Dependence

The discussion above has assumed that the data are independent, but this is rare in practice. Fortunately there is a well-developed probabilistic theory of extremes for stationary dependent continuous time series. To summarize: under mild conditions on the dependence structure, the limiting distribution (1) again emerges as the limit for the maximum, but with a twist. Suppose that $X_1, \ldots, X_k$ are consecutive observations from such a series, that $X_1^*, \ldots, X_k^*$ are independent observations with the same marginal distribution, F, and that M_k and M_k^* are the corresponding maxima. Then it turns out that there exist sequences $\{a_k\}$ and $\{b_k\}$ such that $\left(M_k^* - b_k\right)/a_k$ has limiting distribution H if and only if $(M_k - b_k)/a_k$ has limiting distribution H^θ, where the parameter $\theta \in (0, 1]$ is known as the extremal index (Leadbetter et al. 1983). This quantity has various interpretations, the most direct being that θ^{-1} is the mean size of the clusters of extremes that appear in dependent data. The case $\theta = 1$ corresponds to independence but also covers many other situations: for example, Gaussian autoregressions of order p also have $\theta = 1$. This raises a general problem in the statistics of extremes, that of the relevance of asymptotic arguments to applications: this result indicates that extremely rare events will occur singly, but for levels of interest, there may be appreciable clustering that must be modeled.

Further Reading

The probabilistic basis of extremes is discussed from different points of view by Galambos (1987), Resnick (1987) and de Haan and Ferreira (2006), and Resnick (2006) discusses the closely related topic of heavy-tailed modeling. A historically important book on statistics of extremes is Gumbel (1958). Coles (2001) and Beirlant et al. (2004) give modern accounts, the former focusing exclusively on modeling using likelihood methods, and the latter taking a broader approach. Embrechts et al. (1997) give an discussion oriented towards finance, while Castillo (1988) is turned towards applications in engineering; as mentioned above there is a close connection to the extensive literature on survival analysis and reliability modeling. The essays in Finkenstädt and Rootzén (2004) provide useful overviews of various topics in extremes.

One important topic not discussed above is multivariate extremes, such as the simultaneous occurrence of rare events in many financial time series, or environmental events such as heatwaves or severe rainstorms. Much current research activity is devoted to this domain, which has

obvious implications for ►risk analysis and management. In addition to the treatments in the books cited above, Kotz and Nadarajah (2000) provide extensive references to the early literature on multivariate extremes. Balkema and Embrechts (2007) take a more geometric approach.

The journal *Extremes* (http://www.springer.com/statistics/journal/10687) provides an outlet for both theoretical and applied work on extremal statistics and related topics.

About the Author

Professor Davison is Editor of *Biometrika* (2008–). He is an elected Fellow of the American Statistical Association and the Institute of Mathematical Statistics, an elected member of the International Statistical Institute, and a Chartered Statistician. Professor Davison has published on a wide range of topics in statistical theory, methods and applications. He has also co-written highly-regarded books, including *Bootstrap Methods and their Application* (with D.V. Hinkley, Cambridge University Press, Cambridge, 1997) and *Statistical Models* (Cambridge University Press, Cambridge, 2003). In 2009, Professor Davison was awarded a *laurea honoris causa* in Statistical Science by the University of Padova, Italy.

Cross References

►Environmental Monitoring, Statistics Role in
►Extreme Value Distributions
►Fisher-Tippett Theorem
►Generalized Extreme Value Family of Probability Distributions
►Generalized Weibull Distributions
►Insurance, Statistics in
►Methods of Moments Estimation
►Point Processes
►Poisson Processes
►Quantitative Risk Management
►Statistical Aspects of Hurricane Modeling and Forecasting
►Statistical Modeling of Financial Markets
►Testing Exponentiality of Distribution
►Weibull Distribution

References and Further Reading

Balkema G, Embrechts P (2007) High risk scenarios and extremes. European Mathematical Society, Zürich
Beirlant J, Goegebeur Y, Teugels J, Segers J (2004) Statistics of extremes: theory and applications. Wiley, New York
Castillo E (1988) Extreme value theory in engineering. Academic, New York
Coles SG (2001) An introduction to statistical modeling of extreme values. Springer, New York
de Haan L, Ferreira A (2006) Extreme value theory: an introduction. Springer, New York
Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling extremal events for insurance and finance. Springer, Berlin
Finkenstädt B, Rootzén H (eds) (2004) Extreme values in finance, telecommunications, and the environment. Chapman and Hall/CRC, New York
Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distributions of the largest or smallest member of a sample. Proc Camb Philos Soc 24:180–190
Galambos J (1987) The asymptotic theory of extreme order statistics, 2nd edn. Krieger, Melbourne, FL
Gumbel EJ (1958) Statistics of extremes. Columbia University Press, New York
Kotz S, Nadarajah S (2000) Extreme value distributions: theory and applications. Imperial College Press, London
Leadbetter MR, Lindgren G, Rootzén H (1983) Extremes and related properties of random sequences and processes. Springer, New York
Resnick SI (1987) Extreme values, regular variation and point processes. Springer, New York
Resnick SI (2006) Heavy-tail phenomena: probabilistic and statistical modeling. Springer, New York

Statistics on Ranked Lists

Michael G. Schimek
Professor
Medical University of Graz, Graz, Austria

Introduction

In various fields of application, we are confronted with lists of distinct objects in rank order because we can always rank objects according to their position on a scale. When we have variate values (interval or ratio scale), we might replace them by corresponding ranks. In the latter case, there is a loss of accuracy but a gain in generality. The ordering might be due to a measure of strength of evidence or to an assessment based on expert knowledge or a technical device. Taking advantage of the generality of the rank scale, we are in the position of ranking objects which might otherwise not be comparable across lists, for instance, because of different assessment technologies or levels of measurement error. This is a direct result of the fact that rankings are invariant under the stretching of the scale.

In this article, we focus primarily on statistics for two ranked lists comprising all elements of a set of objects (i.e., no missing elements). Due to limited space, we will not discuss methods for m lists in detail but give an example at the end and some references. Let us assume two (but it could be

up to m) assessors, one of which ranks N distinct objects according to the extent to which a particular attribute is present. The ranking is from 1 to N, without ties. The other assessor also ranks the objects from 1 to N. Historically, the goal of rank order statistics was to have a handle that allows the avoidance of the difficulty of setting up an objective scale in certain applications such as in psychometrics. It all started about 100 years ago with seminal work of the psychologist and statistician Charles E. Spearman (1863–1945) aiming at a measure of association between ranked lists. Nowadays, there are four primary tasks when analyzing rank scale data: (1) measuring association between ranked lists, (2) measuring distance between ranked lists, (3) identification of significantly overlapping sublists (estimation of the point of degeneration of paired rankings into noise), and (4) aggregation of ranked full lists or sublist.

Association between Ranked Lists

Suppose we have $N = 10$ major cities ranked according to a measure of air pollution (e.g., particulate matter) and the prevalence of respiratory disease (Table 1).

We are interested in the degree of association between these two rankings representing air pollution and disease prevalence. Such a measure of association is the Kendall's τ coefficient (Kendall 1938, 1942). Let us consider any pair of objects (o_i, o_j). Is the pair in direct order, we score for this pair +1, is it in inverse order, we score for this pair −1. Then the scores obtained for the two lists for a fixed pair of objects are multiplied, giving a common score. This procedure is performed for all $\frac{1}{2}N(N-1)$ possible pairs (45 in this example). Finally, the total of the positive scores, say P, and of the negative scores, say Q, is calculated. The overall score $S = P + Q$ is divided by the maximum possible score (the value that S takes when all rankings are identical). This heuristic procedure defines the τ coefficient which in our example is $\tau = 0.644$. A zero value would indicate independence (no association). τ takes 1 for complete agreement and −1 for complete disagreement. In practice, there are more efficient ways to calculate τ. The coefficient can be interpreted as a measure of concordance between two sets of N rankings (P is the number of concordant pairs, Q of discordant pairs, and S is the excess of concordant over discordant pairs) as well as a coefficient of disarray (minimum moves necessary to transform the second list into the natural order of the first one by successively interchanging pairs of neighbors).

Statistics on Ranked Lists. Table 1 Example of two rankings for ten cities ordered according to pollution rank

City (object)	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
Pollution	1	2	3	4	5	6	7	8	9	10
Disease	3	1	2	5	8	6	4	9	7	10

Another famous measure of association is Spearman's ρ, also called rank correlation coefficient (Spearman 1904). Let d_i be the difference between the ranks in the two lists for object o_i (for the N objects these differences sum to zero). The coefficient is of the form

$$\rho = 1 - \frac{6\sum_i d_i^2}{N^3 - N}. \quad (1)$$

When two rankings are identical, it follows from (1) that $\rho = 1$, in the case of reverse order we have $\rho = -1$ (in our example $\rho = 0.818$). Q, the total of the negative scores for Kendall's τ coefficient, is equivalent to the number of pairs which occur in different orders in the two lists forming so-called inversions. Thus τ is a linear function of the number of inversions and ρ can be interpreted as a coefficient of inversion when each inversion is weighted. If a pair of ranks (i, j) is inverted $(i < j)$, we score $(j - i)$ for any inversion, then the sum of all such scores totals to V. One can show that

$$\rho = 1 - \frac{12V}{N^3 - N},$$

where V can also be expressed as $\frac{1}{2}\sum_i d_i^2$.

A detailed account of rank correlation methods summarizing the classical literature up to 1990 can be found in Kendall and Gibbons (1990). Around that time there was little interest in procedures for ranked data, some of them, like Spearman's L_1-based footrule (Spearman 1906), were almost unknown in the statistical community because of technical and computational shortcomings, as well as a lack of relevance for common applications. Most recently, there has been a dramatic shift in relevance because of emerging technologies producing huge amounts of ranked lists, such as Web search engines offering selected server-based information and high-throughput techniques in genomics providing insight into gene expression. These and others have given rise to new developments concerning the statistical handling of rank scale information. An essential aspect is the measurement of distance between ranked lists.

Distance between Ranked Lists

The most popular distance measure is Kendall's τ intrinsic to his already introduced measure of association. It is equal to the number of adjacent pairwise exchanges required to convert one ranking to another. Let us have two

permutations τ and τ' of a set O of objects. Then Kendall's τ distance is given by

$$K(\tau,\tau') = \sum_{\{i,j\}\in O} K_{i,j}(\tau,\tau'),$$

where $K_{i,j}(\tau,\tau')$ takes 0 if the orderings of the ranks of objects i and j agree in the two lists and otherwise 1. Its maximum is $\frac{1}{2}N(N-1)$ where N is the list length.

An alternative measure of distance is Spearman's footrule (related to the Manhattan distance for variate values). Let us again assume two permutations τ and τ' of a set O of objects. Spearman's footrule distance is the sum of the absolute differences between the ranks of the two lists over the N elements in O,

$$S(\tau,\tau') = \sum_{i=1}^{N} |R_\tau(o_i) - R_{\tau'}(o_i)|,$$

where $R_\tau(o_i)$ is the rank of object o_i in list τ, and $R_{\tau'}(o_i)$ in list τ', respectively. As can be seen from the above formulae, Spearman's footrule takes the actual rankings of the elements into consideration, whereas, in Kendall's τ only relative rankings matter. The maximum Spearman's distance is $\frac{1}{2}|N|^2$ for N even, and $\frac{1}{2}(|N|+1)(|N|-1)$ for N odd, which corresponds to the situation in which the two lists are exactly the reverse of each other.

For a mathematical theory of distance measures, we refer to Fagin et al. (2003). Recent developments as well as novel applications are discussed in Schimek et al. (2011).

Degeneration of Rankings into Noise

Typically, when the number N of objects is large or even huge, it is unlikely that consensus between two rankings of interest prevails. Only the top-ranked elements might be relevant. For the remainder objects their ordering is more or less at random. This is not only true for surveys of consumer preferences but also for many other applications of topical interest such as the ▶meta-analysis of gene expression data from several laboratories. In many instances, we observe a general decrease of the probability for consensus rankings with increasing distance from the top rank position. Typically, there is reasonable conformity in the rankings for the first, say k, elements of the lists, motivating the notion of *top-k ranked lists*.

The statistical challenge is to identify the length of the top list. So far, heuristics have been used in practice to specify k. Recently Hall and Schimek (2010) could derive a moderate deviation-based inference procedure for random degeneration in paired ranked lists. The result is an estimate $\hat{k}$ for the length of the so-called partial (top-k) list. Such an inference procedure is not straightforward since the degree of correspondence between ranked lists (full or partial) is not necessarily high, due to various irregularities of the assessments.

Let us define a sequence of indicators, where $I_j = 1$ if the ranking given by the second assessor to the object ranked j by the first assessor, is not distant more than δ index positions from j, and otherwise $I_j = 0$. Further, let us assume (1) independent Bernoulli random variables $I_1,\ldots,I_N$, with $p_j \geq \frac{1}{2}$ for each $j \leq j_0 - 2$, $p_{j_0-1} > \frac{1}{2}$, and $p_j = \frac{1}{2}$ for $j \geq j_0$; (2) a general decrease of p_j for increasing j that does not need to be monotone. The index j_0 is the point of degeneration into noise and needs to be estimated ($\hat{j}_0 - 1 = \hat{k}$). Then for a pilot sample size v a constant $C > 0$ is chosen such that $z_v \equiv \left(C v^{-1} \log v\right)^{1/2}$ is a moderate-deviation bound for testing the null hypothesis H_0 that $p_k = \frac{1}{2}$ for v consecutive values of k, versus the alternative H_1 that $p_k > \frac{1}{2}$ for at least one of the values of k. In particular, it is assumed that H_0 applies to the v consecutive values of k in the respective series defined by

$$\hat{p}_j^+ = \frac{1}{v}\sum_{\ell=j}^{j+v-1} I_\ell \quad \text{and} \quad \hat{p}_j^- = \frac{1}{v}\sum_{\ell=j-v+1}^{j} I_\ell,$$

where $\hat{p}_j^+$ and $\hat{p}_j^-$ are estimates of p_j computed from the v data pairs I_ℓ for which ℓ lies immediately to the right of j, or immediately to the left of j, respectively. We reject H_0 if and only if $\hat{p}_j^\pm - \frac{1}{2} > z_v$. Under H_0, the variance of $\hat{p}_j^\pm$ equals $(4v)^{-1}$ (this implies $C > \frac{1}{4}$). Taking advantage of this inference procedure, the complex decision problem is solved via an iterative algorithm, adjustable for irregularity in the rankings.

Aggregation of Ranked Lists

The task of rank aggregation is to provide consensus rankings (majority preferences) of objects across lists, thereby producing a conforming subset of objects O^*. The above described inference procedure facilitates rank aggregation because it helps to specify the partial list length k which means a substantial reduction in the associated computational burden. As a matter of fact, list aggregation by means of brute force is limited to the situation where N is unrealistically small. The approach proposed in Lin and Ding (2009) which we describe below, outperforms most of the aggregation techniques so far but for large sets O, the specification of k beforehand remains crucial. It is a stochastic search algorithm that provides an optimal solution, i.e., a consolidated list of objects, for a given distance measure such as Kendall's τ or Spearman's footrule, to be precise, for their penalized versions because of the partial nature of the input lists (for details see Schimek et al. 2011). Lin's and Ding's algorithm is preferable to those that do not aim to optimize any criterion, thus only providing approximate

solutions under unknown statistical properties (examples are Dwork et al. 2001, DeConde et al. 2006).

Let us assume a random matrix $(\mathbf{X})_{N\times k}$ with elements 0 and 1 with the constraints of its columns summing up to 1 and its rows summing up to, at most, 1. Under this setup, each realization of $\mathbf{X}$, x, uniquely determines an ordered list (permutation) of length k by the position of 1's in each column from left to right. Let $\mathbf{p} = (p_{jr})_{N\times k}$ denote the corresponding probability matrix (each column sums to 1). For each column variable, $\mathbf{X}_r = (X_{1r}, X_{2r}, \ldots, X_{Nr})$, a ►multinomial distribution with sample size 1 and probability vector $\mathbf{p}_r = (p_{1r}, p_{2r}, \ldots, p_{Nr})$ is assumed. Then the probability mass function is of the form

$$P_{\mathbf{v}}(x) \propto \prod_{j=1}^{N}\prod_{r=1}^{k}(p_{jr})^{x_{jr}} I\left(\sum_{r=1}^{k} x_{jr} \le 1, 1 \le j \le N; \sum_{j=1}^{N} x_{jr} = 1, 1 \le r \le k\right).$$

Any realization x of $\mathbf{X}$ uniquely determines the corresponding top-k candidate list without reference to the probability matrix $\mathbf{p}$. The idea is to construct a stochastic search algorithm to find an ordering x^* that corresponds to an optimal τ^* satisfying the minimization criterion. Lin and Ding (2009) use a cross-entropy Monte Carlo technique in combination with an Order Explicit algorithm (since the orders of the objects in the optimal list are explicitly given in the probability matrix $\mathbf{p}$). Cross-entropy Monte Carlo is iterating between two steps: a simulation step in which random samples from $P_{\mathbf{v}}(x)$ are drawn, and an update step producing improved samples increasingly concentrating around an x^* corresponding to an optimal τ^*.

Let us finally illustrate the application of the inference procedure together with rank aggregation as outlined in this paper. We simulated $m = 5$ ranked lists τ_j of gene expression data ($N = 60$ genes) from a known central ranking as outlined in DeConde et al. (2006). The length of the top-k list was set to 10. In Table 2, we display the input lists and the output top-k list for $\delta = 10$ and $v = 16$, applying the (penalized) Kendall's τ distance. We obtained an estimated $\hat{k} = 8$ instead of the true $k = 10$. Most objects ranked in input position 9 and 10 are displaced due to irregular (random) assignments. Therefore our procedure was short-cutting the top-ranked elements for the sake of clear separation. However, a longer partial list could have been obtained by parameter adaptations in the moderate deviation-based inference procedure. All calculations were carried out with the R package TopKLists of the author and collaborators.

Statistics on Ranked Lists. Table 2 Example of the aggregation of five rankings of $N = 60$ objects (genes) and the consensus top-ranking set of $\hat{k} = 8$ objects

	Input lists					Output list
Rank	τ_1	τ_2	τ_3	τ_4	τ_5	τ^*
1	o_8	o_1	o_{12}	o_4	o_3	o_2
2	o_{10}	o_5	o_{45}	o_{10}	o_7	o_5
3	o_4	o_{37}	o_2	o_6	o_5	o_4
4	o_7	o_4	o_5	o_2	o_{46}	o_6
5	o_{50}	o_6	o_9	o_9	o_2	o_8
6	o_6	o_{20}	o_8	o_3	o_8	o_{10}
7	o_{40}	o_2	o_6	o_7	o_{32}	o_3
8	o_{55}	o_{34}	o_3	o_1	o_{41}	o_7
9	o_{33}	o_{47}	o_{28}	o_{40}	o_{44}	–
10	o_{21}	o_{44}	o_{26}	o_{11}	o_1	–
11	o_{15}	o_{19}	o_{60}	o_{46}	o_9	–
12	o_{14}	o_{57}	o_{38}	o_{16}	o_{55}	–
13	o_{54}	o_{46}	o_1	o_{54}	o_{42}	–
14	o_{13}	o_8	o_{41}	o_{43}	o_{40}	–
15	o_{53}	o_{36}	o_{15}	o_{35}	o_{27}	–
⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	o_{35}	o_{16}	o_{39}	o_{48}	o_{39}	–

About the Author

Professor Michael G. Schimek is Past Vice President of the International Association for Statistical Computing, Member of the International Statistical Institute, Fellow and Chartered Statistician of the Royal Statistical Society, as well as Adjunct Professor of Masaryk University in Brno (Czech Republic).

Cross Reference

►Distance Measures
►Kendall's Tau
►Measures of Dependence
►Moderate Deviations
►Ranks

References and Further Reading

DeConde RP et al (2006) Combined results of microarray experiments: a rank aggregation approach. Stat Appl Genet Mol Biol 5(1):Article 15

Dwork C et al (2001) Rank aggregation methods for the Web. http://www10.org/cdrom/papers/577/

Fagin R, Kumar R, Sivakumar D (2003) Comparing top-k lists. SIAM J Discrete Math 17:134–160

Hall P, Schimek MG (2010) Moderate deviation-based inference for random degeneration in paired rank lists. Submitted manuscript

Kendall M (1938) A new measure of rank correlation. Biometrika 30:91–93

Kendall M (1942) Note on the estimation of a ranking. J R Stat Soc A 105:119–121

Kendall M, Gibbons JD (1990) Rank correlation methods. Edward Arnold, London

Lin S, Ding J (2009) Integration of ranked lists via Cross Entropy Monte Carlo with applications to mRNA and microRNA studies. Biometrics 65:9–18

Schimek MG, Lin S, Wang N (2011) Statistical integration of genomic data. Springer, New York (forthcoming)

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101

Spearman C (1906) A footrule for measuring correlation. Brit J Psychol 2:89–108

Statistics Targeted Clinical Trials Stratified and Personalized Medicines

Aboubakar Maitournam
University Abdou Moumouni of Niamey, Niamey, Niger

The rapid breakthroughs in genomics-based technologies like DNA sequencing, microarrays for gene expression and mRNA transcript profiling, comparative genomic hybridization (CGH), and mass spectrometry for protein characterization and identification of metabolic and regulatory pathways and networks announce the advent of stratified medicine and its immediate corollary called personalized medicine. Both stratified and personalized medicine are in their infancy. But, they already raise statistical and stochastic modeling challenges partially handled by the growing multidisciplinary field of ▶bioinformatics.

Statistics, Targeted Clinical Trials, and Stratified Medicine

With the actual progress in the burgeoning field of genomic science, most of the common diseases like cancer can be stratified at the molecular level. The aim is to refine disease taxonomies and to allocate patients to molecularly targeted therapy subgroups based on prognostic and predictive biomarkers. This will improve the efficiency of the treatment by adapting it to the patient prognostic profile. However, molecularly targeted therapy benefits only a subset of patients (Betensky et al. 2002). The refinement of the disease classification is based on gene expression transcript profiling, and the prediction of which patients will be more responsive to the experimental treatment than to the control regimen may be based on a molecular assay measuring, for example, expression of targeted proteins.

For stratified medicine, both molecular signatures of patients and of the diseases can be used, firstly for stratification of patients into responder and nonresponder groups and, secondly, in the near future also for individualized therapy. Stratification of patients into responder or nonresponder groups based on theranostics (molecular diagnosis assays) is the basis of stratified medicine. This implies that the first steps toward stratified medicine are randomized clinical trials for the evaluation of molecularly targeted therapy called targeted clinical trials (Simon 2004). Targeted clinical trials have eligibility restricted to patients predicted to be responsive to the molecularly targeted drug.

In a modeling of phase III randomized clinical trials for the evaluation of molecularly targeted therapy, (Maitournam and Simon 2004 and Simon and Maitournam 2004) established that the targeted clinical trial design is more efficient than a conventional untargeted design with broad eligibility. They evaluated relative efficiencies, e_1 and e_2, of the two designs, respectively, with respect to the number of patients required for randomization $\left(e_1 = \frac{n}{n_T}\right)$ and relatively to the number required for screening

$$\left(e_2 = n \Big/ \left(n_T \Big/ \left((1-\lambda_{spec})\gamma + \lambda_{sens}(1-\gamma) \right) \right) \right),$$

where $2n$ is the total number of randomized patients for untargeted design, $2n_T$ is that of targeted design, λ_{spec} and λ_{sens} are the specificity and the sensitivity of the molecular diagnosis assay, and γ is the proportion of not responders in the referral population. Indeed, for untargeted design, n patients are allocated to control group and n other patients to treatment group. Consequently, the total number of randomized patients for untargeted design is $2n$. In the same way, for targeted design the total number of randomized patients is $2n_T$. Thus, the relative efficiencies are respectively

$$e_1 = \frac{2n}{2n_T} = \frac{n}{n_T}$$

and

$$e_2 = \frac{2n}{\left(2n_T/((1-\lambda_{spec})\gamma + \lambda_{sens}(1-\gamma))\right)} = \frac{n}{\left(n_T/((1-\lambda_{spec})\gamma + \lambda_{sens}(1-\gamma))\right)}$$

They derived explicit formulas for calculating the above relative efficiencies, in the case of continuous outcome based on normal mixture, and in the binary case by using the Ury and Fleiss formula. In the continuous case, outcomes are also compared by using a two-sample Wilcoxon test, and in that nonparametric setting relative efficiencies are evaluated by Monte Carlo simulation. Online efficiency calculation for binary case is available at (http://linus.nci.nih.gov/brb/samplesize/td.html).

However, some statistical challenges related to the design of targeted clinical trials remain. For example, analytical expressions of relative efficiencies of targeted versus untargeted clinical trial designs for continuous outcomes are not trivial in the nonparametric and Bayesian settings. Furthermore, the conventional statistical challenges raised by genomics and microarrays (see Simon et al. 2003 and Sebastini et al. 2003) like experimental design, data quality, normalization, choice of data analysis method, correction of multiple hypotheses testing, validation of cluster, and classifier (see Simon et al. 2003 for a comprehensive synthesis) slow the progress of theranostics and subsequently that of targeted clinical trials and stratified medicine. The latter announces the advent of Personalized Medicine.

Statistics and Personalized Medicine

Personalized medicine (Langreth and Waldholz 1999) is in a restrictive and ideal sense, the determination of the right dose at the right time for the right patient or the evaluation of his predisposition to disease by using genomics-based technologies and his genomic makeup. More precisely, personalized medicine relies on patient polymorphic markers like single nucleotide polymorphisms (SNPs), variable number of tandem repeats (VNTR), short tandem repeats (STRs), and other mutations (Bentley 2004). Personalized medicine is sometimes mistaken as stratified medicine. In fact, stratified medicine is the precursor of personalized medicine.

Personalized medicine is opening huge opportunities for mathematical formalization sketched, for example, for molecular biology of DNA (Carbone and Gromov, 2001). Indeed, the upcoming era of personalized medicine coincides with the actual era of data (Donoho 2000) characterized by massive records of various individual data generated almost continuously. Individual i will thus be identified as a high-dimensional heterogeneous vector $(X_{i1}, \ldots, X_{im})$, where m is an integer, the $X_{ij}, j = 1, \ldots, m$; are deterministic or random qualitative and quantitative variables. The latter are for instance: biometric and genomic fingerprints, family records, age, gender, height, weight, diseases status, diet, medical images, personal medical history, family history, conventional prognostic profiles, and so on.

However, as personalized medicine will rely on huge technological infrastructures, it will generate a lot of data at the individual level. This will lead to enormous problems of:

- Correlation
- Multiple hypotheses testing
- Sensitivity and specificity of molecular diagnosis tools
- Choice of metrics for comparisons between individuals and between individuals and databases
- Integration of heterogeneous data and, subsequently, qualitative and quantitative standardization.

Acknowledgment

The author thanks Dr. Carmen Buchrieser, Senior Researcher at Pasteur Institute, for reviewing the manuscript.

About the Author

Dr. Aboubakar Maitournam held several positions at Pasteur Institute as postdoc and researcher at the interface of Genomics, Imaging, Bioinformatics, and Statistics. He contributed to the publication of *Listeria monocytogenes* genome and to the setting up of statistical methods for analysis of gene expression data at the Genopole of Pasteur Institute. His latest works at National Institute of Health, Biometric Research Branch (Bethesda, USA) focused on the statistical design of targeted clinical trials for the evaluation of molecularly targeted therapies. Currently Dr. Aboubakar Maitournam is a faculty member of department of Mathematics and Computer Sciences (University Abdou Moumouni of Niamey, Niger). Dr. Aboubakar Maitournam was also appointed as Director of Statistics (2008–2010) at the Ministry of Competitiveness and Struggle Against High Cost Life (Niger). Dr. Maitournam is a member of SPAS (Statistical Pan African Society). He contributes regularly to a Nigerien weekly newspaper called *Le Républicain* with papers for general public related to information era, statistical process control and competitiveness, genomics and statistics, mathematics, and society.

Cross References

►Clinical Trials: An Overview
►Clinical Trials: Some Aspects of Public Interest

►Medical Research, Statistics in
►Monte Carlo Methods in Statistics

References and Further Reading

Bentley DR (2004) Genomes for medicine. Nature 429:440–445
Betensky RA, Louis DN, Cairncross JG (2002) Influence of unrecognized molecular heterogeneity on randomized trials. J Clin Oncol 20(10):2495–2499
Carbone A, Gromov M (2001) Mathematical slices of molecular biology. La Gazette des Mathématiciens, Société Mathématique de France, special edition, 11–80
Donoho D (2000) High-dimensional data analysis: the curses and blessings of dimensionality. Aide-Mémoire, Stanford University
Langreth R, Waldholz M (1999) New era of personalized medicine-targeting drugs for each unique genetic profile. Oncologist 4:426–427
Maitournam A, Simon R (2004) On the efficiency of targeted clinical trials. Stat Med 24:329–339
Sebastini P, Gussoni E, Kohane IS, Ramoni MF (2003) Statistical challenges in functional genomics. Stat Sci 18(1):33–70
Simon R (2004) An agenda for clinical trials: clinical trials in the genomic era. Clin Trials 1:468–470
Simon R, Maitournam A (2004) Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 10: 6759–6763
Simon RM, Korn EI, McShane LM, Radmacher MD, Wright GW, Zhao Y (2003) Design and analysis of DNA microarray investigations. Springer, New York

Statistics, History of

Oscar Sheynin (assisted by Miodrag Lovric)
Berlin, Germany
Faculty of Economics, University of Kragujevac, Kragujevac, Serbia

Statistics: Origin of that Term

Many authors discussed this, notably Karl Pearson (1978). It is widely believed that the term statistics originated from the Latin *Status* (situation, condition) of population and economics; in late Latin, the same term meant State. Another root of the term comes from the Italian word *stato* (state), and a *statista* (a person who deals with affairs of state). According to Kendall (1960:447) the first use of the word statistics "occurs in a work by an Italian historian Girolamo Ghilini, who in 1589 refers to an account of *civile, politica, statistica e militare scienza*." In 1587 Giovanni Botero described the political structure of several states in his *Della ragione di stato* (English translation 1956) latinized as *De Disciplina status*. Humboldt (1815) wrote "political arithmetic (see Staatswissenschaft and Political Arithmetic) or, in latino-barbare (late Latin), statistics."

None of the above belonged to statistics or statisticians in the modern sense and the same is true for later sources: Shakespeare's *Hamlet* (1601), Helenus Politanus' (1672) *Microscopium statisticum*, and for Hermann Conring's lectures (from 1660, published 1673).

In English, the word *statist* appeared in Shakespeare's *Hamlet*, Act V, Scene 2 (c. 1601), and *Cymbeline*, Act II, Scene 4 (c. 1610), and the word *statistics* was first introduced into English in 1770 by W. Hooper in his translation of J. F. Von Bielfeld's *The elements of universal erudition, Containing an analytical argument of the sciences, polite arts, and belles letters* (3 vols): "The science, that is called *statistics, teaches us what is the political arrangement of all the modern states of the known world*." (vol 3, p 269). The word statistics was used again in this old sense in 1787 by E. A. W. Zimmermann in his book *A Political Survey of the Present State of Europe*. According to Karl Pearson (1978:2), John Sinclair was the first who had attached modern meaning to the word statistics in *The statistical account of Scotland drawn up from the communications of the ministers of the different parishes* (21 vols, 1791–1799).

Staatswissenschaft and Political Arithmetic

The Staatswissenschaft or University statistics was born in Germany in the mid-seventeenth century and a century later Achenwall established its Göttingen school which described various aspects of a given state, mostly without use of numbers. His successor Schlözer (1804:86) coined a pithy saying: *History is statistics flowing, and statistics is history standing still*. His followers adopted it as the definition of statistics (which did not involve studies of causes and effects).

Also during that time political arithmetic had appeared (Graunt, Petty). It widely used numbers and elementary stochastic considerations and discussed causes and relations, thus heralding the birth of statistics. Graunt (1662/1899) stated that it was necessary to know "how many people there be" of each sex, age, religion, trade, etc. (p. 396), provided appropriate estimates (sometimes quite wrongly), especially concerning ►medical statistics. He was able to use sketchy and unreliable statistical data for estimating the population of London and England as well as the influence of various diseases on mortality and attempted to discover regularities in the movement of population. Contradicting the prevailing opinion, he established that both sexes were approximately equally numerous and derived a rough estimate of the sex ratio

at birth (p. 389). Graunt also reasonably noted that mortality from syphilis was underestimated because of moral considerations (p. 356). Graunt doubted, however, that statistical investigations were needed for anyone except the King and his main ministers (p. 397).

He also compiled the first ever mortality table (p. 387); although rather faulty but of great methodological importance, it was applied by Jakob Bernoulli and Huygens.

One of the main subjects of political arithmetic was indeed population statistics, and it certainly confirmed that "In a multitude of people is the glory of a king, but without people a prince is ruined" (Proverbs 14:28). And here is another link between the Old Testament and that new discipline: Moses sent spies to the land of Canaan to find out "whether the people [there] are strong or weak, whether they are few or many, […] whether the land is rich or poor […]" (Numbers 13: 17–20).

Tabular statistics which appeared in the mid-eighteenth century could have served as a link between the two new disciplines, but its representatives were being scorned as "slaves of tables" (Knies 1850:23). However, in the 1680s Leibniz recommended to compile "statistical tables" with or without numbers and wrote several papers belonging to both those disciplines. They were first published in the nineteenth century, then reprinted (Leibniz 1986).

Numerical description of phenomena without studying causes and effects also came into being. The London Statistical Society established in 1834 declared that all conclusions "shall admit of mathematical demonstrations" (which was too difficult to achieve), and stipulated that statistics did not discuss causes and effects (which was impossible to enforce) (see Anonymous 1839). Louis (1825) described the *numerical method* which was actually applied previously. Its partisans (including D'Alembert) advocated compilation of numerical data on diseases, scarcely applied probability, and believed that theory was hardly needed.

A similar attitude had appeared in other natural sciences; the astronomer Proctor (1872) plotted 324 thousand stars on his charts wrongly stating that no underlying theory was necessary. Compilation of statistical yearbooks, star catalogues, etc., can be mentioned as positive examples of applying the same method, but they certainly demand preliminary discussion of data. Empiricism underlying the numerical method was also evident in the Biometric school (The Two Streams of Statistical Thought).

The Staatswissenschaft continued to exist, although in a narrower sense; climate, for example, fell away. At least in Germany it is still taught at universities, certainly includes numerical data, and studies causes and effects. It thus is partly the application of the statistical method to various disciplines and a given state. Chuprov's opinion (1909/1959:50, 1922:339) that the Staatswissenschaft will revive, although with an emphasis on numbers, and determine the essence of statistics was partly wrong: that science did not at all die, neither does it determine statistics.

Statistics and the Statistical Method: The Theory of Errors

Kolmogorov and Prokhorov 1982 defined mathematical statistics as a branch of mathematics devoted to systematizing, processing, and utilizing statistical data, i.e., the number of objects in some totality. Understandably, they excluded the collection of data and their exploratory analysis. The latter is an important stage of theoretical statistics which properly came into being in the mid-twentieth century. Debates about mathematical versus theoretical statistics can be resolved by stating that both data analysis and collection of data only belong to the latter and determine the difference between it and the former.

The first definition of the theory of statistics (which seems to be almost the same as theoretical statistics) worth citing is due to Butte (1808:XI): It is a science of understanding and estimating statistical data, their collection, and systematization. It is unclear whether Butte implied applications of statistics as well. Innumerable definitions of statistics (without any adjectives) had been offered beginning with Schlözer (Staatswissenschaft and Political Arithmetic), but the above suffices, and I only adduce the definition of its aims due to Gatterer (1775:15) which seems partly to describe both political arithmetic and the new Staatswissenschaft (Staatswissenschaft and Political Arithmetic): To understand the state of a nation by studying its previous states.

The statistical method is reasoning based on mathematical treatment of numerical data and the term is mostly applied to data of natural sciences. The method underwent two previous stages. During the first one, statements based on unrecorded general notions were made, witness an aphorism (Hippocrates 1952): Fat men are apt (!) to die earlier than others. Such statements express qualitative correlation quite conforming to the qualitative nature of ancient science.

The second stage was distinguished by the availability of statistical data (Graunt). The present, third stage began by the mid-nineteenth century when the first stochastic criteria for checking statistical inferences had appeared (Poisson, see Sheynin 1978, Sect. 5.2). True, those stages are not really separated one from another: even ancient astronomers had collected numerical observations.

Most important discoveries were made even without such criteria. Mortality from cholera experienced by those whose drinking water was purified was eight times lower than usual (Snow 1855:74–86) which explained the spread

of cholera. Likewise, smallpox vaccination (Jenner 1798) proved absolutely successful.

The theory of errors belongs to the statistical method. Its peculiar feature is the use of the "true value" of the constants sought. Fourier (1826/1890:533–534) defined it as the limit of the arithmetic mean of observations which is heuristically similar to the frequentist definition of probability and which means that residual systematic errors are included in that value.

From its birth in the second half of the eighteenth century (Simpson, Lambert who also coined that term (1765, Sect. 321)) to the 1920s it constituted the main field of application for the probability theory, and mathematical statistics borrowed its principles of maximal likelihood (Lambert 1760, Sect. 303) and least variance (Gauss 1823, Sect. 17) from it (from the theory of errors).

Gauss' first justification of the method of ▶least squares (1809) for adjusting "indirect observations" (of magnitudes serving as free terms in a system of redundant linear algebraic equations with unknowns sought and coefficients provided by the appropriate theory) was based on the (independently introduced) principle of maximum likelihood and on the assumption that the arithmetic mean of the "direct observations" was the best estimator of observations. He abandoned that approach and offered a second substantiation (1823), extremely difficult to examine, which rested on the choice of least variance. Kolmogorov (1946) noted in passing that it was possible to assume as the starting point minimal sample variance (whose formula Gauss had derived) – with the method of least squares following at once!

Gauss (1823, Sect. 2) stated that he only considered random errors. Quite a few authors had been favoring this second substantiation; best known is Markov (1899/1951:247) who (p. 246) nevertheless declared that the method of least squares was not optimal in any sense. On the contrary, in case of normally distributed errors it provides jointly efficient estimators (Petrov 1954).

One of the previous main methods for treating indirect observations was due to Boscovich (Cubranic 1961, 1962; Sheynin 1971) who participated in the measurement of a meridian arc. In a sense it led to the median. Already Kepler (Sheynin 2009, Sect. 1.2.4) indirectly considered the arithmetic mean "the letter of the law." When adjusting indirect observations, he likely applied elements of the minimax method (choosing a "solution" of a redundant system of equations that corresponded to the least maximal absolute residual free term) and of statistical simulation: He corrupted observations by small arbitrary "corrections" so that they conform to each other. Ancient astronomers regarded observations as their private property, did not report rejected results, and chose any reasonable estimate. Errors of observation were large, and it is now known that with "bad" distributions the arithmetic mean is not better (possibly worse) than a separate observation.

Al-Biruni, the Arab scholar (10*th*–11*th* cc.) who surpassed Ptolemy, did not yet keep to the arithmetic mean but chose various estimators as he saw fit (Sheynin 1992).

There also exists a determinate theory of errors which examines the entire process of measurement without applying stochastic reasoning and which is related to the exploratory data analysis and experimental design. Ancient astronomers selected optimal conditions for observation, when errors least influenced the end result (Aaboe and De Solla Price 1964). Bessel (1839) found out where should the two supports of a measuring bar be situated to ensure the least possible change of its length due to its weight. At least in the seventeenth century, natural scientists including Newton gave much thought to suchlike considerations. Daniel Bernoulli (1780) expressly distinguished random and systematic errors. Gauss and Bessel originated a new stage in experimental science by assuming that each instrument was faulty unless and until examined and adjusted.

Another example: the choice of the initial data. Some natural scientists of old mistakenly thought that heterogeneous material could be safely used. Thus, the English surgeon Simpson (1847–1848/1871:102) vainly studied mortality from amputations performed in many hospitals during 45 years. On the other hand, conclusions were sometimes formulated without any empirical support. William Herschel (1817/1912:579) indicated that the size of a star randomly chosen from many thousands of them will hardly differ much from their mean size. He did not know that stars enormously differed in size so that their mean size did not really exist and in any case nothing follows from ignorance: *Ex nihilo nihil*!

Jakob Bernoulli, De Moivre, Bayes: Chance and Design

The theory of probability emerged in the mid-seventeenth century (Pascal, Fermat) with an effective introduction of expectation of a random event. At first, it studied games of chance, then (Halley 1694) tables of mortality and insurance, and (Huygens 1699) problems in mortality. Halley's research, although classical, contained a dubious statement. Breslau, the city whose population he studied, had a yearly rate of mortality equal to 1/30, the same as in London, and yet he considered it as a statistical standard. If such a concept is at all appropriate, there should be standards of several levels.

Equally possible cases necessary for calculating chances (not yet probabilities) were lacking in those applications, and Jakob Bernoulli (1713, posthumously) proved

that posterior statistical chances of the occurrence of an event stochastically tended to the unknown prior chances. In addition, his law of large numbers (the term was due to Poisson) determined the rapidity of that process; Markov (1900/1924:44–52) improved Bernoulli's crude intermediate calculations and strengthened his estimate. Pearson (1925) achieved even better results, but only by applying the Stirling formula unknown to Bernoulli (as did Markov providing a parallel alternative improvement on pp 102–115). Pearson also unreasonably compared Bernoulli's estimate with the wrong Ptolemaic system of the world. He obviously did not appreciate theorems of existence (of the limiting property of statistical chances).

Statisticians never took notice of that rapidity, neither did they cite Bernoulli's law if not sure that the prior probability really existed and they barely recognized the benefits of the theory of probability (and hardly mentioned the more powerful forms of that law due to Poisson and Chebyshev). They did not know or forgot that mathematics as a science did not depend on the existence of its objects of study. The actual problem was to investigate whether the assumptions of the *Bernoulli trials* (their mutual independence and constancy of the probability of the studied event) were obeyed, and it was Lexis (The Two Streams of Statistical Thought) who formulated it. The previous statement of Cournot (1843; Sect. 86), whose outstanding book was not duly appreciated, that prior probability can be replaced by statistics in accord with *the Bernoulli's principle* was unnoticed.

The classical definition of probability, due to De Moivre (1738, Introduction) rather than to Laplace, with its equally possible cases is still with us. The axiomatic approach does not help statisticians and, moreover, practitioners have to issue from data, hence from the Mises frequentist theory developed in the 1930s which is not, however, recognized as a rigorous mathematical discovery.

Arbuthnot (1712) applied quite simple probability to prove that only Divine Providence explained why during 82 years more boys were invariably born in London than girls since the chances of a random occurrence of that fact were quite negligible. Cf. however the D'Alembert–Laplace problem: a long word is composed of printer's letters; was the composition random? Unlike D'Alembert, Laplace (1814/1995:9) decided that, although all the arrangements of the letters were equally unlikely, the word had a definite meaning, and therefore composed with an aim. His was a practical solution of a general and yet unsolved problem: to distinguish between a random and a determinate finite sequence of unities and zeros.

Arbuthnot could have noticed that Design was expressed by the binomial law, but it was still unknown. Even its introduction by Jakob Bernoulli and later scientists failed to become generally accepted: philosophers of the eighteenth century almost always only understood randomness in the "uniform" sense.

While extending Arbuthnot's study of the sex ratio at birth, De Moivre (1733) essentially strengthened the law of large numbers by proving the first version of the central limit theorem (see ►Central Limit Theorems) thus introducing the normal distribution, as it became called in the end of the nineteenth century. Laplace offered a somewhat better result, and Markov (1914/1951:511) called their proposition the *De Moivre–Laplace theorem*.

De Moivre devoted the first edition of his *Doctrine of Chances* (1718) to Newton, and there, in the Dedication, reprinted in 1756 (p. 329), we find his understanding of the aims of the new theory: separation of chance from Divine design, not yet the study of various and still unknown distributions, etc.

Such separations were being made in everyday life even in ancient India in cases of testimonies (Bühler 1886/1967:267). A misfortune encountered by a witness during a week after testifying was attributed to Divine punishment for perjury and to chance otherwise.

Newton himself (manuscript 1664–1666/1967:58–61) considered geometric probability and statistical estimation of the probability of various throws of an irregular die.

Bayes (1763), a memoir with a supplement published next year (Price and Bayes 1764), influenced statistics not less than Laplace. The so-called ►Bayes' theorem actually introduced by Laplace (1814/1995:10) was lacking there, but here is in essence his pertinent problem: a_i urns $(i = 1, 2)$ contain white and black balls in the ratio of α_i/β_i. A ball is extracted from a randomly chosen urn, determine the probability of its being white. The difficulty here is of a logical nature: may we assign a probability to an isolated event? This, however, is done, for example, when considering a throw of a coin. True, prior probabilities such as $\alpha_i/(\alpha_i + \beta_i)$ are rarely known, but we may keep to Laplace's principle (1803:xi): adopt a hypothesis and repeatedly correct it by new observations – if available!

Owing to these difficulties English and American statisticians for about 30 years had been abandoning the Bayes approach, but then (Cornfield 1967) the Bayes theorem *had returned from the cemetery*.

The main part of the Bayes memoir was his stochastic estimation of the unknown prior probability of the studied event as the number of *Bernoulli trials* increased. This is the inverse problem as compared with the investigations of Bernoulli and De Moivre, and H. E. Timerding, the Editor of the German translation of Bayes (1908), presented his result as a limit theorem. Bayes himself had not done it

for reasons concerned with rigor: unlike other mathematicians of his time (including De Moivre), he avoided the use of divergent series. Bayes' great discovery also needed by statisticians was never mentioned by them. Great, because it did not at all follow from previous findings and concluded the creation of the initial version of the theory of probability.

Both Bernoulli and De Moivre estimated the statistical probability given its theoretical counterpart and declared that they had at the same time solved the inverse problem (which Bayes expressly considered). Actually, the matter concerned the study of two different random variables with differing variances (a notion introduced by Gauss 1823), and only Bayes understood that the De Moivre formula did not ensure a good enough solution of the inverse problem.

Statistics in the Eighteenth Century

Later statisticians took up De Moivre's aim (Jakob Bernoulli, De Moirre, Bayes: Chance and Design) who actually extended Newton's idea of discovering the Divinely provided laws of nature. They, and especially Süssmilch, made the next logical step by attempting to discover the laws of the movement of population, hence to discern the pertinent Divine design. Euler essentially participated in compiling the most important chapter of the second edition, 1761–1762, of Süssmilch (1741), and Malthus (1798) picked up one of its conclusions, viz., that population increases in a geometric progression.

Süssmilch also initiated moral statistics by studying the number of marriages, of children born out of wedlock, etc. Its proper appearance was connected with A. M. Guerry and A. Quetelet (1830s and later).

Euler published a few elegant and methodically important memoirs on population statistics and introduced such concepts as increase in population and period of its doubling (see Euler 1923). Also methodically interesting were Lambert's studies of the same subject. When examining the number of children in families he (1772, Sect. 108) arbitrarily increased by a half their total number as given in his data likely allowing for stillbirths and mortality.

Most noteworthy were Daniel Bernoulli's investigations of several statistical subjects. His first memoir was devoted to inoculation (1766), to not a quite safe communication of a mild form of the deadly smallpox from one person to another (Jenner introduced vaccination of smallpox at the turn of that century) and proved that it lengthened mean life by two years plus and was thus highly beneficial (in the first place, for the nation). Then, he investigated the duration of marriages (1768), which was necessary for insurance depending on two lives. Finally, he (1770–1771) turned to the sex ratio at birth. He evidently wished to discover the *true value* of the ratio of male/female births (which does not really exist) but reasonably hesitated to make a final choice. However, he also derived the normal distribution although without mentioning De Moivre whose statistical work only became known on the Continent by the end of the nineteenth century.

Laplace (1812, Chapter 6) estimated the population of France by sampling (New Times: Great Progress and the Soviet cul-de-sac) and studied the sex ratio at birth. In this latter case he introduced *functions of very large numbers* (of births a and b) $x^a(1-x)^b$ and managed to integrate them. As usual, he had not given thought to thoroughly presenting his memoirs. While calculating the probability that male births will remain prevalent for the next 100 years, he did not add *under the same conditions of life*; and the final estimate of France's population was stated carelessly: Poisson, who published a review of that classic, mistakenly quoted another figure. Laplace's *Essai philosophique* (1814) turned general attention to probability and statistics.

The Theory of Probability and Statistics: Quetelet

Both Cournot (1843) and Poisson (1837) thought that mathematics should be the base of statistics. Poisson with coauthors (1835) were the first to state publicly that statistics was "the functioning mechanism of the calculus of probability" and had to do with mass observations. The most influential scholars of the time shared the first statement and likely the second as well. Fourier, in a letter to Quetelet (1869, t. 1, p 103) written around 1820, declared that statistics must be based on *mathematical theories*, and Cauchy (1845/1896:242) maintained that statistics provided means for judging doctrines and institutions and should be applied "avec tout la rigueur."

However, Poisson and Gavarret, his former student who became a physician and the author of the first book on medical statistics (1840), only thought about large numbers (e.g., when comparing two empirical frequencies) and a German physician Liebermeister (ca. 1877) complained that the alternative, i.e., the mathematical statistical approach was needed.

The relations between statistics and mathematics remained undecided. The German statistician Knapp (1872:116–117) declared that placing colored balls in Laplacean urns was not enough for shaking scientific statistics out of them. Much later mathematicians had apparently been attempting to achieve something of the

sort since Chuprov (1922:143) remarked that "Mathematicians playing statistics can only be overcome by mathematically armed statisticians." In the nineteenth, and the beginning of the twentieth century statisticians had still been lacking such armament.

Quetelet, who dominated statistics for several decades around the mid-nineteenth century, popularized the theory of probability. He tirelessly treated statistical data, attempted to standardize population statistics on an international scale, initiated anthropometry, declared that statistics ought to help foresee how various innovations will influence society, and collected and systematized meteorological data. Being a religious person, he (1846:259) denied any evolution of organisms which to some extent explains why Continental statisticians were far behind their English colleagues in studying biological problems. And Quetelet was careless in his writings so that Knapp (1872:124) stated that his spirit was rich in ideas but unmethodical and therefore un-philosophical. Thus, Quetelet (1836, t. 1, p 10) stated without due justification that the crime rate was constant although he reasonably but not quite expressly added: under invariable social conditions.

Quetelet paid attention to preliminary treatment of data and thus initiated elements of the exploratory data analysis (Statistics and the Statistical Method: The Theory of Errors); for example, he (1846:278) maintained that a too detailed subdivision of the material was a *charlatanisme scientifique*. He (1848:38) introduced the concept of Average man both in the impossible physical sense (e.g., mean stature and mean weight cannot coexist) and in the moral sphere, attributed to him mean inclinations to crime (1836, t. 2, p 171) and marriage (1848, p 77) and declared that that fictitious being was a specimen of mankind (1832, p 1).

Only in passing did he mention the Poisson law of large numbers, so that even his moral mean was hardly substantiated. Worse, he had not emphasized that the inclinations should not be attributed to individuals, and after his death German statisticians, without understanding the essence of the matter, ridiculed his innovations (and the theory of probability in general!) which brought about the downfall of *Queteletism*.

Fréchet (1949) replaced the Average man by *homme typique*, by an individual closest to the average. In any case, an average man (although not quite in Quetelet's sense) is meant when discussing per capita economic indications.

New Times: Great Progress and the Soviet cul-de-sac

In the main states of Europe and America statistical institutions and/or national statistical societies, which studied and developed population statistics, came into being during the first five decades of the nineteenth century. International statistical congresses aiming at unifying official statistical data had been held from 1851 onward, and in 1885 the still active International Statistical Institute was established instead.

A century earlier Condorcet initiated and later Laplace and Poisson developed the application of probability for studying the administration of justice. The French mathematician and mechanician Poinsot (1836) declared that calculus should not be applied to subjects permeated by imperfect knowledge, ignorance, and passions, and severe criticism was leveled at applications to jurisprudence for tacitly assuming independence of judges or jurors: "In law courts people behave like the*moutons de Panurge*" (Poincaré 1912:20). Better known is Mill's declaration (1843/1886:353): Such applications disgrace mathematics. Laplace (1812, Supplement of 1816/1886:523) only once and in passing mentioned that assumption.

However, stochastic reasoning can provide a "guideline" for determining the number of witnesses and jurors (Gauss, before 1841/1929:201–204) and the worth of majority verdicts. Poisson (1837:4) introduced the mean prior (statistically justified) probability of the defendant's guilt, not to be assigned to any individual and akin to Quetelet's inclination to crime. Statistical data was also certainly needed here. Quetelet (1836, t. 2, p 313) studied the rate of conviction as a function of the defendant's personality, noted that in Belgium the rate of conviction was considerably higher than in France (1833:18) and correctly explained this by the absence, in the former, of the institution of jurors (1846:334).

Statistical theory was also invariably involved in jurisprudence in connection with errors of the first and second kind. Thus (Sheynin 2009:17), the Talmud stipulated that a state of emergency (leading to losses) had to be declared in a town if a certain number of its inhabitants died during three consecutive days. Another example pertaining to ancient India is in Jakob Bernoulli, De Moirre, Bayes: Chance and Design.

A number of new disciplines belonging to natural science and essentially depending on statistics had appeared in the nineteenth century. *Stellar statistics* was initiated earlier by William Herschel (1784:162) who attempted to catalogue all the visible stars and thus to discover the form of our (finite, as he thought at the time) universe. In one section of the Milky Way he replaced counting by sample estimation (p. 158). He (1783) also estimated the parameters of the Sun's motion by attributing to it the common component of the proper motion of a number of stars. Galileo (1613) applied the same principle for estimating the

period of rotation of the Sun about its axis: he equated it with the (largely) common period of rotation of sunspots.

Most various statistical studies of the solar system (Cournot 1843) and the starry heaven (F. G. W. Struve, O. Struve, Newcomb) followed in the mid-nineteenth century and later (Kapteyn). Newcomb (Sheynin 2002) processed more than 62 thousand observations of the Sun and the planets and revised astronomical constants. His methods of treating observations were sometimes quite unusual. Hill and Elkin (1884:191) concluded that the "great Cosmical questions" concerned not particular stars, but rather their average parallaxes and the general relations between star parameters.

Daniel Bernoulli was meritorious as the pioneer of *epidemiology* (Statistics in the Eighteenth Century). It came into being in the nineteenth century mostly while studying cholera epidemics. The other new disciplines were *public hygiene* (the forerunner of ecology), *geography of plants, zoogeography, biometry, and climatology*.

Thus, in 1701 Halley published a chart of North Atlantic showing (contour) lines of equal magnetic declination, and Humboldt (1817) followed suit by inventing lines of equal mean yearly temperatures (isotherms) replacing thousands of observations and thus separating climatology from meteorology. These were splendid examples of exploratory data analysis (Statistics and the Statistical Method: The Theory of Errors). Also in meteorology, a shift occurred from studying mean values (Humboldt) to examining deviations from them, hence to temporal and spatial distributions of meteorological elements.

Statistics ensured the importance of public hygiene. Having this circumstance in mind, Farr (1885:148) declared that "Any deaths in a people exceeding 17 in 1,000 annually are unnatural deaths." Data pertaining to populations in hospitals (*hospitalism*, mortality due to bad hygienic conditions), barracks, and prisons were collected and studied, causes of excessive mortality indicated and measures for preventing it made obvious.

At least medicine had not submitted to statistics without opposition since many respected physicians did not understand its essence or role. A staunch supporter of "rational" statistics was Pirogov, a cofounder of modern surgery and founder of military surgery. He stressed the difficulty of collecting data under war conditions and reasonably interpreted them.

Around the mid-nineteenth century, statistics essentially fostered the introduction of anesthesia since that new procedure sometimes led to serious complications. Another important subject statistically studied was the notorious hospitalism, see above.

Biometry indirectly owed its origin to Darwin, witness the Editorial in the first issue of *Biometrika* in 1902: "The problem of evolution is a problem of statistics. […] Every idea of Darwin […] seems at once to fit itself to mathematical definition and to demand statistical analysis."

Extremely important was the recognition of the statistical laws of nature (theory of evolution, in spite of Darwin himself), kinetic theory of gases (Maxwell), and stellar astronomy (Kapteyn). And the discovery of the laws of heredity (Mendel 1866) would have been impossible without statistics. Methodologically these laws were based on the understanding that randomness in individual cases becomes regularity in mass (Kant, Laplace, and actually all the stochastic laws).

Laplace (1814; English translation 1995:2) declared that randomness was only occasioned by our failure to comprehend all the natural forces and by the imperfection of analysis, and he was time and time again thought only to recognize determinism. However, the causes he mentioned were sufficiently serious; he expressly formulated *statistical determinism* (e.g., stability of the relative number of dead letters, an example of transition from randomness to regularity); and his work in astronomy and theory of errors was based on the understanding of the action of random errors. It is also opportune to note here that randomness occurs in connection with unstable movement (Poincaré) and that a new phenomenon, chaotic behavior (an especially unpleasant version of instability of motion), was discovered several decades ago. Finally, Laplace was not original: Maupertuis (1756:300) and Boscovich (1758, Sect. 385) preceded him.

In the nineteenth century, but mostly perhaps in the twentieth, the statistical method penetrated many other sciences and disciplines beyond natural sciences so that it is now difficult to say whether any branch of knowledge can manage without it.

There are other points worth mentioning. *Correlation theory* continued to be denied even in 1916 (Markov), actually because it was not yet sufficiently developed. Its appearance (Galton, Pearson) was not achieved at once. In 1865–1866 the German astronomer and mathematician Seidel quantitatively estimated the dependence of the number of cases of typhoid fever on the level of subsoil water and precipitation but made no attempt to generalize his study. And in the 1870s several scientists connected some terrestrial phenomena with solar activity but without providing any such estimates.

According to Gauss (1823:18), for series of observations to be independent, it was necessary for them not to contain common measurements, and geodesists without referring to him have been intuitively keeping to his viewpoint.

For two series of about m observations each, n of them common to both, the measure of their interdependence was thought to be n/m. Kapteyn (1912) made the same proposal without mentioning anyone.

Estimation of precision was considered superfluous (Bortkiewicz 1894–1896, Bd 10, pp 353–354): it is a *luxury* as opposed to the statistical feeling. *Sampling* met with protracted opposition although even in 1812 the German statistician Lueder (Lueder 1812:9) complained about the appearance of "legions" of numbers. In a crude form, it existed long ago, witness the title of Stigler (1977). In the seventeenth century in large Russian estates it was applied for estimating the quantity of the harvested grain, and, early in the next century Marshal Vauban, the *French Petty*, made similar estimations for France as a whole.

No wonder that Laplace, in 1786, had estimated the population of France by sampling, and, much more important, calculated the ensuing error. True, Pearson (1928) discovered a logical inconsistency in his model. As a worthy method, sampling penetrated statistics at the turn of the nineteenth century (the Norwegian statistician Kiaer) and Kapteyn (1906) initiated the study of the starry heaven by stratified sampling, but opposition continued (Bortkiewicz 1901).

The *study of public opinion* and *statistical control of quality of industrial production*, also based on sampling, had to wait until the 1920s (true, Ostrogradsky (1848) proposed to check samples of goods supplied in batches), and *econometrics* was born even later, in the 1930s.

A curious side issue of statistics, *sociography*, emerged in the beginning of the twentieth century. It studies ethnic, religious, etc., subgroups of society, does not anymore belong solely to statistics, and seems not yet to be really scientific. And in sociology it became gradually understood that serious changes in the life of a society or a large commercial enterprise should be based on preliminary statistical studies.

Soviet statistics became a dangerous pseudoscience alienated from the world (Sheynin 1998). Its main goal was to preserve appearances by protecting Marxist dogmas from the pernicious influence of contemporary science and it frustrated any quantitative studies of economics and banished mathematics from statistics. In 1909, Lenin called Pearson a Machian and an enemy of materialism which was more than enough for Soviet statisticians to deny the work of the Biometric school lock, stock, and barrel.

Culmination of the success in that direction occurred in 1954, during a high-ranking conference in Moscow. Its participants even declared that statistics did not study mass random phenomena which, moreover, did not possess any special features. Kolmogorov, who was present at least for his own report, criticized Western statisticians for adopting unwarranted hypotheses…

Soviet statisticians invariably demanded that quantitative investigations be inseparably linked with the qualitative content of social life (read: subordinated to Marxism), but they never repeated such restrictions when discussing the statistical method as applied to natural sciences.

The Two Streams of Statistical Thought

Lexis (1879) proposed a distribution-free test for the equality of probabilities of the studied event in a series of observations, the ratio Q of the standard deviation of the frequency of the occurrence of the studied event, as calculated by the Gauss formula, to that peculiar to the ►binomial distribution. That ratio would have exceeded unity had the probability changed; been equal to unity otherwise, all this taking place if the trials were independent; and been less than unity for interdependent trials. Lexis (1879, Sect. 1) also qualitatively isolated several types of statistical series and attempted to define stationarity and trend.

Bortkiewicz initiated the study of the expectation of Q and in 1898 introduced his celebrated law of small numbers which actually only essentially popularized the barely remembered Poisson distribution. In general, his works remain insufficiently known because of his pedestrian manner, excessive attention to detail, and bad composition which he refused to improve. Winkler (1931:1030) quoted his letter (date not given) stating that he expected to have five readers (!) of his (unnamed) contribution.

Markov and mostly Chuprov (1918–1919) refuted the applicability of Q but anyway Lexis put into motion the Continental direction of statistics by attempting to base statistical investigations on a stochastic basis. Lexis was not, however, consistent: even in 1913 he held that the law of large numbers ought to be justified by empirical data. Poisson can be considered the godfather of the new direction.

On the other hand, the Biometric school with its leader Pearson was notorious for disregarding stochastic theory and thus for remaining empirical. Yet he developed the principles of correlation theory and contingency, introduced *Pearsonian* curves for describing asymmetrical distributions, devised the most important chi-square test (see ►Chi-Square Tests), and published many useful statistical tables. To a large extent his work ensured the birth of mathematical statistics.

Pearson successfully advocated the application of the new statistics in various branches of science and studied his own discipline in the context of general history (1978, posthumous). There (p 1) we find: "I do feel how wrongful

it was to work for so many years at statistics and neglect its history." He acquired many partisans and enemies (including Fisher). Here is Newcomb in a letter to Pearson of 1903 (Sheynin 2009, Sect. 10.9.4) and Hald (1998:651): "You are the one living author whose production I nearly always read when I have time […] and with whom I hold imaginary interviews […]"; "Between 1892 and 1911 [he] created his own kingdom of mathematical statistics and biometry in which he reigned supremely, defending its ever expanding frontiers against attacks."

Nevertheless, the work of his school was scorned by Continental scientists, especially Markov, the apostle of rigor. Chuprov, however, tirelessly, although without much success, strove to unite the two streams of statistical thought. Slutsky also perceived the importance of the Biometric school. He (1912) expounded its results and, although only in a letter to Markov of 1912, when he was not yet sufficiently known, remarked that Pearson's shortcomings will be overcome just as it happened with the nonrigorous mathematics of the seventeenth and eighteenth centuries.

Chuprov also achieved important results, discovering for example finite exchangeability (Seneta 1987). He mainly considered problems of the most general nature, hence inevitably derived unwieldy and too complicated formulas, and his contributions were barely studied. In addition, his system of notations was horrible. In one case he (1923:472) applied two-storey superscripts and, again, two-storey subscripts in the same formula!

Markov, the great mathematician, was to some extent a victim of his own rigidity. Even allowing for the horrible conditions in Russia from 1917 to his death in 1922, it seems strange that he failed, or did not wish to notice the new tide of opinion in statistics (and even in probability theory).

Mathematical Statistics

In what sense is mathematical statistics different from biometry? New subjects have been examined such as sequential analysis, the treatment of previously studied problems (sampling, time series, hypothesis testing) essentially developed, links with probability theory greatly strengthened (Pearson's empirical approach is not tolerated anymore). New concepts have also appeared and this seems to be a most important innovation. Fisher (1922) introduced statistical estimators with such properties as consistency, efficiency, etc., some of which go back to Gauss who had used and advocated the principle of unbiased minimum variance.

It is known that the development of mathematics has been invariably connected with its moving ever away from Nature (e.g., to imaginaries) and that the more abstract it was becoming, the more it benefited natural sciences. The transition from true values to estimating parameters was therefore a step in the right direction. Nevertheless, the former, being necessary for the theory of errors, are still being used in statistics, and even for objects not existing in Nature, see Wilks (1962, Sect. 10.1), also preceded by Gauss (1816, Sects. 3 and 4) in the theory of errors.

Rao (*Math. Rev.* 2005k:62007) noted that modern statistics has problems with choosing models, measuring uncertainty, testing hypotheses, and treating massive sets of data, and, in addition, that statisticians are not acquiring sufficient knowledge in any branch of natural science.

About the Author

Oscar Sheynin was born in Moscow, 1925. He graduated from the Moscow Geodetic Institute and Mathematical-Mechanical Faculty of Moscow State University, and he is Candidate of Sciences, Physics and Mathematics. He was working as a geodesist in the field, then taught mathematics, notably at the Plekhanov Institute for National Economy (Moscow) as Dozent. From 1962 to this day, he independently studies history of probability and statistics and since 1991 he has been living in Germany. He is a Member of International Statistical Institute (1975), Full Member, International Academy of History of Science (1995), and of the Royal Statistical Society. He has published more than 130 papers including 25 in the *Archive for History of Exact Sciences* and a joint paper on probability in the nineteenth century with Boris Gnedenko. Much more can be found at www.sheynin.de.

Cross References

►Astrostatistics
►Bayes' Theorem
►Foundations of Probability
►Laws of Large Numbers
►Least Squares
►Medical Statistics
►Normal Distribution, Univariate
►Poisson Distribution and Its Application in Statistics
►Probability, History of
►Sex Ratio at Birth
►Statistical Publications, History of

References and Further Reading

Aaboe A, De Solla Price DJ (1964) Qualitative measurements in antiquity. In: Mélanges A. Koyré, t. 1: L'aventure de la science. Hermann, Paris, pp 1–20

Anchersen JP (1741) Descriptio statuum cultiorum in tabulis. Otto Christoffer Wenzell, Copenhagen/Leipzig

Anonymous (1839) Introduction. J Stat Soc Lond 1:1–5
Arbuthnot J (1710/1712) An argument for Divine Providence taken from the constant regularity observed in the birth of both sexes. Philos Trans R Soc Lond (repr Kendall MG, Plackett RL (eds) (1997) Studies in the history of statistics and probability, vol 2. Griffin, High Wycombe, pp 30–34)
Bayes T (1763, published 1764) An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. Philos Trans R Soc Lond 53:370–418
Bayes T (1908) Versuch zur ösung eines problems der wahrscheinichkeitsrechnung. Herausgeber, Timeding HE. Ostwald Klassiker No. 169. Leipzig:Engelmann
Bernoulli D (1766) Essai d'une nouvelle analyse de la mortalité cause par la petite vérole etc. In: Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2, pp 235–267
Bernoulli D (1768) De duratione media matrimoniorum etc. In: Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2, pp 290–303; Sheynin O (2004) Probability and statistics. Russian Papers. Berlin, pp 17–31 (translated from Russian)
Bernoulli D (1770–1771) Mensura sortis ad fortuitam successionem rerum naturaliter contingentium applicata. In: Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2, pp 326–360
Bernoulli D (1780) Specimen philosophicum de compensationibus horologicis etc. In: Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2, pp 376–390
Bernoulli D (1982) Die Werke von Daniel Bernoulli, Bd 2. Basel
Bernoulli J (1713) Ars Conjectandi. Werke, Bd 3 (1975, Birkhäuser, Basel, pp 107–259); German trans: (1899) Wahrscheinlichkeitsrechnung (1999, Thun/Frankfurt am Main); English trans of pt 4: Bernoulli J (2005) On the law of large numbers. Berlin. Available at http://www.sheynin.de
Bessel FW (1839) Einfluß der Schwere auf die Figur eines . . . Stabes. In: Bessel FW (1876) Abhandlungen, Bd 3. Wilhelm Engelmann, Leipzig, pp 275–282
Bortkiewicz L (1894–1896) Kritische Betrachtungen zur theoretischen Statistik. Jahrbücher f. Nationalökonomie u. Statistik, 3. Folge, 8:641–680, 10:321–360, 11:701–705
Bortkiewicz L (1898) Das Gesetz der kleinen Zahlen. Leipzig
Bortkiewicz L (1904) Anwendung der Wahrscheinlichkeitsrechnung auf Statistik. Enc Math Wiss 1:821–851
Boscovich R (1758, in Latin/1966) Theory of natural philosophy. MIT Press, Cambridge. Translated from edition of 1763
Bühler G (ed) (1886) Laws of Manu. Clarendon Press, Oxford (repr 1967)
Butte W (1808) Die Statistik als Wissenschaft. Landshut
Cauchy AL (1845) Sur le secours que les sciences du calcul peuvent fournir aux sciences physiques ou même aux sciences morales. Oeuvr Compl 1 (1896), t. 9. Paris, pp 240–252
Chuprov (Tschuprow) AA (1909, in Russian) Essays on the theory of statistics. Sabashnikov, Saint Petersburg (repr State Publishing House, Moscow, 1959)
Chuprov (Tschuprow) AA (1918–1919) Zur Theorie der Stabilität statistischer Reihen. Skand Aktuarietidskrift 1:199–256; 2: 80–133
Chuprov (Tschuprow) AA (1922) Review of books. Nordisk Statistisk Tidskrift 1:139–160, 329–340
Chuprov (Tschuprow) AA (1923) On the mathematical expectation of the moments of frequency distributions in the case of correlated observations, Metron 2:461–493, 646–683
Cornfield J (1967) The Bayes theorem. Rev Inter Stat Inst 35:34–49
Cournot AA (1843) Exposition de la théorie des chances et des probabilités. Hachette, Paris (repr 1984)
Cubranic N (1961) Geodetski rad R. Boscovica. Zagreb
Cubranic N (1962) Geodätisches Werk R. Boscovic's. In: Actes Symp. Intern. Boscovic. Beograd, 1962, pp 169–174
De Moivre A (1718) Doctrine of chances. W. Pearson, London (2nd edn: 1738, 3rd edn: 1756; last edn repr Chelsea, New York, 1967)
De Moivre A (1733, in Latin) A method of approximating the sum of the terms of the binomial $(a + b)^n$ expanded into a series from whence are deduced some practical rules to estimate the degree of assent which is to be given to experiments. Translated by De Moivre A and inserted in his book (1738, 1756), pp 243–254 in 1756
Euler L (1923) Opera omnia 1, t. 7. Leipzig
Farr W (1885) Vital Statistics: A memorial volume of selections from the reports and writings of William Farr MD, DCL, CB, F.R.S.N. (N A Humphreys, ed.). Sanitary Institute of London, London. (Reprinted 1975 by Scarecrow Press, Metuchen, NJ.)
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc A 222:309–368
Fourier JBJ (1826) Sur les resultats moyen déduits d'un grand nombre d'observations. Oeuvr (1890), t. 2. Paris, pp 525–545
Fréchet M (1949) Réhabilitation de la notion statistique de l'homme moyen. In: Fréchet M (1955) Les mathématiques et les concret. Presses Universitaires de France, Paris, pp 317–341
Galilei G (1613, in Italian) History and demonstrations concerning sunspots etc. In: Galilei G (1957) Discoveries and opinions of Galilei. Garden City, pp 88–144
Gatterer JC (1775) Ideal einer allgemeinen Weltstatistik. Göttingen
Gauss CF (1809) Theoria Motus Corporum Coelestium in Sectionibus Conicis solem Ambientum. Perthes und Besser, Hamburg (English trans: Davis CH (1857) Theory of the motion of the heavenly bodies moving about the sun in conic sections. Little, Brown, Boston (repr Mineola, Dover, (2004))
Gauss CF (1816) Bestimmung der Genauigkeit der Beobachtungen (repr (1880) Carl Friedrich Gauss Werke 4, 109–117. Königliche Gesellschaft der Wissenschaften, Göttingen; English trans: David HA, Edwards AWF (2001) The determination of the accuracy of observations. In: Annotated readings in the history of statistics. Springer, New York, pp 41–50)
Gauss CF (1823) Theoria combinationis observationum erroribus minimis obnoxiae (repr (1880) Carl Friedrich Gauss Werke 4, 1–53. Königliche Gesellschaft der Wissenschaften, Göttingen; English trans: Stewart GW (1995) Theory of the combination of observations least subject to errors. SIAM, Philadelphia)
Gauss CF (1887) Abhandlungen zur Methode der kleinsten Quadrate (repr Vaduz, 1998), Berlin
Gauss CF (1929) Werke, Bd 12. Göttingen/Berlin
Gavarret J (1840) Principes généraux de statistique médicale. Paris
Graunt J (1662) Natural and political observations made upon the bills of mortality. In: Petty W (1899) Economic writings, vol 2, pp 317–435 with Graunt's additions of 1665 and 1676. The Writings were reprinted: Fairfield, 1986; London, 1997. Many other editions of Graunt, e.g., Baltimore, 1939
Hald A (1998) History of mathematical statistics from 1750 to 1930. New York
Halley E (1694) An Estimate of the degrees of mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw; with an attempt to ascertain the price of annuities upon lives Philosophical Transactions of the Royal Society of

London (17): 596–610 and 654–656. (Reprinted, edited with an introduction by Reid LJ, Baltimore, MD: The Johns Hopkins Press 1942)
Herschel W (1783) On the proper motion of the Sun. In: Herschel W (1912) Scientific Papers, vol 1. London, pp 108–130
Herschel W (1784) Account of some observations. In: Herschel W (1912) Scientific Papers, vol 1. London, pp 157–166
Herschel W (1817) Astronomical observations and experiments etc. In: Herschel W (1912) Scientific Papers, vol 2. London, pp 575–591
Hill D, Elkin WL (1884) Heliometer-determination of stellar parallax. Mem R Astron Soc 48, the whole pt 1
Hippocrates (1952) Aphorisms. In: Great books of the western world, vol 10. Encyclopaedia Britannica, Chicago, pp 131–144
Humboldt A (1815) Prolegomena. In: Bonpland A, Humboldt A, Kunth KS. Nova genera et species plantarum etc., vol 1. Russian trans: 1936, Paris
Humboldt A (1817) Des lignes isothermes. Mém Phys Chim Soc Arcueil 3:462–602
Huygens C (1699) Correspondence. Oeuvr Compl (1895), t. 14. La Haye
Jenner E (1798) An inquiry into the causes and effects of the variolae vaccinae, a disease discovered in some of the Western counties of England, particularly Gloucestershire, and known by the name of the cow pox. Sampson Low, London, for the author. In: The three original publications on vaccination against smallpox, vol XXXVIII, pt 4: the Harvard Classics (1909–1914). P.F. Collier, New York
Kapteyn JC (1906) Plan of selected areas. Groningen
Kapteyn JC (1912) Definition of the correlation-coefficient. Monthly Notices R Astron Soc 72:518–525
Kendall MG (1960) Studies in the history of probability and statistics. X. Where shall the history of statistics begin? Biometrika 47(3–4):447–449
Knapp GF (1872) Quetelet als Statistiker. Jahrbücher f. Nationalökonomie u. Statistik 18:89–124
Knies CGA (1850) Die Statistik als selbstständige Wissenschaft. Kassel
Kolmogorov AN (1946, in Russian) Justification of the method of least squares. Selected works (1992), vol 2. Kluwer, Dordrecht, pp 285–302
Kolmogorov AN, Prokhorov (1982 in Russian) Mathematical Statistics In: Vinogradov IM (ed) Soviet Matematicheskaya entsiklopediya (Encyclopaedia of Mathematics), vol 3, Moscow, 576–581
Lambert JH (1760) Photometria (in Latin). Augsburg. Cited statement omitted from German translation
Lambert JH (1765) Anmerkungen und Zusätze zur practischen Geometrie. In: Lambert JH. Beyträge, Tl. 1, pp 1–313
Lambert JH (1765–1772) Beyträge zum Gebrauche der Mathematik und deren Anwendung, Tl. 1–3. Berlin
Lambert JH (1772) Anmerkungen über die Sterblichkeit, Todtenlisten, Geburthen and Ehen. In: Lambert JH. Beyträge, Tl. 3, pp 476–569
Laplace PS (ca. 1803) Traité de Mécanique Céleste, t. 3. Oeuvr Compl (1878), t. 3. Paris. Translation by Bowditch N (1832) Celestial mechanics. New York, 1966
Laplace PS (1812) Théorie analytique des probabilités. Oeuvr Compl (1886), t. 7. Paris
Laplace PS (1814) Essai philosophique sur les probabilités. Oeuvr Compl (1886), t. 7, No. 1, separate paging (English trans: New York, 1995)
Leibniz GW (1986) Sämmtl. Schriften und Briefe, 4, Bd 3. Berlin
Lexis W (1879) Über die Theorie der Stabilität statistischer Reihen. Jahrbücher f. Nationalökonomie u. Statistik 32:60–98 (repr Lexis W (1903) Abhandlungen zur Theorie der Bevölkerungs- und Moralstatistik. Jena, pp 170–212)
Lexis W (1913) Review of book. Schmollers Jahrbuch f. Gesetzgebung, Verwaltung u. Volkswirtschaft im Deutschen Reich 37:2089–2092
Liebermeister C (ca. 1877) Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. Sammlung klinischer Vorträge No. 110 (Innere Med. No. 39), Leipzig, pp 935–962
Louis PCA (1825) Recherches anatomico-pathologiques sur la phtisie. Paris
Lueder AF (1812) Kritik der Statistik und Politik. Göttingen
Malthus TR (1798) Essay on the principle of population. Works (1986), vol 1. Pickering, London
Markov AA (1899, in Russian) On the law of large numbers and the method of least squares. In: Izbrannye Trudy (Selected works). Academy of Sciences, USSR, pp 231–251
Markov AA (1900, in Russian) Calculus of probability. Later editions: 1908, 1913 and posthumous, Moscow, 1924 (German trans: Leipzig/Berlin, 1912) Academy of Sciences, St. Peterburg
Markov AA (1914, in Russian) On Jakob Bernoulli's problem. In: Izbrannye Trudy (Selected works). Academy of Sciences, USSR, pp 511–521
Markov AA (1916, in Russian) On the coefficient of dispersion. In: Izbrannye Trudy (Selected works). Academy of Sciences, USSR, pp 523–535
Markov AA (1951) Izbrannye Trudy (Selected works). Academy of Sciences, USSR
Maupertuis PLM (1756) Sur la divination. Oeuvres, t. 2. Lyon, pp 298–306
Mendel JG (1866, in German) Experiments in plant hybridization. In: Bateson W (1909) Mendel's Principles of Heredity. Cambridge University Press, Cambridge, pp 317–361 (repr 1913)
Mill JS (1843) System of logic. London (repr 1886)
Newton I (1967) Mathematical papers, vol 1. Cambridge University Press, Cambridge
Ostrogradsky MV (1848) Sur une question des probabilités. Bull Cl Phys-Math Acad Imp Sci St Pétersb 6(21–22):321–346
Pearson K (1925) James Bernoulli's theorem. Biometrika 17:201–210
Pearson K (1928) On a method of ascertaining limits to the actual number of individuals etc. Biometrika 20A:149–174
Pearson K (1978) History of statistics in the 17th and 18th centuries against the changing background of intellectual, scientific, and religious thought. Pearson ES (ed) Lectures 1921–1933. Griffin, London
Petrov VV (1954, in Russian) On the method of least squares and its extreme properties. Uspekhi Matematich Nauk 1:41–62
Poinsot L (1836) A remark, in Poisson (1836) p 380 http://www.archive.org/stream/comptesrendusheb02acad#page/380/mode/2up
Poisson S-D (1836) Note sur la loi des grands nombres. C r Acad Sci 2:377–382 http://www.archive.org/stream/comptesrendusheb02acad#page/380/mode/2up
Poisson S-D (1837) Recherches sur la probabilité des jugements etc. Paris (repr Paris, 2003)

Poisson S-D, Dulong PL, Double (1835) Rapports: Recherches de Statistique sur l'affection calculeuse, par M. Le docteur Civiale. C r Acad Sci Paris 1:167–177 (Statistical research on conditions caused by calculi by Doctor Civiale translated for the Int J Epidemiol by Swaine Verdier A, 2001, 30:1246–1249)
Politanus H (1672) Microscopium statisticum quo status imperii Romano-Germanici cum primis extraordinarius, ad vivum repraesentatur
Price RA, Bayes T (1764, published 1765) A demonstration of the second rule in the essay towards the solution of a problem in the doctrine of chances. Published in the Philosophical Transactions, Vol. LIII. Communicated by the Rev. Mr. Richard Price, in a letter to Mr. John Canton, M. A. F. R. S. Philos Trans 54:296–325
Proctor RA (1872) On star-grouping. Proc R Instn Gr Brit 6:143–152
Quetelet A (1832) Recherches sur la loi de la croissance de l'homme. Mém Acad R Sci Lettre Beaux-Arts Belg 7, pp 32
Quetelet A (1833) Statistique des tribunaux de la Belgique. Bruxelles (Coauthor, Smits E)
Quetelet A (1836) Sur l'homme, tt. 1–2. Bruxelles
Quetelet A (1846) Lettres sur la théorie des probabilités. Bruxelles
Quetelet A (1848) Du système social. Paris
Quetelet A (1869) Physique sociale etc., tt. 1–2. Bruxelles (Bruxelles, 1997)
Schlözer AL (1804) Theorie der Statistik. Göttingen
Seidel L (1865) Über den Zusammenhang zwischen den Häufigkeit der Typhus-Erkrankungen und dem Stande des Grundwassers. Z Biol 1:221–236
Seidel L (1866) Vergleichung der Schwankung der Regenmengen mit den Schwankungen in der Häufigkeit des Typhus. Z Biol 2: 145–177
Seneta E (1987) Chuprov on finite exchangeability, expectation of ratios and measures of association. Hist Math 14:243–257
Sheynin O (1971) O dva neobjavljena spisa R. Boskovica iz teorije verovatnoce. Dijalektika 2(Godina 6):85–93
Sheynin O (1978) Poisson's work in probability. Arch Hist Ex Sci 18:245–300
Sheynin O (1992) Al-Biruni and the mathematical treatment of observations. Arabic Sci Philos 2:299–306
Sheynin O (1998) Statistics in the Soviet epoch. Jahrbücher f. Nationalökonomie u. Statistik 217:529–549
Sheynin O (1999) Statistics, definitions of. In: Kotz S (ed) Encyclopedia of statistical sciences, update vol 3. New York, pp 704–711 (repr 2nd edn (2006) of that encyclopedia, vol 12. Hoboken, pp 8128–8135)
Sheynin O (2009) Theory of probability. Historical essay. Berlin. Available at: http://www.sheynin.de and Google
Simpson JY (1847–1848) Anaesthesia. In: Simpson JY (1871) Works, vol 2. Adam and Charles Black, Edinburgh, pp 1–288
Slutsky EE (1912, in Russian) Theory of correlation etc. Kiev
Snow J (1855) On the mode of communication of cholera. Churchill, London (repr (1965) Snow on cholera. Hafner, New York, pp 1–139)
Stigler SM (1977) Eight centuries of sampling inspection: the trial of the pyx. J Am Stat Assoc 72:493–500
Süssmilch JP (1741) Göttliche Ordnung. Berlin. Many later editions
Wilks SS (1962) Mathematical statistics. Wiley, New York
Winckler W (1931) Ladislaus von Bortkiewicz. Schmollers Jahrbuch f. Gesetzgebung, Verwaltung u. Volkswirtschaft im Deutschen Reich 55:1025–1033
Yule GU (1905) The introduction of the words "statistics", "statistical" into the English language. J R Stat Soc 68:391–396

Statistics: An Overview

David Hand
Professor, President of the Royal Statistical Society (2008–2009, 2010)
Imperial College, London, UK

One can define statistics in various ways. My favorite definition is bipartite:

- *Statistics is both the science of uncertainty and the technology of extracting information from data.*

This definition captures the two aspects of the discipline: that it is about understanding (and indeed manipulating) chance, and also about collecting and analyzing data to enable us to understand the world around us. More specifically, of course, statistics can have different aims, including prediction and forecasting, classification, estimation, description, summarization, decision-making, and others.

Statistics has several roots, which merged to form the modern discipline. These include (1) the theory of probability, initially formalized around the middle of the seventeenth century in attempts to understand games of chance, and then put on a sound mathematical footing with Kolmogorov's axioms around 1930; (2) surveys of people for governmental administrative and economic purposes, as well as work aimed at constructing life tables (see ▶Life Table) for insurance purposes (see ▶Insurance, Statistics in); and (3) the development of arithmetic methods for coping with measurement errors in areas like astronomy and mechanics, by people such as Gauss, in the eighteenth and nineteenth centuries.

This diversity of the roots of statistics has been matched by the changing nature of discipline. This is illustrated by, for example, the papers which have appeared in the journal of the Royal Statistical Society (the journal was launched in 1838). In the earlier decades, there was a marked emphasis on social matters, which gradually gave way around the turn of the century, to more mathematical material. The first half of the twentieth century then saw the dramatic development of deep and powerful ideas of statistical inference, which continue to be refined to the present day. In more recent decades, however, the computer has had an equally profound impact on the discipline. Not only has this led to the development of entirely new classes of methods, it has also put powerful tools into the hands of statistically unsophisticated users – users who do not understand the deep mathematics underlying the tools. As might be expected, this can be a mixed blessing: powerful tools in hands which understand and know how to use them properly can be a tremendous asset, but those

same tools in hands which can misapply them may lead to misunderstandings.

Although the majority of statisticians are still initially trained in university mathematics departments (with statistics courses typically being part of a mathematics degree), statistics should not be regarded as a branch of mathematics – just as physics, engineering, surveying, and so on have a mathematical base but are not considered as branches of mathematics. Statistics also has a mathematical base, but modern statistics involves many other intrinsically non-mathematical ideas.

An illustration of this difference is given by the contrast between probability (properly considered as a branch of mathematics – based on an axiom system) and statistics (which is not axiomatic). Given a system or process which is producing data, probability theory tells us what the data will be like. If we repeatedly toss a fair coin, for example, probability theory tells us about the properties of the sequences of heads and tails we will observe. In contrast, given a set of data, statistics seeks to tell us about the properties of the system which generated the data. Since, of course, many different systems could typically have generated any given data set, statistics is fundamentally *inductive*, whereas probability is fundamentally *deductive*.

At its simplest level, statistics is used to describe or summarize data. A set of 1,000 numerical values can be summarized by their mean and dispersion – though whether this simple two-value summary will be adequate will depend on the purpose for which the summary is being made. At a much more sophisticated level, official statistics are used to describe the properties of the entire population and economy of a country: the distribution of ages, how many are unemployed, the Gross National Product, and so on. The effective governance of a country, management of a business, operation of an education system, running of a health service, and so on, all depend on accurate descriptive statistics, as well as on statistical extrapolations of how things are likely to change in the future.

Often, however, mere descriptions are not enough. Often the observed data are not the entire population, but are simply a sample from this population, and the aim is to infer something about the entire population. Indeed, often the "entire population" may not be well-defined; what, for example, would be the entire population of possible measurements of the speed of light in repeated experiments? In such cases, the aim is to use the observed sample of values as the basis for an estimate of the "true underlying" value (of the speed of light in this example).

A single "point" estimate is all very well, but we must recognize that if we had chosen a different sample of values we would probably have obtained a different estimate – there is uncertainty associated with our estimate. A point estimate can be complemented by indicating the range of this uncertainty: indicating how confident we can be that the true unobserved value lies in a specified interval of values. Basic rules of probability tell us that increasing the sample size allows us to narrow down this range of uncertainty (provided the sample is collected in a certain way), so that we can be as confident as we wish (or as we can afford) about the unknown true value.

Estimation is one aspect of statistics, but often one has more pointed questions. For example, one might be evaluating a new medicine, and want to test whether it is more effective than the current drug of choice. Or one might want to see how well the data support a particular theory – that the speed of light takes a certain specified value, for example. Since, in the first example, people respond differently, and, in the second, measurement error means that repeated observations will differ, the data will typically consist of several observations – a sample, as noted above – rather than just one. Statistical *tests*, based on the sample, are then used to evaluate the various theories. *Hypothesis testing* methods (Neyman-Pearson hypothesis tests) are used for comparing competing explanations for the data (that the proposed new medicine is more effective than or is as effective as the old one, for example). Such tests use probability theory to calculate the chance that some summary statistic of the data will take values in given ranges. If the observed value of the summary statistic is very unlikely under one hypothesis, but much more likely under the other, one feels justified in rejecting the former and accepting the latter. *Significance testing* methods (Fisherian tests) are used to see how well the observed data match a particular given hypothesis. If probability calculations show that one is very unlikely to obtain a value at least as extreme as the observed value of the summary statistic then this is taken as evidence against the hypothesis.

Such testing approaches are not uncontroversial. Intrinsic to them is the calculation of how often one would expect to obtain such results in repeated experiments, assuming that the data arose from a distribution specified by a given hypothesis. They are thus based on a particular interpretation of probability – the *frequentist* view. However, one might argue that hypothetical repeated experiments are all very well, but in reality we have just the one observed set of data, and we want to draw a conclusion using that one set. This leads to ►Bayesian statistics. Bayesian statistics is based on a different interpretation of probability – the *subjective* view. In this view, probability is regarded as having no external reality, but rather as a degree of belief. In particular, in the testing context, the different values of the parameters of the distribution producing the data are themselves assumed to take some

distribution. In this approach to inference, one then uses the data to refine one's beliefs about the likely form of the distribution of the parameters, and hence of the distribution from which the data were generated.

The *likelihood function* plays an important role in all schools of inference; it is defined as the probability of obtaining the observed data, viewed as a function of the parameters of the hypothesized distribution. The likelihood function is used in Bayesian inference to update one's initial beliefs about the distribution of the parameters. A further school of statistics, the *likelihood* school, focuses attention on the likelihood function, on the grounds that it is this which contains all the relevant information in the data. Comparative discussions of the various schools of inference, along with the various profound concepts involved, are given by Barnett (1999) and Cox (2006).

The choice of the term "Bayesian" to describe a particular school of inference is perhaps unfortunate: ►Bayes' theorem is accepted and used by all schools. The key distinguishing feature of Bayesian statistics is the subjective interpretation of probability and the interpretation of the parameters of the distributions as random variables themselves.

The differences between the various schools of inference have stimulated profound, and sometimes fierce debates. Increasingly, however, things seem to be moving towards a recognition that different approaches are suited to different questions. For example, one might distinguish between what information the data contain, what we should believe after having observed the data, and what action we should take after having observed the data.

Thus far I have been talking about data without mentioning how it was collected. But data collection is a key part of statistical science. Properly designed data collection strategies lead to faster, cheaper collection, and to more accurate results. Indeed, poorly designed data collection strategies can completely invalidate the conclusions. For example, an experiment to compare two medicines in which one purposively gave one treatment to the sicker patients is unlikely to allow one to decide which is the more effective treatment. Sub-disciplines of statistics such as *experimental design* and *survey sampling* are concerned with effective data collection strategies. Experimental design studies situations in which it is possible to manipulate the subject matter: one can choose which patient will get which treatment, one can control the temperature of a reaction vessel, etc. Survey design is concerned with situations involving observational data, in which one studies the population as it is, without being able to intervene: in a salary survey, for example, one simply records the salaries. Observational data are weaker in the sense that causality cannot be unambiguously established: with such data there is always the possibility that other factors have caused an observed correlation. With experimental data, on the other hand, one can ensure that the only difference between two groups is a controlled difference, so that this must be the cause of any observed outcome difference. Key notions in experimental design are control groups, so that like is being compared with like, and random assignment of subjects to different treatments. A key notion in survey sampling is the random selection of the sample to be analyzed. In both cases, ►randomization serves the dual roles of reducing the chance of biases which could arise (even subconsciously) if purposive selection were to be used (as in the example of giving one treatment to sicker patients), and permitting valid statistical inference.

Once the data set has been collected, one has to analyze it. There exist a huge number of statistical data analysis tools. A popular misconception is that one can think of these tools as constituting a toolbox, from which one chooses that tool which best matches the question one wishes to answer. This notion has probably been promoted by the advent of powerful and extensive software packages, such as SAS and SPSS, which have modules structured around particular analytic techniques. However, the notion is a misleading one: in fact, statistical techniques constitute a complex web of related ideas, with, for example, some being special cases of others, and others being variants applied to different kinds of data. Rather than a toolbox, it is better to think of statistics as a language, which enables one to construct a way to answer any particular scientific question. This perspective is illustrated by statistical languages such as Splus and R. Statistical tools are underwritten by complex and powerful theory, which ties them together in various ways. For example:

- We can compare two groups using a t-test.
- If we are uneasy about the t-test assumptions, we might use a nonparametric alternative, or perhaps a ►randomization test.
- The t-test can be generalized to deal with more than two groups, as in ►analysis of variance.
- And it can be generalized to deal with a continuous "independent" variable in regression.
- Analysis of variance and regression are each special cases of ►analysis of covariance.
- And all these are examples of linear models.
- Linear models can be extended by generalizing the assumed distributional forms, in ►generalized linear models.
- Analysis of variance itself can be generalized to the multivariate situation in multivariate analysis

of variance (see ►Multivariate Analysis of Variance (MANOVA)) and the general linear model (see ►General Linear Models).
- And linear discriminant analysis (see ►Discriminant Analysis: An Overview, and ►Discriminant Analysis: Issues and Problems) can be regarded as a special case of multivariate analysis of variance.
- Linear discriminant analysis is a special case of supervised classification, with other such tools being ►logistic regression, ►neural networks, support vector machines, recursive partitioning classifiers, and so on.
- And on and on.

There are some very important subdomains of statistics which have been the focus of vast amounts of work, because of the importance of the problems with which they deal. These include (but are certainly not limited to) areas such as time series analysis, supervised classification, nonparametric methods, latent variable models, neural networks, belief networks, and so on.

Certain important theoretical ideas pervade statistical thinking. I have already referred to the likelihood function as a central concept in inference. Another example is the concept of overfitting. When one seeks to model a sample of observations with a view to understanding the mechanism which gave rise to it, it is important to recognize that the sample is just that, a sample. A different sample would probably be rather different from the observed sample. What one is really seeking to do is capture the common underlying characteristics of the various possible samples, not the peculiar characteristics of the sample one happens to have drawn. Too close a fit of a model to the observed data risks capturing the idiosyncrasies of these data. There are various strategies for avoiding this, including smoothing a model, using a weaker model, averaging multiple models based on subsets of the data or random perturbations of it, adding a penalization term to the measure of goodness of fit of the model to the data so that overfitting is avoided, and others.

I have already noted how the discipline of statistics has evolved over the past two centuries. This evolution is continuing, driven by the advent of new application areas (e.g., ►bioinformatics, retail banking, etc.) and, perhaps especially, the computer. The impact of the computer is being felt in many ways. A significant one is the appearance of very large data sets – in all domains, from telecommunications, through banking and supermarket sales, to astronomy, genomics, and others. Such large data sets pose new challenges. These are not merely housekeeping ones of keeping track of the data, and of the time required to analyze them, but also new theoretical challenges. Closely related to the appearance of these very large data sets is the growth of interest in *streaming* data: data which simply keep on coming, like water from a hose. Again, such data sets are ubiquitous, and typically require real-time analysis.

The computer has also enabled significant advances through computer intensive methods, such as ►bootstrap methods and ►Markov chain Monte Carlo. Bootstrap methods approximate the relationship between a sample and a population in terms of the observed relationship between a subsample and the sample. They are a powerful idea, which can be used to explore properties of even very complex estimators and procedures. Markov chain Monte Carlo methods (see ►Markov Chain Monte Carlo) are simulation methods which have enabled the practical implementation of Bayesian approaches, which were otherwise stymied to a large extent by impractical mathematics.

Graphical displays have long been a familiar staple of statistics – on the principle that a picture is worth a thousand words, provided it is well-constructed. Computers have opened up the possibility of interactive dynamic graphics for exploring and displaying data. However, while some exciting illustrations exist, the promise has not yet been properly fulfilled – though this appears to be simply a matter of time.

Another important change driven by the computer has been the advent of other data analytic disciplines, such as machine learning, ►data mining, image processing, and pattern recognition (see ►Pattern Recognition, ►Aspects of and Statistical Pattern Recognition Principles). All of these have very considerable overlaps with statistics – to the extent that one might regard them as part of "greater statistics," to use John Chambers's phrase (Chambers 1993). Such disciplines have their own emphasis and flavor (e.g., data mining being concerned with large data sets, machine learning with an emphasis on algorithms rather than models, etc.) but it is futile to try to draw sharp distinctions between them and statistics.

From an external perspective, perhaps the single most striking thing about statistics is how pervasive it is. One cannot run a country effectively without measures of its social and economic characteristics, without knowing its needs and resources. One cannot run a corporation successfully without understanding its customer base, its manufacturing and service operations, and its workforce. One cannot develop new medicines without rigorous clinical trials. One cannot control epidemics without forecasting and extrapolation models. One cannot extract information from physics or chemistry experiments without proper statistical techniques for analyzing the resulting data. And so on and on. All of these require measurements, projections, and understanding based on statistical analysis. The fact is that the modern world is a very complex place. Statistical methods are vital tools for understanding

its complexity, grasping its subtleties, and coping with its ambiguities and uncertainties.

An excellent overview of statistics is given by Wasserman (2004), and a short introduction describing the power and fascination of the modern discipline is given by Hand (2008). Aspects of the modern discipline are set in context in Hand (2009).

About the Author

David Hand is Professor of Statistics at Imperial College, London. He previously held the Chair of Statistics at the Open University. Professor Hand is a Fellow of the Royal Statistical Society and of the British Academy, an Honorary Fellow of the Institute of Actuaries, and a Chartered Statistician. He is a past-president of the International Federation of Classification Societies, and was president of the Royal Statistical Society for the 2008–2009 term, and again in 2010. He is the second person to serve twice since Lord George Hamilton, in 1915. He was Joint Editor of the Journal of the Royal Statistical Society Series C, Applied Statistics (1989–1992). He is founding editor of *Statistics and Computing* (1991–2001). David Hand has received various awards and prizes for his research including, the Thomas L. Saaty Prize for Applied Advances in the Mathematical and Management Sciences (2001), the Royal Statistical Society's Guy Medal in Silver (2002), the IEEE International Conference on Data Mining award for Outstanding Contributions (2004) and a Royal Society Wolfson Research Merit Award (2006–2010). Professor Hand has (co-)authored over 300 papers and 26 books.

Cross References

►Agriculture, Statistics in
►Astrostatistics
►Banking, Statistics in
►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Statistics
►Bayesian Versus Frequentist Statistical Reasoning
►Bioinformatics
►Biopharmaceutical Research, Statistics in
►Biostatistics
►Business Statistics
►Careers in Statistics
►Chemometrics
►Components of Statistics
►Computational Statistics
►Confidence Interval
►Decision Theory: An Overview
►Demography
►Econometrics
►Economic Statistics
►Environmental Monitoring, Statistics Role in
►Estimation: An Overview
►Federal Statistics in the United States, Some Challenges
►Fraud in Statistics
►Industrial Statistics
►Information Theory and Statistics
►Insurance, Statistics in
►Marine Research, Statistics in
►Medical Statistics
►Misuse and Misunderstandings of Statistics
►National Account Statistics
►Philosophical Foundations of Statistics
►Prior Bayes: Rubin's View of Statistics
►Promoting, Fostering and Development of Statistics in Developing Countries
►Psychiatry, Statistics in
►Psychology, Statistics in
►Rise of Statistics in the Twenty First Century
►Role of Statistics
►Significance Testing: An Overview
►Social Statistics
►Sociology, Statistics in
►Sport, Statistics in
►Statistical Distributions: An Overview
►Statistical Fallacies: Misconceptions, and Myths
►Statistical Genetics
►Statistical Inference
►Statistical Inference in Ecology
►Statistical Inference: An Overview
►Statistical Methods in Epidemiology
►Statistical Modeling of Financial Markets
►Statistical Modelling in Market Research
►Statistical Quality Control
►Statistical Software: An Overview
►Statistics and Climate Change
►Statistics and Gambling
►Statistics and the Law
►Statistics Education
►Statistics, History of
►Statistics: Controversies in Practice
►Statistics: Nelder's view
►Tourism Statistics

References and Further Reading

Chambers JM (1993) Greater or lesser statistics: a choice for future research. Stat Comput 3:182–184

Hand DJ (2008) Statistics: a very short introduction. Oxford University Press, Oxford

Hand DJ (2009) Modern statistics: the myth and the magic (RSS Presidential Address). J R Stat Soc A 172:287–306

Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge

Barnett V (1999) Comparative statistical inference, 3rd edn. Wiley, Chichester
Wasserman L (2004) All of statistics: a concise course in statistical inference. Springer, New York

Statistics: Controversies in Practice

William Notz
Professor
The Ohio State University, Columbus, OH, USA

Controversies may arise when statistical methods are applied to real problems. The reasons vary, but some possible sources are (1) the user fails to appreciate the limitations of the methods and makes claims that are not justified, (2) the use of statistical methods is affected by non-statistical considerations, and (3) researchers disagree on the appropriate statistical methods to use. In what follows, we provide examples of controversies involving all these sources. The references allow readers to explore these examples in more detail. We hope that this article will help readers identify and assess controversies that they encounter in practice.

Example 1: Web Surveys

Using the Internet to conduct "Web surveys" is becoming increasingly popular. Web surveys allow one to collect large amounts of survey data at lower costs than traditional methods. Anyone can put survey questions on dedicated sites offering free services, thus large-scale data collection is available to almost every person with access to the Internet. Some argue that eventually Web surveys will replace traditional survey methods.

Web surveys are not easy to do well. Problems faced by those who conduct them include (1) participants may be self-selected, (2) certain members of the target population may be systematically underrepresented and (3) nonresponse. These problems are not unique to Web surveys, but how to overcome them in Web surveys is not always clear. For a more complete discussion, see Couper (2000).

Controversy arises because those who do Web surveys may make claims about their results that are not justified. The controversy can be seen in the Harris Poll Online. The Harris Poll Online has created an online research panel of over 6 million volunteers, consisting "of a diverse cross-section of people residing in the United States, as well as in over 200 countries around the world" (see www.harrispollonline.com/question.asp). When the Harris Poll Online conducts a survey, a probability sample is selected from the panel and statistical methods are used to weight the responses and provide assurance of accuracy and representativeness. As a result, the Harris Poll Online believes their results generalize to some well-defined larger population. But the panel members (and hence participants) are self-selected, and no weighting scheme can account for all the ways in which the panel is different from the target population.

Example 2: Accessibility of Data

Research in many disciplines involves the collection and analysis of data. In order to assess the validity of the research, it may be important for others to verify the quality of the data and its analysis. Scientific journals, as a rule, require that published experimental findings include enough information to allow other researchers to reproduce the results. But how much information is enough? Some argue that all data that form the basis for the conclusions in a research paper should be publicly available, or at least available to those who review the research for possible publication.

Controversy arises because of non-statistical considerations. Data collection can be time consuming and expensive. Researchers expect to use the data they collect as the basis for several research papers. They are reluctant to make it available to others until they have a chance to fully exploit the data themselves.

An example of this controversy occurred when mass spectrometry data from a sample of a fossilized femur of a *Tyrannosaurus rex* indicated that fragments of protein closely matched sequences of collagen, the most common protein found in bones, from birds (see Asara et al. 2007 and Schweitzer et al. 2007). This was the first molecular confirmation of the long-theorized relationship between dinosaurs and birds. Many researchers were skeptical of the results (see, for example, Pevzner et al. 2008). They questioned the quality of the data, the statistical analyses, and doubted that collagen could survive so long, even partially intact. Critics demanded that all the data be made publicly available. Eventually researchers posted all the spectra in an online database. Although there was evidence that some of the data may have been contaminated, a reanalysis (see Bern et al. 2009) supported the original findings.

Example 3: Placeboes in Surgery

Randomized, double-blind, placebo-controlled trials are the gold standard for evaluating new medical interventions and are routinely used to assess new medical therapies. However, only a small percentage of studies of surgery use randomized comparisons. Surgeons think their operations succeed, but even if the patients are helped, the placebo

effect may be responsible. To find out, one should conduct a proper experiment that includes a "sham surgery" to serve as a placebo. See Freeman et al. (1999) and Macklin (1999) for discussion of the use of placeboes in surgery trials.

The use of placeboes in surgery trials is controversial. Arguments against the use of placeboes include nonstatistical considerations. Placebo surgery always carries some risk, such as postoperative infection. A fundamental principle is that "the interests of the subject must always prevail." Even great future benefits cannot justify risks to subjects today unless those subjects receive some benefit. No doctor would do a sham surgery as ordinary therapy, because there is some risk. If we would not use it in medical practice, it is not ethical to use it in a clinical trial. Do these arguments outweigh the acknowledged benefits of a proper experiment?

Example 4: Hypothesis Testing in Psychology Research

Research studies in many fields rely on tests of significance. Custom may dictate that results should be significant at the 5% level in order to be published. Overreliance on statistical testing can lead to bad habits. One simply formulates a hypothesis, decides on a statistical test, and does the test. One may never look carefully at the data. The limitations of tests are so severe, the risks of misinterpretation so high, and bad habits so ingrained, that some critics in psychology have suggested significance tests be banned from professional journals in psychology.

Here the controversy involves the appropriate statistical method. To help resolve the controversy, the American Psychological Association appointed a Task Force on Statistical Inference. The Task Force did not want to ban tests. Its report (see Wilkinson 1999) discusses good statistical practice in general. Regarding hypothesis testing, the report states "It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p-value or, better still, a confidence interval. … Always provide some effect-size estimate when reporting a p-value." Although banning tests might eliminate some abuses, the committee thought there were enough counterexamples to justify forbearance.

About the Author

William Notz has served as Editor of *Technometrics* and Editor of the *Journal of Statistics Education*. He is a Fellow of the American Statistical Association. He is co-author (with David Moore) of the book, *Statistics Concepts and Controversies* (W.H. Freeman and Company 7th edition). He has served as Acting Chair of the Department of Statistics and Associate Dean of the College of Mathematical and Physical Sciences at the Ohio State University.

Cross References

►Clinical Trials: An Overview
►Effect Size
►Frequentist Hypothesis Testing: A Defense
►Internet Survey Methodology: Recent Trends and Developments
►Misuse of Statistics
►Null-Hypothesis Significance Testing: Misconceptions
►Psychology, Statistics in
►P-Values

References and Further Reading

Asara JM, Schweitzer MH, Freimark LM, Phillips M, Cantley LC (2007) Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. Science 316(5822):280–285

Bern M, Phinney BS, Goldberg D (2009) Reanalysis of Tyrannosaurus rex mass spectra. J Proteome Res, Article ASAP DOI: 10.1021/pr900349r, Publication Date (Web): July 15, 2009

Couper MP (2000) Web surveys: a review of issues and approaches. Public Opin Quart 64:464–494

Freeman TB, Vawter DE, Leaverton PE, Godbold JH, Hauser RA, Goetz CG, Olanow CW (1999) Use of placebo surgery in controlled trials of a cellular-based therapy for Parkinson's disease. New Engl J Med 341:988–992

Macklin R (1999) The ethical problems with sham surgery in clinical research. New Engl J Med 341:992–996

Pevzner PA, Kim S, Ng J (2008) Comment on Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. Science 321(5892):1040

Schweitzer MH, Suo Z, Avci R, Asara JM, Allen MA, Arce FT, Horner JR (2007) Analyses of soft tissue from *Tyrannosaurus rex* suggest the presence of protein. Science 316(5822):277–280

Wilkinson L, Task Force on Statistical Inference, American Psychological Association, Science Directorate, Washington, DC, US (1999). Statistical methods in psychology journals: guidelines and explanations. Am Psychol 54:594–604

Statistics: Nelder's View

John Nelder†
Formerly Visiting Professor
Imperial College, London, UK

Statistical Science is a Wonderful Subject

Many scientists in their training take a basic course in statistics, and from it most of them learn almost nothing

that will be useful to them in the practice of their science. In the wider world statistics has a bad name:

- There are lies, dams lies, and statistics
  You can prove anything with statistics
  and so on.

I give here a personal view of my subject, what its components are, and what can be done with it. It should have not have a bad name; rather it should be regarded as a wonderful subject in which there are many new discoveries to be made.

Statistical Science

"Statistics" is an unfortunate term, because it can refer both to data and methods used to analyze those data. I, therefore, propose to use the term "Statistical science." It embraces all the techniques that can be used to make sense of figures. In principle it can be useful in the analysis of data from any scientific experiment or survey. It is above all a scientifically useful activity. A good statistical analysis will reveal; it will not obscure. I shall use the term "statistician" as a short form of "statistical scientist."

The Components of Statistical Science

Mathematics

The statistician must know some mathematics. Certain components are vital; for example, matrix algebra and methods for describing the structures of data. Remember always that mathematics, or parts of it, are tools for the statistician in his work. One part of mathematics is special and will be described separately.

Probability Theory

The statistician's use of probability theory is primarily for the construction of statistical models. These involve the use of probability distributions of one or more random variables to describe the assumed random components in the data, that is those aspects of the data that can only be described by their mass behavior. In addition models include what are described as fixed effects, that is effects that are assumed to stay constant across different data sets. In their statistics course scientists are usually introduced to the idea of statistical significance. Many come to believe that the sole purpose of a statistical analysis is to show that a difference between the effects of two treatments applied in an experiment is significant. The statistician knows that the size of a significant difference depends both on the size of the effect, the size of the sample and the underlying variation in the measurements. It is of course important that an experiment should be big enough to show clearly the differences it is sought to measure. Why is there this mistaken stress on the idea of statistical significance? I believe that it is because it gives the lecturer an opportunity to prove some mathematical theorems from probability theory.

Very often the lecturer is only interested in probability theory, whereas the statistician's interests are much wider. It is very important to stress that statistical science is not the same as probability theory.

Statistical Inference

Here we reach what I believe to be the heart of statistical science, namely what inferences may be legitimately made from the data we are analyzing. The components are the data we have, past data on a similar topic, and a statistical model for describing the data (we hope). When we have data from a number of experiments we shall be looking for effects that are constant across these experiments, in other word looking for statistical sameness rather than statistical differences. If we can find such effects we have extended the scope of our inferences about the effects in question.

How do we come by the statistical model that drives our inferences?

Sometimes there is a standard model from past work that has stood the test of time, but quite often the statistician has to draw on his own experience to formulate a suitable model.

The inference problem then becomes "given this model, defined by a set of unknown parameters, which values of those parameters do the data point to?"

The basic idea here is that of ▶likelihood, first introduced by Fisher in the 1920s. A likelihood is not a probability, and so requires new methods for its manipulation, not covered by probability theory. Unlike random variables, which can be integrated over any part of their distribution, likelihoods can be compared only at different ordinates. Fisher introduced the idea of maximum likelihood for defining the most likely values of the parameters given the data. However it may be that the model is unsuitable for describing the data; then the inference will be false. The statistician can test this by defining a goodness-of-fit statistic and testing the statistic against its null distribution. Models can be extended by adding terms, deleting terms, or exchanging terms, or by replacing a linear term by a smooth curve driven by the data, etc.

The Experimental Cycle

Both experiments and surveys need to be designed, and the statistician can be helpful at this stage if he is knowledgeable. The need for design is still not known as well as it should be: remember that double-blind trials, now

widely used in medicine, took 30 years to become accepted! The next stage is the execution of an experiment or survey. Many things can go wrong at this stage, biases introduced and so on, and a good statistician will be aware of such things. It is a good thing if all statisticians in their training actually do at least one experiment themselves, so that they get first-hand experience of the difficulties an experimenter may encounter. After execution comes analysis, which is often of major concern to the statistician. Output from analysis will include estimates of the effects of interest to the experimenter, together with estimates of their uncertainty.

Experiments rarely stand on their own, so finally a stage of consolidation is required, where results from the current experiment are compared with previous experiments on the same topic. This is often called ►meta-analysis, though I prefer the older combination of information. This completes the experimental cycle except for writing-up of the results; then the whole cycle can start again.

The Status of Bayesian Statistics

Many statisticians espouse methods based on Bayes's theorem for the analysis of experiments. In this framework there are no fixed effects and every parameter in the model is assigned a prior probability distribution. Much has been written about making these prior distributions uninformative etc., and some Bayesians regard these as purely subjective assessments. Given data, there is still no way of checking these prior assumptions. Various theorems can be proved from the Bayesian specification, but in my view these have nothing to do with the problems of scientific inference. Indeed I regard the problem given by Bayes in his original paper as much better described by a two-stage likelihood, than by a prior probability.

The Statistician and His Clients

A statistician will usually be working with other scientists who have statistical problems in the analysis of their data. The statistician must establish a close working relation with those he is helping, and to do this it is essential to learn some of the scientist's jargon. In my first job I had to learn at least six different jargons. The statistician should encourage his clients to learn something of his own jargon, so that his methods are not thought of as being some kind of magic!

Conclusion

Statistical science has a wider scope than any other science, because the idea of inference is not subject-dependent.

Its scope is therefore huge and its processes are continually both challenging and interesting. Remember only that statistical science is not the same as probability theory; it is much wider and (I think) much more interesting.

About the Author

John Ashworth Nelder was born on 8 October 1924 in Dulverton, Somerset, England. In 1950, he was appointed Head of the statistics section at the National Vegetable Research Station at Wellesbourne. In 1968, John succeeded Frank Yates as Head of statistics at Rothamsted Experimental Station, Harpenden. At Rothamsted, he started a collaboration with Robert Wedderburn that resulted in a seminal paper on Generalized Linear Models that has enormous impact on statistical analysis. During his time at Rothamsted, he was appointed as a visiting professor at Imperial College London (1972), which led to his collaboration with Peter McCullagh in writing a book, *Generalized Linear Models* (Chapman and Hall, 2nd edition 1989). Since his retirement in 1984, he had continued as a visiting professor in the Department of Mathematics at Imperial College, London. John Nelder had received many honors for his statistical work. He was awarded the Guy Medal in Silver of the Royal Statistical Society in 1977, and elected a Fellow of the Royal Society in 1981. He had served as President both the International Biometrics Society (1978–1979) and the Royal Statistical Society (1985–1986). In 1981, the Universitè Paul Sabatier, Toulouse, granted him an Honorary Doctorate. He had published over a 100 papers. Professor Nelder is also known for his contribution to statistical computing through designing and directing the development of the statistical software packages Genstat and GLIM. Professor Nelder received the Royal Statistical Society's Guy Medal in Gold in 2005.

John Nelder died on 7th August 2010 in Luton & Dunstable Hospital, Luton, UK, where he was recovering from a fall. He had sent his contributed entry on September 17 2009, adding the following text to his email: "Here is a first draft. I hope it may be useful."

"John Nelder was one of the most influential statisticians of his generation, having an impact across the entire range of statistics, from data collection, through deep theory, to practical implementation of software." (David Hand "Professor John Nelder FRS – Obituary", Reporter, Imperial College, London, 27 August 2010.)

Cross References

►Bayesian Analysis or Evidence Based Statistics?
►Bayesian Versus Frequentist Statistical Reasoning

►Clinical Trials: An Overview
►Components of Statistics
►Effect Size
►Likelihood
►Medical Research, Statistics in
►Meta-Analysis
►Model Selection
►Prior Bayes: Rubin's View of Statistics
►Probability Theory: An Outline
►Statistical Consulting
►Statistical Design of Experiments (DOE)
►Statistical Inference
►Statistics: An Overview

Stem-and-Leaf Plot

Vesna Bucevska
Associate Professor, Faculty of Economics
Ss. Cyril and Methodious University, Skopje, Macedonia

A stem-and-leaf plot (or simply stemplot), was invented by John Tukey (The idea behind the stemplot can be traced back to the work of Arthur Bowley in the early 1900s.) in his paper "Some Graphic and Semigraphic Displays" in 1972. It is a valuable tool in exploratory data analysis, since it displays the relative density and shape of data. Therefore, it is used as an alternative to the histogram. In order to construct a stem-and-leaf plot the following steps have to be taken:

1. The data have to be sorted in ascending order.
2. The stem-and-leaf units have to be determined. This means that we have to define what will be the stems and what will be the leaves for observations of interest. Each stem can consist of any number of digits, but each leaf can have only a single digit.

Data are grouped according to their leading digits, called stems, which are placed on the left side of the vertical line, while on the right hand side of the vertical line in ascending order follow the final digits of each observation called leaves. We can illustrate the way to construct a stem-and-leaf plot using the following data set for number of customers per day in a shop:

5 12 8 20 14 16 17 23 27 22 22 25 31 34 42 39 44 53 44 50 62

First we have to sort data in ascending order:

5 8 12 14 16 17 20 22 22 23 25 27 31 34 39 42 44 44 50 53 62.

Let us decide that the stem unit is 10, and the leaf unit is 1. Thus, the stem-and-leaf has the following appearance:

Stem	Leaf
0 |	5 8
1 |	2 4 6 7
2 |	0 2 2 3 5 7
3 |	1 4 9
4 |	2 4 4
5 |	0 3
6 |	2

If a stem-and leaf is turned on its side, it looks like a histogram constructed from the digits of the data. It is important to list each stem even they do not have associated leaves. If a larger number of bins is desired then there may be two stems for each digit.

If some of the observations are not integers then these numbers have to be rounded. If there are some negative numbers in data set then a minus sign has to be put in front of the stem unit.

Typically in statistical software packages (like Minitab or Statgraphics) stem-and-leaf display is preceded by another column of numbers to the left of the plot. It represents depths, which give cumulative counts from the top and bottom of the table, stopping at the row that contains the median, and the number for this row is given in parentheses. Recalling the example given above, we obtain

	Stem	Leaf
2	0 |	5 8
4	1 |	2 4 6 7
(6)	2 |	0 2 2 3 5 7
9	3 |	1 4 9
6	4 |	2 4 4
3	5 |	0 3
1	6 |	2

Although the stem and leaf plot is very similar to histogram it has some advantages over it. First, it keeps data in

their original form and the values of each individual data can be recovered from the plot. Second, it can be easily constructed without using computer, especially when the data set we are dealing with is not very large (in a range from 15 to 150 data points). For very large data set the histogram is prefered.

Cross References

►Exploratory Data Analysis
►Sturges' and Scott's Rules

References and Further Reading

Becker WE, Harnett DK (1987) Business and economics statistics with computer applications. Addison-Wesley, Reading

Montgomery D (2005) Introduction to statistical quality control, 5th edn. Wiley, New York

Newbold P, Carlson WL, Thorne B (2007) Statistics for business and economics, 6th edn. Pearson, New Jersey

Tukey JW (1972) Some graphic and semigraphic displays. In: Bancroft TA (ed) Statistical papers in honor of George W. Snedecor. Iowa State University Press, Ames, pp 293–316

Step-Stress Accelerated Life Tests

Mohamed T. Madi
Professor of Statistics, Associate Dean
UAE University, Al-Ain, United Arab Emirates

Introduction

To ascertain the service life and reliability of a product, or to compare alternative manufacturing designs, life testing at normal conditions is clearly the most reliable method. Due to continual advances in engineering science and improvement in manufacturing designs, one often deals with products that are highly reliable with a substantially long life span. Electronic products and devices (e.g., toasters, washers, and electronic chips), for example, are expected to last over a period of time much longer than what laboratory testing would allow. In these situations, the standard life testing methods may require long and prohibitively expensive testing time in order to get enough failure data necessary to make inferences about its relationship with external stress variables.

In order to shorten the testing period, test units are subjected to conditions more severe than normal. Such accelerated life testing (ALT) results in shorter lives than would be observed under normal conditions. Commonly, each test unit is run to failure at a constant stress, then a model for the relationship between the life of the unit and the constant stress is fitted to the data. This relationship is then extrapolated to estimate the life distribution of the product and get the desired information on its performance under normal use. Stress factors can include humidity, temperature, vibration, voltage, load, or any other factor affecting the life of the units. For a recent account of work on accelerated testing and test plans, we refer the reader to Nelson (2005a, b).

When constant-stress testing is considered too lengthy, step-stress testing may be used to reduce the times to failure still further. Such testing involves starting a test unit at a specified low stress. If the unit does not fail in a specified time, then the stress on it is raised to a higher value and held for another specified time. The stress is repeatedly increased and held this way until failure occurs. The time in the step-stress pattern when a test unit fails is recorded as the data on that unit. Applications of this type of testing include metal fatigue under varying load in service, cryogenic cable insulation, and electronics applications to reveal failure modes (elephant testing), so they can be designed out of the product.

When more constraints on the length of a life test are present, some form of censoring is commonly adopted. If for example, removing unfailed items from the life test at prespecified times is adopted, we have type I censoring. Instead, if we terminate the life test at the time of a failed item and remove all remaining unfailed items from the test, we have type II censoring.

One advantage of step-stress accelerated life testing (SSALT) is that the experimenters need not start with a high stress that could be harsh for the product, hence avoiding excessive extrapolation of test results. The obvious drawback is that it requires stronger assumptions and more complex analysis, compared to constant-stress ALT.

The simplest form of SSALT is the partial ALT introduced by DeGroot and Goel (1979) and in which the products are first tested under use conditions for a period of time before the stress is increased and maintained at the higher level throughout the test. They modeled the effect of switching the stress from normal conditions stress to the single accelerated stress by multiplying the remaining lifetime of the item by some unknown factor $\alpha > 0$. They studied the issues of estimation and optimal design in the framework of Bayesian decision theory.

Another formulation of this type of ALT, called the cumulative exposure (CE) model, was proposed by Nelson (1980). It assumes that the remaining life of test units depends on the current cumulative fraction failed and current stress. Survivors will fail according to the cdf for that stress but starting at the previously accumulated fraction failed. Nelson (1980) and Miller and Nelson (1983) studied

maximum likelihood estimation (MLE) under this type of parametric model when the underlying distribution is taken to be the Weibull and exponential, respectively.

Bhattacharyya and Soejoeti (1988) proposed the tampered failure rate (TFR) model for SSALT. Their model assumes that a change in the stress has a multiplicative effect on the failure rate function over the remaining life. In the special setting of a two-step partially accelerated life test, and assuming that the initial distribution belongs to a two-parameter Weibull family, they studied MLE and derived the Fisher information matrix.

There are mainly two types of SSALTs: a simple SSALT where there is a single change of stress during the test and mutliple-step SSALT where change of the stress occurs more than once. Madi (1993) generalized the TFR model from the simple step-stress model to the multiple step-stress model.

Acceleration Models and Lifetime Distributions

Stress Functions

Unless a nonparametric approach is used (see Shaked and Singpurwalla (1983), McNichols and Padgett (1988), and Tyoskin and Krivolapov (1996)), an SSALT model (ALT model in general) consists of a theoretical life distribution whose parameters are functions of accelerating stress and unknown coefficients to be estimated from the test data. These simple relationships, called stress functions, are widely used in practice, and special cases include the Arrhenius, inverse power, and Eyring laws (see Nelson (1990)). For example, Nelson (1980) used the Weibull with parameters (α, β), as the lifetime distribution, where the scale parameter α depends on stress according to an inverse power law $\alpha(V) = (V_0/V)^p$.

Lifetime Distribution Under Step-stress Pattern

The Cumulative Exposure Model

The basic idea for this model, introduced by Nelson (1980), is to assume that the remaining life of specimens depends only on the current cumulative fraction failed and current stress, regardless of how the fraction accumulated. Specifically, if we let F_i denote the cumulative distribution function (cdf) of the time to failure under stress s_i, the cdf of the time to failure under a step-stress pattern, F_0, is obtained by considering that the lifetime t_{i-1}under s_{i-1} has an equivalent time u_i under s_i such that $F_{i-1}(t_{i-1}) = F_i(u_i)$. Then the model is built as follows:

We assume that the population cumulative fraction of specimens failing under stress s_1, in Step 1, is

$$F_0(t) = F_1(t),\ 0 \le t \le t_1$$

In Step 2, we write $F_2(u_1) = F_1(t_1)$ to obtain u_1 that is the time to failure that would have produced the population cumulative fraction failing under s_2. The population cumulative fraction of specimens failing in Step 2 by total time t is

$$F_0(t) = F_2(t - t_1 + u_1),\ t_1 \le t \le t_2$$

Similarly, in Step 3, the unit has survived Step 2 and we consider an equivalent time u_2 under s_3 such that

$$F_3(u_2) = F_2(t_2 - t_1 + u_1)$$

where $t_2 - t_1 + u_1$ is an equivalent time under s_2. Then we have

$$F_0(t) = F_3(t - t_2 + u_2),\ t_2 \le t \le t_3$$

In general, Step i has the equivalent start time u_{i-1} that is the solution of

$$F_i(u_{i-1}) = F_{i-1}(t_{i-1} - t_{i-2} + u_{i-2})$$

and

$$F_0(t) = F_i(t - t_{i-1} + u_{i-1}),\ t_{i-1} \le t \le t_i$$

Finally, the CE model can then be written as

$$F_0(t) = \begin{cases} F_1(t), & 0 \le t \le t_1 \\ F_2(t - t_1 + u_1), & t_1 \le t \le t_2 \\ F_3(t - t_2 + u_2), & t_2 \le t \le t_3 \\ \quad \cdots & \quad \cdots \\ \quad \cdots & \quad \cdots \\ F_i(t - t_{i-1} + u_{i-1}), & t_{i-1} \le t \le t_i \end{cases}$$

$u_0 = t_0 = 0$ and u_i is the solution of $F_{i+1}(u_i) = F_i(t_i - t_{i-1} + u_{i-1})$, for $i = 1, \ldots, m-1$.

If the stress function is taken to be the inverse power law and F_i is a ►Weibull distribution, then the cdf for the fraction of specimens failing by time t for the constant stress V_i is

$$F_i(t) = 1 - \exp[-\{t(V_i/V_0)^p\}^\beta],$$

and for $t_{i-1} \le t \le t_i$,

$$F_0(t) = 1 - \exp[-\{(t - t_{i-1} + u_{i-1})(V_i/V_0)^p\}^\beta].$$

The Tampered Failure Rate Model

Consider the experiment in which n units are simultaneously put on test at time $t_0 = 0$ to a stress setting x_1. Starting at time $t_2 > 0$, the surviving units are subjected to a higher stress level x_2 while in the time interval $[t_1, t_2)$. At time t_2, the stress is increased on the surviving units to x_3 over $[t_2, t_3)$ and so on until the kth and last time interval $[t_{k-1}, \infty)$, where the remaining units are subjected to x_k until they all fail. The TFR model assumes that the effect of changing the stress from x_{i-1} to x_i is to multiply the failure rate function by α_{i-1}. The resulting step-stress failure rate function is given by

$$\lambda^*(t) = \left(\prod_{i=0}^{j-1} \alpha_i\right) \lambda_1(t), \; t_{j-1} \le t \le t_j, \quad j = 1, \ldots, k$$

where $t_0 = 0, t_k = \infty$ and $\alpha_{-1} = \alpha_0 = 1$. The corresponding survival function is

$$\overline{F}^*(t) = \left(\prod_{i=0}^{j-1} \overline{F}(t_i)^{(1-\alpha_i)\prod_{l=-1}^{i-1}\alpha_l}\right) \overline{F}(t)^{\prod_{i=0}^{j-1}\alpha_i},$$
$$t_{j-1} \le t \le t_j, \quad j = 1, \ldots, k$$

Substituting the Weibull survival function with scale parameter θ and shape parameter β, $\overline{F}(t) = \exp[-(y/\theta)^\beta]$, $\overline{F}^*(t)$ becomes

$$\overline{F}^*(t) = \left(\prod_{i=0}^{j-1} \exp\left\{\left(\prod_{l=-1}^{i} \alpha_l\right)\left(\frac{t_i}{\theta}\right)^\beta - \left(\prod_{l=-1}^{i-1} \alpha_l\right)\left(\frac{t_i}{\theta}\right)^\beta\right\}\right) \times \exp\left\{-\left(\prod_{i=0}^{j-1}\alpha_i\right)\left(\frac{t}{\theta}\right)^\beta\right\}$$

Putting $\delta_j = \theta\left(\prod_{i=0}^{j-1}\alpha_i\right)^{-\beta^{-1}}$, we have

$$\overline{F}^*(t) = \left(\prod_{i=0}^{j-1} \exp\left\{(t_i/\delta_{i+1})^\beta - (t_i/\delta_i)^\beta\right\}\right) \times \exp\left\{-(t/\delta_j)^\beta\right\},$$

which can be rewritten as

$$\overline{F}^*(t) = \exp\left\{\sum_{i=0}^{j-1}\left((t_i/\delta_{i+1})^\beta - (t_i/\delta_i)^\beta\right)\right\} \times \exp\left\{-(t/\delta_j)^\beta\right\}, \quad t_{j-1} \le t \le t_j, \quad j = 1, \ldots, k$$

Inference

Different fitting methods can be used in the context of SSALT. They include maximum likelihood estimation, ►least squares, best linear unbiased, and graphical methods. MLE is used frequently because it is straightforward and yields approximate variances and confidence limits for the parameters and percentiles. The major drawback is the computational complexity. The estimators are rarely obtained in closed form and extensive iterative methods must be used to determine the MLE.

Recent inferential work based on maximum likelihood for the CE model under different censoring schemes include Gouno et al. (2004), Zhao and Elsayed (2005), Wu et al. (2006), Balakrishnan and Xie (2007a, b), and Balakrishnan and Han (2008). Madi (1993) considered the MLE for the multiple step-stress TFR model when the life distribution under constant stress is Weibull.

Optimal Designs

Different optimization criteria have been used to design SSALT plans. Most are based on the variance of the MLE of the parameter of interest (variance optimality) or the determinant of the Fisher information matrix (D-optimality). One question arising is is on the duration that items need to be exposed to each stress level.

For example, Miller and Nelson (1983) presented optimal design for simple SSALT under the assumption of an exponential distribution. Their optimization criterion is to minimize the asymptotic variance of the MLE of the mean at a specified design stress. This criterion leads to optimizing the levels of the first and the second test stresses and the time of stress change. Bai et al. (1989) extended their work to the case in which a prescribed censoring time is involved. Gouno et al. (2004) considered the multiple SSALT with equal duration steps τ and progressive type I censoring and addressed the problem of optimizing τ using variance optimality as well as D-optimality.

About the Author

Dr. Mohamed Madi is a Professor, Department of Statistics, and Associate Dean, College of Business and Economics, UAE University, United Arab Emirates. He was the Assistant Dean for Research and Director of the UAEU Research Affairs Unit for Internally Funded Projects (2005–2008). He has authored and coauthored more than 30 papers and one book. Professor Madi has received the College of Business and Economics 2008 Outstanding Senior Research Award. He is Associate editor for the *Journal of Statistical Theory & Applications*, USA, and the *Jordan Journal of Mathematics and Statistics*, Jordan.

Cross References

►Accelerated Lifetime Testing
►Censoring Methodology

►Degradation Models in Reliability and Survival Analysis
►Generalized Weibull Distributions
►Industrial Statistics
►Modeling Survival Data
►Ordered Statistical Data: Recent Developments
►Parametric and Nonparametric Reliability Analysis
►Significance Testing: An Overview
►Survival Data

References and Further Reading

Bai DS, Kim MS, Lee SH (1989) Optimum simple step-stress accelerated life tests with censoring. IEEE Trans Reliab 38:528–532
Balakrishnan N, Han D (2008) Exact inference for a simple step-stress model with competing risks for failure from exponential distribution under Type-II censoring. J Stat Plan Infer 138(12):4172–4186
Balakrishnan N, Xie Q (2007a) Exact inference for a simple step-stress model with Type-II hybrid censored data from the exponential distribution. J Stat Plan Infer 137(8):2543–2563
Balakrishnan N, Xie Q (2007b) Exact inference for a simple step-stress model with Type-I hybrid censored data from the exponential distribution. J Stat Plan Infer 137(11):3268–3290
Bhattacharyya GK, Soejoeti Z (1988) A tampered failure rate model for step-stress accelerated test. Commun Stat Theory Meth 18(5):1627–1643
DeGroot MH, Goel PK (1979) Bayesian estimation and optimal design in partially accelerated life testing. Nav Res Logist Q 26:223–235
Gouno E, Sen A, Balakrishnan N (2004) Optimal step-stress test under progressive Type-I censoring. IEEE Trans Reliab 53:383–393
Madi MT (1993) Multiple step-stress accelerated life test; the tampered failure rate model. Commun Stat Theory Meth 22(9):2631–2639
McNichols DT, Padgett WJ (1988) Inference for step-stress accelerated life tests under arbitrary right-censorship. J Stat Plan Infer 20(2):169–179
Miller R, Nelson W (1983) Optimum simple step-stress plans for accelerated life testing. IEEE Trans Reliab 32:59–65
Nelson W (1980) Accelerated life testing: step-stress models and data analysis. IEEE Trans Reliab 29:103–108
Nelson W (1990) Accelerated testing: statistical models, test, plans and data analyses. Wiley, New York
Nelson WB (2005a) A bibliography of accelerated test plans. IEEE Trans Reliab 54:194–197
Nelson WB (2005b) A bibliography of accelerated test plans. Part II. IEEE Trans Reliab 54:370–373
Shaked M, Singpurwalla ND (1983) Inference for step-stress accelerated life tests. J Stat Plan Infer 7(4):295–306
Tyoskin OI, Krivolapov SY (1996) Nonparametric model for step-stress accelerated life testing. IEEE Trans Reliab 45:346–350
Wu SJ, Lin YP, Chen YJ (2006) Planning step-stress life test with progressively type I group-censored exponential data. Stat Neerl 60:46–56
Zhao W, Elsayed EA (2005) A general accelerated life model for step-stress testing. IIE Trans 37:1059–1069

Stochastic Difference Equations and Applications

Alexandra Rodkina[2], Cónall Kelly[1,2]
[1]Professor and Head
University of the West Indies,
Mona Campus, Kingston, Jamaica
[2]University of the West Indies, Mona Campus, Kingston, Jamaica

A first-order difference equation of the form

$$x_{n+1} = F(n, x_n), \quad n \in \mathbb{N}, \tag{1}$$

may be used to describe phenomena that evolve in discrete time, where the size of the each generation is a function of that preceding. But the real world often refuses to conform to such a neat mathematical representation. Unpredictable effects can be included in the form of a sequence of random variables $\{\xi_n\}_{n\in\mathbb{N}}$, and the result is a *stochastic difference equation*:

$$X_{n+1} = F(n, X_n) + G(n, X_n)\xi_{n+1}, \quad n \in \mathbb{N}. \tag{2}$$

The solution of (2) is a discrete time stochastic process adapted to the natural filtration of $\{\xi_n\}_{n\in\mathbb{N}}$. Stochastic difference equations also arise as discretizations of ►stochastic differential equations, though their asymptotic properties can be harder to analyze. Although a thorough introduction to the theory of deterministic difference equations can be found in Elaydi (2005) (for example), no comparable text exists for their stochastic counterparts. Nonetheless the recent development of powerful analytic tools is driving research efforts forward, and our understanding of discrete stochastic dynamics is growing. This has implications both for the modeling of real-world phenomena that evolve in discrete time, and the analysis of numerical methods for stochastic differential equations. Both are discussed in this article.

Mathematical biology is a good place to look for real-world phenomena that evolve in discrete time (see Murray 2002). Certain species, for example periodic cicadas and fruit flies, reproduce in non-overlapping generations, and the change in biomass from one generation to the next may be represented as a stochastic difference equation of the form

$$X_{n+1} = X_n\left[N(X_n) + Q(X_n)\xi_{n+1}\right], \quad n \in \mathbb{N}. \tag{3}$$

Notice that the form of (3) guarantees the existence of an equilibrium solution at $X \equiv 0$, corresponding to absence of the species. The sequence of random variables $\{\xi_n\}_{n\in\mathbb{N}}$

captures random influences like disease and natural variability in fecundity between generations. In order to model predator-prey interaction, competition or mutualism, it is essential to have a good understanding of the role of the coefficient functions N and Q in the dynamics of systems of such equations. For example, an equilibrium solution displaying almost sure asymptotic stability indicates that a species is not viable in the long run, as its biomass will decay to an unsustainable level over time. In the stochastic context, *almost sure* means *with probability one* and is usually written *a.s.*

Theoretical tools for investigating the a.s. asymptotic stability of the equilibrium of the similar equation

$$X_{n+1} = X_n\left[1 + R(X_n) + Q(X_n)\xi_{n+1}\right], \quad n \in \mathbb{N}, \tag{4}$$

were developed in Appleby et al. (2009a), in the form of a semi-martingale convergence theorem and a discrete form of the Itô formula. It turns out that the relative speed of decay of R and Q close to equilibrium determines the a.s. asymptotic stability of the equilibrium. One consequence of this is that an unstable equilibrium in a deterministic system may be stabilized by an appropriate perturbation coefficient Q. In the special case where R and Q are polynomials, a more detailed description is possible. If the a.s. stability is a result of a dominant R then solutions decay at an exact power law rate, however if the system has been stabilized by a dominant Q no such rate is possible. Moreover, solutions can be shown to change sign a random (though finite) number of times, indicating that discrete equations with stabilizing noise may be inappropriate in the context of a population model: biomass is inherently nonnegative. The closely related question of the role played by R and Q in the oscillatory behavior of solutions of (4) was investigated in Appleby et al. (2010).

The influence of random perturbations can be hidden from any observer of a single trajectory. In Rodkina (2009) it was shown that when R and Q are polynomial, there exist solutions of (4) that, with arbitrarily high probability, converge to zero monotonically and inside a well-defined deterministic envelope. The fluctuations that ordinarily characterize the presence of random noise are absent. This phenomenon is impossible in continuous time, since solutions of stochastic differential equations have trajectories that are non-differentiable almost everywhere.

Stochastic difference equations also find applications in economic modeling. Consider a self regulating island economy in the tropics, and suppose one wishes to model the effects of the annual hurricane season on economic activity. The essential mechanism underlying dynamic equilibrium in an idealized model of such an economy can be represented by the equation

$$x_{n+1} = x_n + f(x_n), \quad n \in \mathbb{N}, \tag{5}$$

under appropriate conditions on f (see Appleby et al. (2008) for details).

The degree to which activity during a hurricane season influences such a model varies randomly from year to year, depending on the number and intensity of storm systems, and how close the centre of each storm passes to the island. These effects may be incorporated by adding the term $\sigma_n\xi_{n+1}$ at each iteration, where again $\{\xi_n\}_{n\in\mathbb{N}}$ is a sequence of independent random variables, and each σ_n represents intensity of seasonal activity. Notice that including a state-independent perturbation in the model destroys the equilibrium.

In Appleby et al. (2008) it was shown that, if (5) is globally asymptotically stable, the perturbed model will eventually return to the vicinity of the former equilibrium, provided the intensity of seasonal activity converges to zero sufficiently quickly. However, no matter how effective the self-regulatory property of the system, if the seasonal activity fades out more slowly than a critical rate, which depends on the "heaviness" of the tails of the distribution of each ξ_n, then the system will not return to the former equilibrium. Hence (in this model), even if seasonal activity lessens each year, the economy may be prevented from settling back to near-equilibrium if the storms that do occur tend to be extremely violent. For models which are only locally stable in the absence of perturbations, the potential exists for an external shock to push a fundamentally stable economic situation over into instability.

Stochastic difference equations arise in numerical analysis, since they are the end product of the discretization of a stochastic differential equation. Consider

$$dX(t) = f(X(t))dt + g(X(t))dB(t), \quad t \geq 0, \tag{6}$$

where B is a standard Brownian motion. In general, solutions of (6) cannot be written in closed form; to explore their properties we can try to simulate them on a computer. Since computers are finite-state machines we must discretize the time set of (6) with, for example, a one-step Euler-Maruyama numerical scheme on a uniform mesh. This yields the stochastic difference equation

$$X_{n+1} = X_n + hf(X_n) + \sqrt{h}g(X_n)\xi_{n+1}, \quad n \in \mathbb{N}, \tag{7}$$

where $\{\xi_n\}_{n\in\mathbb{N}}$ is a sequence of i.i.d. standard normal random variables, and h is the mesh size. A good discussion of numerical methods for stochastic differential equations may be found in Kloeden and Platen (1992).

But discretization can alter the very properties of (6) that we are trying to examine. For example, a geometric Brownian motion (see ►Brownian Motion and Diffusions) with positive initial value remains positive with probability one. However, the Euler-Maruyama discretization does not: discrete processes can jump across equilibrium given a sufficiently large input from the stochastic component. This is a concern as geometric Brownian motion is often used to model asset prices in financial markets, which (like biomass in the population model) are inherently nonnegative. However, the probability of positivity can be increased over a finite simulation interval by increasing the density of mesh-points.

Nonetheless, any practical simulation must be carried out with a fixed non-zero stepsize h, so it is also necessary to study the effect of discretization with fixed h on carefully chosen test equations with known dynamics. A linear stability analysis seeks to discover when the asymptotic stability of an equilibrium solution of the test equation is preserved after discretization. Direct analysis of solutions of the stochastic difference equation arising from the discretization is necessary. Since these solutions are stochastic processes, asymptotic stability may be defined in several ways, each of which speaks to a difference aspect of the process. For example, a.s. asymptotic stability is a property of almost all trajectories, whereas mean-square asymptotic stability is a property of the distribution.

The literature surrounding mean-square stability analysis of stochastic numerical methods is extensive. For example an analysis of the stochastic θ-method using a scalar geometric Brownian motion as test equation may be found in Higham (2000), with an extension to systems of two equations in Saito and Mitsui (2002), using a technique outlined in Kloeden and Platen (1992). By contrast, developments in a.s. asymptotic stability analysis are more recent: Rodkina and Schurz (2005) have investigated a.s. asymptotic stability for the θ-method applied to a scalar stochastic differential equation, and Higham et al. (2007) have shown that a.s. exponential asymptotic stability in systems of equations with linearly bounded coefficients can be recovered in a θ-discretisation for sufficiently small h. We anticipate an expansion of the literature in the coming years.

Finally, we comment that it is often possible to reproduce a specific continuous-time dynamic in a discrete stochastic process by through careful manipulation of the mesh, presenting two examples from the literature. First, a.s. oscillatory behavior in linear stochastic differential equations with a fading point delay has been reproduced in Appleby and Kelly (2006) using a pre-transformation of the differential equation and a mesh that contracts at the same rate as the delay function. Second, state-dependent meshes have been used to reproduce finite-time explosions in a discretization of (6) (see for example Dávila et al., 2005).

About the Authors

Professor Alexandra Rodkina is Head of the Department of Mathematics, University of the West Indies, Jamaica. She has authored and co-authored more than 200 papers and three books, and is a member of the Editorial Board of the *International Journal of Difference Equations*.

Dr. Cónall Kelly is a lecturer at the same department, and is the author of 13 papers. Together, Professor Rodkina and Dr. Kelly have published four papers and organized special sessions at three conferences.

Cross References

►Brownian Motion and Diffusions
►Statistical Aspects of Hurricane Modeling and Forecasting
►Stochastic Differential Equations

References and Further Reading

Appleby JAD, Kelly C (2006) Oscillation of solutions of a nonuniform discretisation of linear stochastic differential equations with vanishing delay. Dy Contin Discret Impuls Syst A 13B(suppl):535–550

Appleby JAD, Berkolaiko G, Rodkina A (2008) On local stability for a nonlinear difference equation with a non-hyperbolic equilibrium and fading stochastic perturbations. J Differ Equ Appl 14(9):923–951

Appleby JAD, Berkolaiko G, Rodkina A (2009a) Non-exponential stability and decay rates in nonlinear stochastic difference equations with unbounded noise. Stochastics 81(2):99–127

Appleby JAD, Kelly C, Mao X, Rodkina A (2010) On the local dynamics of polynomial difference equations with fading stochastic perturbations. Dy Contin Discret Impuls Syst A 17(3):401–430

Appleby JAD, Rodkina A, Schurz H (2010) Non-positivity and oscillations of solutions of nonlinear stochastic difference equations with state-dependent noise. J Differ Equ Appl 6(7):807–830

Dávila J, Bonder JF, Rossi JD, Groisman P, Sued M (2005) Numercial analysis of stochastic differential equations with explosions. Stoch Anal Appl 23(4):809–825

Elaydi S (2005) An introduction to difference equations, 3rd edn. Undergraduate Texts in Mathematics. Springer, New York

Hasminski RZ (1981) Stochastic stability of differential equations. Sijthoff and Noordhoff, Alpen aan den Rijn – Germantown, Md

Higham DJ (2000) Mean-square and asymptotic stability of the stochastic theta method. SIAM J Numer Anal 38:753–769

Higham DJ, Mao X, Yuan C (2007) Almost sure and moment exponential stability in the numerical simulation of stochastic differential equations. SIAM J Numer Anal 45:592–609

Kelly C, Rodkina A (2009) Constrained stability and instability of polynomial difference equations with state-dependent noise. Discret Contin Dyn Syst B 11(4):913–933

Kloeden PE, Platen E (1992) Numerical solution of stochastic differential equations. Springer, Berlin
Murray JD (2002) Mathematical biology 1: an introduction. Interdisciplinary applied mathematics, vol 17. Springer, New York
Rodkina A, Schurz H (2005) Almost sure asymptotic stability of drift implicit θ-methods for bilinear ordinary stochastic differential equation in R1. J Comput Appl Math 180:13–31
Saito Y, Mitsui T (2002) Mean-square stability of numerical schemes for stochastic differential systems. Vietnam J Math 30:551–560

Stochastic Differential Equations

Peter E. Kloeden
Professor
Institut fur Mathematik, Frankfurt, Germany

A scalar stochastic differential equation (SDE)

$$dX_t = f(t, X_t)\, dt + g(t, X_t)\, dW_t \tag{1}$$

involves a the Wiener process W_t, $t \geq 0$, which is one of the most fundamental ▶stochastic processes and is often called a Brownian motion (see ▶Brownian Motion and Diffusions). A Wiener process is a Gaussian process with $W_0 = 0$ with probability 1 and $\mathcal{N}(0, t-s)$-distributed increments $W_t - W_s$ for $0 \leq s < t$ where the increments $W_{t_2} - W_{t_1}$ and $W_{t_4} - W_{t_3}$ on non-overlapping intervals, (i.e., with $0 \leq t_1 < t_2 \leq t_3 < t_4$) are independent random variables. It follows from the Kolmogorov criterion that the sample paths of a Wiener process are continuous. However, they are nowhere differentiable.

Consequently, an SDE is not a differential equation at all, but only a symbolic representation for the stochastic integral equation

$$X_t = X_{t_0} + \int_{t_0}^{t} f(s, X_s)\, ds + \int_{t_0}^{t} g(s, X_s)\, dW_s,$$

where the first integral is a deterministic Riemann integral for each sample path. The second integral cannot be defined pathwise as a Riemann-Stieltjes integral because the sample paths of the Wiener process do not have even bounded variation on any bounded time interval, but requires a new type of stochastic integral. An Itô stochastic integral $\int_{t_0}^{T} f(t) dW_t$ is defined as the mean-square limit of sums of products of an integrand f evaluated at the left end point of each partition subinterval times $[t_n, t_{n+1}]$ the increment of the Wiener process, i.e.,

$$\int_{t_0}^{T} f(t) dW_t := \text{m.s.} - \lim_{N_\Delta \to \infty} \sum_{j=0}^{N_\Delta - 1} f(t_n)\ (W_{t_{n+1}} - W_{t_n}),$$

where $t_{n+1} - t_n = \Delta / N_\Delta$ for $n = 0, 1, \ldots, N_\Delta - 1$. The integrand function f may be random or even depend on the path of the Wiener process, but $f(t)$ should be independent of future increments of the Wiener process, i.e., $W_{t+h} - W_t$ for $h > 0$.

The Itô stochastic integral has the important properties (the second is called the Itô isometry) that

$$\mathbb{E}\left[\int_{t_0}^{T} f(t) dW_t\right] = 0,$$
$$\mathbb{E}\left[\left(\int_{t_0}^{T} f(t) dW_t\right)^2\right] = \int_{t_0}^{T} \mathbb{E}\left[f(t)^2\right] dt.$$

However, the solutions of Itô SDE satisfy a different chain rule to that in deterministic calculus, called the Itô formula, i.e.,

$$U(t, X_t) = U(t_0, X_{t_0}) + \int_{t_0}^{t} L^0 U(s, X_s)\, ds + \int_{t_0}^{t} L^1(s, X_s)\, dW_s,$$

where

$$L^0 U = \frac{\partial U}{\partial t} + f \frac{\partial U}{\partial x} + \frac{1}{2} g^2 \frac{\partial^2 U}{\partial x^2}, \qquad L^1 U = g \frac{\partial U}{\partial x}.$$

An immediate consequence is that the integration rules and tricks from deterministic calculus do not hold and different expressions result, e.g.,

$$\int_0^T W_s\, dW_s = \frac{1}{2} W_T^2 - \frac{1}{2} T.$$

There is another stochastic integral called the Stratonovich integral, for which the integrand function is evaluated at the mid-point of each partition subinterval rather than at the left end point. It is written with $\circ dW_t$ to distinguish it from the ▶Itô integral. A Stratonovich SDE is thus written

$$dX_t = f(t, X_t)\, dt + g(t, X_t) \circ dW_t.$$

Note that the Itô and Stratonovich versions of an SDE may have different solutions, e.g.,

$$dX_t = X_t\, dW_t \quad \Rightarrow X_t = X_0 e^{W_t - \frac{1}{2}t} \qquad \text{Itô}$$
$$dX_t = X_t \circ dW_t \Rightarrow X_t = X_0 e^{W_t} \qquad \text{Stratonovich}$$

However, the Itô SDE (1) has the same solutions as the Stratonovich SDE with the modified drift coefficient, i.e.,

$$dX_t = \underline{f}(t, X_t)\, dt + g(t, X_t) \circ dW_t, \qquad \underline{f} := f - \frac{1}{2} g \frac{\partial g}{\partial x}.$$

In particular, the Itô and Stratonovich versions of an SDE with additive noise, i.e., with g independent of x, are the same.

Stratonovich stochastic calculus has the same chain rule as deterministic calculus, which means that Stratonovich SDE can be solved with the same integration tricks as for ordinary differential equations. However, Stratonovich stochastic integrals do not satisfy the nice properties above for Itô stochastic integrals, nor does the Stratonovich SDE have the same direct connection with diffusion process theory as the Itô SDE, e.g., the coefficient of the Fokker-Planck equation correspond to those of the Itô SDE (1), i.e.,

$$\frac{\partial p}{\partial t} + f\frac{\partial}{\partial x} + \frac{1}{2}g^2\frac{\partial^2 p}{\partial x^2} = 0.$$

The Itô and Stratonovich stochastic calculi are both mathematically correct. Which one should be used is really a modeling issue, but once one has been chosen, the advantages of the other can be used through the above drift modification.

The situation for vector valued SDE and vector valued Wiener processes is similar. Details can be found in the given references.

About the Author

Peter Kloeden graduated with a B.A. (with First Class Honors) in Mathematics from Macquarie University in Sydney, Australia. He received his Ph.D. in Mathematics from the University of Queensland in 1975 under the supervision of Rudolf Vyborny. In 1995, he also received a Doctor of Science in Mathematics from the University of Queensland. After 20 years of teaching at various universities in Australia, he was appointed in 1997 to the Chair in Applied and Instrumental Mathematics at the Johann Wolfgang Goethe University in Frankfurt. Professor Kloeden received the 2006 W.T. and Idalia Reid Prize by the Society of Industrial and Applied Mathematics, USA, for his fundamental contributions to the theoretical and computational analysis of differential equations. In 2009 he was elected a Fellow of the Society of Industrial and Applied Mathematics. He is Associate editor of a number of journals including: *Journal of Nonlinear Analysis: Theory, Methods and Applications*, *Journal of Stochastic Analysis*, *SINUM*, *Discrete and Continuous Dynamical Systems – Series B*, *Nonlinear Dynamics and Systems Theory*, *Advances in Dynamical Systems and Applications (ADSA)*, *Stochastics and Dynamics*, *Journal of Stochastic Analysis and Applications*, *Advanced Nonlinear Studies* and *International Journal of Dynamical Systems and Differential Equations*.

Cross References

►Brownian Motion and Diffusions
►Gaussian Processes
►Itô Integral
►Numerical Methods for Stochastic Differential Equations
►Optimal Statistical Inference in Financial Engineering
►Sampling Problems for Stochastic Processes
►Stochastic Difference Equations and Applications
►Stochastic Modeling Analysis and Applications
►Stochastic Modeling, Recent Advances in
►Stochastic Processes
►Stochastic Processes: Classification

References and Further Reading

Kloeden PE, Platen E (1992) The numerical solution of stochastic differential equations. Springer, Berlin (3rd revised edition, 1999)

Øksendal B (2003) Stochastic differential equations. an introduction with applications. Springer, Berlin (6th edition, Corr. 4th printing, 2007)

Stochastic Global Optimization

Anatoly Zhigljavsky
Professor, Chair in Statistics
School of Mathematics, Cardiff University, Cardiff, UK

Stochastic global optimization methods are methods for solving a global optimization problem incorporating probabilistic (stochastic) elements, either in the problem data (the objective function, the constraints, etc.), or in the algorithm itself, or in both.

Global optimization is a very important part of applied mathematics and computer science. The importance of global optimization is primarily related to the applied areas such as engineering, computational chemistry, finance and medicine amongst many other fields. For the state of the art in the theory and methodology of global optimization we refer to the "Journal of Global Optimization" and two volumes of the "Handbook of Global Optimization" (Horst and Pardalos 1995; Pardalos and Romeijn 2002). If the objective function is given as a "black box" computer code, the optimization problem is especially difficult. Stochastic approaches can often deal with problems of this kind much easier and more efficiently than the deterministic algorithms.

The problem of global minimization. Consider a general minimization problem $f(x) \to \min_{x \in X}$ with objective

function $f(\cdot)$ and feasible region X. Let x^* be a global minimizer of $f(\cdot)$; that is, x^* is a point in X such that $f(x^*) = f_*$ where $f_* = \min_{x \in X} f(x)$. Global optimization problems are usually formulated so that the structure of the feasible region X is relatively simple; this can be done on the expense of increased complexity of the objective function.

A global minimization algorithm is a rule for constructing a sequence of points $x_1, x_2, \ldots$ in X such that the sequence of record values $y_{on} = \min_{i=1\ldots n} f(x_i)$ approaches the minimum f_* as n increases. In addition to approximating the minimal value f_*, one often needs to approximate at least one of the minimizers x_*.

Heuristics. Many stochastic optimization algorithms where randomness is involved have been proposed heuristically. Some of these algorithms are based on analogies with natural processes; the well-known examples are evolutionary algorithms (Glover and Kochenberger 2003) and simulated annealing (Van Laarhoven and Aarts 1987). Heuristic global optimization algorithms are very popular in applications, especially in discrete optimization problems. Unfortunately, there is a large gap between practical efficiency of stochastic global optimization algorithms and their theoretical rigor.

Stochastic assumptions about the objective function. In deterministic global optimization, Lipschitz-type conditions on the objective function are heavily exploited. Much research have been done in stochastic global optimization where stochastic assumptions about the objective function are used in a manner similar to how the Lipschitz condition is used in deterministic algorithms. A typical example of a stochastic assumption of this kind is the postulation that $f(\cdot)$ is a realization of a certain stochastic process. This part of stochastic optimization is well described in Zhigljavsky and Zilinskas (2008), Chap. 4 and will not be pursued in this article.

Global random search (GRS). The main research in stochastic global optimization deals with the so-called global random search (GRS) algorithms which involve random decisions in the process of choosing the observation points. A general GRS algorithm assumes that a sequence of random points $x_1, x_2, \ldots, x_n$ is generated where for each $j \geq 1$ the point x_j has some probability distribution P_j. For each $j \geq 2$, the distribution P_j may depend on the previous points $x_1, \ldots, x_{j-1}$ and on the results of the objective function evaluations at these points (the function evaluations may not be noise-free). The number of points n, $1 \leq n \leq \infty$ (the stopping rule) can be either deterministic or random and may depend on the results of function evaluation at the points $x_1, \ldots, x_n$.

Three important classes of GRS algorithms. In the algorithm which is often called 'pure random search' (PRS) all the distributions P_j are the same (that is, $P_j = P$ for all j) and the points x_j are independent. In Markovian algorithms the distribution P_j depends only on the previous point x_{j-1} and $f(x_{j-1})$, the objective function value at x_{j-1}. In the so-called population-based algorithms the distributions P_j are updated only after a certain number of points with previous distribution have been generated.

Attractive features of GRS. GRS algorithms are very popular in both theory and practice. Their popularity is owed to several attractive features that many global random search algorithms share: (a) the structure of GRS algorithms is usually simple; (b) these algorithms are often rather insensitive to the irregularity of the objective function behavior, to the shape of the feasible region, to the presence of noise in the objective function evaluations, and even to the growth of dimensionality; (c) it is very easy to construct GRS algorithms guaranteeing theoretical convergence.

Drawbacks of GRS. Firstly, the practical efficiency of the algorithms often depends on a number of parameters, but the problem of the choice of these parameters frequently has little relevance to the theoretical results concerning the convergence of the algorithms. Secondly, for many global random search algorithms an analysis on good parameter values is lacking or just impossible. Thirdly, the convergence rate can be painfully slow, see discussion below. Improving the convergence rate (or efficiency of the algorithms) is a problem that much research in the theory of global random search is devoted to.

Main principles of GRS. A very large number of specific global random search algorithms exist, but only a few main principles form their basis. These principles can be summarized as follows:

(1) Random sampling of points at which $f(\cdot)$ is evaluated,
(2) Random covering of the space,
(3) Combination with local optimization techniques,
(4) The use of different heuristics including cluster-analysis techniques to avoid clumping of points around a particular local minima,
(5) Markovian construction of algorithms,
(6) More frequent selection of new trial points in the vicinity of "good" previous points,
(7) Use of statistical inference, and
(8) Decrease of randomness in the selection rules for the trial points.

In constructing a particular global random search method, one usually incorporates several of these principles, see Zhigljavsky and Zilinskas 2008 where all these principles are carefully considered.

Convergence of GRS. To establish the convergence of a particular GRS algorithm, the classical Borel-Cantelli theorem (see ►Borel–Cantelli Lemma and Its Generalizations) is usually used. The corresponding result can be formulated as follows, see Zhigljavsky and Zilinskas 2008, Theorem 2.1. Assume that $X \subseteq \mathbb{R}^d$ with $0 < \mathrm{vol}(X) < \infty$ and $\sum_{j=1}^{\infty} \inf P_j(B(x,\varepsilon)) = \infty$ for all $x \in X$ and $\varepsilon > 0$, where $B(x,\varepsilon) = \{y \in X : \|y-x\|_2 \le \varepsilon\}$ and the infimum is taken over all possible locations of previous points $x_1, \ldots, x_{j-1}$ and the results of the objective function evaluations at these points. Then with probability one, the sequence of points $x_1, x_2, \ldots$ falls infinitely often into any fixed neighborhood of any global minimizer.

In practice, a very popular rule for selecting the sequence of probability measures P_j is $P_j = \alpha_j P_0 + (1-\alpha_j)Q_j$, where $0 \le \alpha_j \le 1$, P_0 is the uniform distribution on X and Q_j is an arbitrary probability measure on X. In this case, the corresponding GRS algorithm converges if $\sum_{j=1}^{\infty} \alpha_j = \infty$.

Rate of convergence of PRS. Assume $X \subseteq \mathbb{R}^d$ with $\mathrm{vol}(X) = 1$ and the points $x_1, x_2, \ldots, x_n$ are independent and have uniform distribution on X (that is, GRS algorithm is PRS). The rate of convergence of PRS to the minimizer x_* is the fastest possible (for the worst continuous objective function) among all GRS algorithms. To guarantee that PRS reaches the ε-neighborhood $B(x_*,\varepsilon)$ of a point x_* with probability at least $1-\gamma$, we need to perform at least $n_* = \left\lceil -\log(\gamma) \cdot \Gamma\left(\frac{d}{2}+1\right) / \left(\pi^{\frac{d}{2}} \varepsilon^d\right) \right\rceil$ iterations, where $\Gamma(\cdot)$ is the Gamma-function. This may be a very large number even for reasonable values of d, ε and γ. For example, if $d = 10$ and $\varepsilon = \gamma = 0.1$ then $n_* \simeq 0.9 \cdot 10^{10}$. See Sect. 2.2.2 in Zhigljavsky and Zilinskas (2008) for an extensive discussion on convergence and convergence rates of PRS and other GRS algorithms.

Markovian GRS algorithms. In a Markovian GRS algorithm, the distribution P_j depends only on the previous point x_{j-1} and its function value $f(x_{j-1})$; that is, the sequence of points $x_1, x_2, \ldots$ constitutes a Markov chain (see ►Markov Chains). The most known Markovian GRS algorithms are the simulated annealing methods (Van Laarhoven and Aarts 1987). If a particular simulated annealing method creates a time-homogeneous Markov chain then the corresponding stationary distribution of this Markov chain is called Gibbs distribution. Parameters of the simulated annealing can be chosen so that the related Gibbs distribution is concentrated in a narrow neighborhood of the global minimizer x_*. The convergence to the Gibbs distribution can be very slow resulting in a slow convergence of the corresponding simulated annealing algorithm. The convergence of all Markovian GRS algorithms is generally slow as the information about the objective function obtained during the search process is used ineffectively.

Population-based methods. Population-based methods are very popular in practice (Glover and Kochenberger 2003). These methods generalize the Markovian GRS algorithms in the following way: rather than to allow the distribution P_j of the next point x_j to depend on the previous point x_{j-1}, it is now the distribution of a population of points (descendants, or children) depends on the previous population of points (parents) and the objective function values at these points. There are many heuristic arguments associated with these methods (Glover and Kochenberger 2003). There are also various probabilistic models of the population-based algorithms (Zhigljavsky 1991).

Statistical inference in GRS. The use of statistical procedures can significantly accelerate the convergence of GRS algorithms. Statistical procedures can be especially useful for defining the stopping rules and the population sizes in the population-based algorithms. These statistical procedures are based on the use of the asymptotic theory of extreme order statistics and the related theory of record moments. As an example, consider PRS and the corresponding sample $S = \{f(x_j), j = 1, \ldots, n\}$. This is an independent sample of values from the distribution with c.d.f. $F(t) = \int_{f(x) \le t} P(dx)$ and the support $[f_*, f^*]$, where $f^* = \sup_{x \in X} f(x)$. It can be shown that under mild conditions on f and P, this distribution belongs to the domain of attraction of the ►Weibull distribution, one of the ►extreme value distributions. Based on this fact, one can construct efficient statistical procedures for f_* using several minimal order statistics from the sample S.

For the theory, methodology and the use of probabilistic models and statistical inference in GRS, we refer to Zhigljavsky and Zilinskas (2008) and Zhigljavsky (1991).

Cross References

►Borel–Cantelli Lemma and Its Generalizations
►Markov Chains
►Weibull Distribution

About the Author

Professor Zhigljavsky is Director of the Center for Optimization and Its Applications at Cardiff University.

References and Further Reading

Glover F, Kochenberger GA (2003) Handbook on metaheuristics. Kluwer Academic, Dordrecht
Horst R, Pardalos P (eds) (1995) Handbook of global optimization. Kluwer Academic, Dordrecht
Pardalos P, Romeijn E (eds) (2002) Handbook of global optimization, vol 2. Kluwer Academic, Dordrecht

Van Laarhoven PJM, Aarts EHL (1987) Simulated annealing: theory and applications. D. Reidel, Dordrecht
Zhigljavsky A, Zilinskas A (2008) Stochastic global optimization. Springer, New York
Zhigljavsky A (1991) Theory of global random search. Kluwer Academic, Dordrecht

Stochastic Modeling, Recent Advances in

Christos H. Skiadas
Professor, Director of the Data Analysis and Forecasting Laboratory
Technical University of Crete, Chania, Greece

The term Stochastic Modeling is related to the theory and applications of probability in the modeling of phenomena in real life applications. Stochastic is a term coming from the ancient Greek period and is related to "*stochastes*" (people who are philosophers or intellectuals, scientists in recent notation) and "*stochazomai*" (I am involved in highly theoretical and intellectual issues as are philosophy and science).

The term model accounts for the representation of the reality (a real situation) by a verbal, logical or mathematical form. It is clear that the model includes a part of the main characteristics of the real situation. As far as the real situation is better explained the model will be termed as successful or not.

The science or even the art to construct and apply a model to real situations is termed as modeling. It includes model building and model adaptation, application to specific data and even simulation; that is producing a realization of a real situation.

It is clear that it is essential to organise and apply a good method or even process of collecting, restoring, classifying, organising and fitting data related to the specific case; that is to develop the "data analysis" scientific field.

Modeling is related to the use of past data to express the future developments of a real system. To this end modeling accounts for two major intellectual and scientific schools; the school of determinism and the school of probabilistic or stochastic modeling.

Deterministic modeling is related to determinism; that is the expression of the reality with a modeling approach that uses the data from the past and could lead to a good and even precise determination of the future paths of a natural system. Determinism was a school of thought that was the basis of very many developments in various scientific fields last centuries. Deterministic models of innovation diffusion appear in (Skiadas 1985, 1986, 1987).

From the other part, it was clear from the very beginning even from the rising of philosophy and science from the ancient Greek period that the future was unpredictable (probabilistic) or even chaotic. However, the successful solutions of several problems last centuries, especially in physics, straighten determinism as a school of thought. Probabilistic methods came more recently with many applications. Of course the basic elements were developed during the last centuries but with only few applications. Some of the famous contributors are P.-S. Laplace and J.C.F. Gauss. A main development was done by studying and modeling the heat transfer by proposing and solving a partial differential equation for the space and time propagation of heat (see Fourier (1822, 1878) and Fick (1855)). However radical progress came by modeling the Brownian motion, Brown (1828), (see the seminal paper by Einstein (1905) followed by Smoluchowski (1906)). (See also ▶Brownian Motion and Diffusions)

Modeling by Stochastic Differential Equations

Time was needed to understand and introduce probabilistic ideas into differential equations; thus called ▶stochastic differential equations. This was achieved only during the twentieth century. Even more some very important details were missing. One important point had to do with calculus and how to apply calculus in stochastic differential equations. The solution came with Itô and his postulate that the infinitesimal second order terms of a stochastic process do not vanish thus accepting to apply rules of what is now called as the Itô calculus or stochastic calculus. Stochastic calculus is also proposed by others differentiating their work from Itô's calculus on the summation process applied in defining the stochastic integral (R.L. Stratonovich, P. Malliavin). Itô's proposition can be given in the following form useful to apply in stochastic differential equations, Oksendal (1989), Gardiner (1990):

$$df(x_t,t) = \frac{\partial f(x_t,t)}{\partial x_t}dx_t + \frac{1}{2}\frac{\partial^2 f(x_t,t)}{\partial {x_t}^2}(dx_t)^2$$

where x_t is a stochastic process over time t and $f(x_t,t)$ is a stochastic function of the specific process.

The above form for the function $f(x_t,t)$ usually is used as a transformation function to reduce a nonlinear stochastic differential equation to a linear one and thus finding a closed form solution.

Although the first proposal of a probabilistic differential equation is merely due to P. Langevin, in recent years it

was generally accepted the following stochastic differential equations form:

$$dx_t = \mu(x_t, t)dt + \sigma(x_t, t)dw_t,$$

where w_t is the so-called Wiener process. This is a stochastic process with mean value zero and variance 1 and is usually termed as the standard Wiener process with $N(0,1)$ property, the process is characterized by independent increments normally distributed.

By using the above Itô's rule and the appropriate transformation function the exact solutions of several nonlinear stochastic differential equations arise. Except of the usefulness of the exact solutions of stochastic differential equations when dealing with specific cases and applications their use is very important in order to check how precise the approximate methods of solution of stochastic differential equations are. A general method of solution was proposed by Kloeden et al. (1992, 1999, 2003). Related theoretical solutions with applications can be found in Skiadas et al. (1993, 1994), Giovanis and Skiadas (1995), Skiadas and Giovanis (1997), Skiadas (2010).

The main stochastic differential equations solved can be summarized into two categories: The stochastic differential equations with a multiplicative error term of the form: $dx_t = \mu(x_t, t)dt + \sigma(t)x_t dw_t$, frequently used in market applications, and the stochastic differential equations with non-multiplicative or additive error term of the form: $dx_t = \mu(x_t, t)dt + \sigma(t)dw_t$. In the later case there appear applications with a constant σ.

The most known model with a multiplicative error term is the so-called Black and Scholes (1973) model in finance: $dx_t = \mu x_t dt + \sigma x_t dw_t$ (in most applications x_t is replaced by S_t).

The famous Ornstein–Uhlenbeck (1930) process is the most typical model with an additive error term: $dx_t = \vartheta(\mu - x_t)dt + \sigma dw_t$.

There are very many stochastic differential equations that could find interesting applications. As it was shown (Skiadas-Katsamaki 1995) even a general stochastic exponential model could give realistic paths especially during the first stages of a diffusion process: $dx_t = \mu(x_t)^b dt + \sigma dw_t$. In the same paper three methods for estimating the parameter σ are given.

Modeling using stochastic differential equations has several applications but also faces the problems arising from the introduction of stochastic theory. First of all, a stochastic differential equation gives a solution which may provide several stochastic paths during a simulation. However, one cannot find one final path as it is the case in a deterministic process. In most cases the deterministic solution arises by eliminating the error term. An infinite number of stochastic paths could provide the mean value of the stochastic process as a limit of a summation. When there exists an exact solution of the stochastic differential equation it can be estimated the mean value and if possible the variance. More useful, after estimating the mean value and the variance, is the estimation of the confidence intervals, thus informing regarding the limits of the real life application modeled.

Acknowledgments

For biography *see* the entry ►Chaotic Modelling.

Cross References

►Brownian Motion and Diffusions
►Chaotic Modelling
►Ito Integral
►Numerical Methods for Stochastic Differential Equations
►Optimal Statistical Inference in Financial Engineering
►Probability Theory: An Outline
►Stochastic Differential Equations
►Stochastic Modeling Analysis and Applications
►Stochastic Models of Transport Processes
►Stochastic Processes

References and Further Reading

Kloeden PE, Schurz H, Platten E, Sorensen M (1992). On effects of discretization on estimators of drift parameters for diffusion processes. Research Report no. 249, Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81(3):637–654

Brown R (1828) A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. Philos Mag 4:161–173

Einstein A (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Annalen der Physik 17:549–560

Fick A (1855) Über Diffusion. Poggendorff's Annalen 94:59–86

Fick A (1855b) On liquid diffusion. Philos Mag J Sci 10:31–39

Fourier J (1822) Theorie analytique de la chaleur. Firmin Didot, Paris

Fourier J (1878) The analytical theory of heat. Cambridge University Press, Cambridge

Gardiner CW (1990) Handbook of stochastic methods for physics, chemistry and natural science, 2nd edn. Springer, Berlin

Giovanis AN, Skiadas CH (1995) Forecasting the electricity consumption by applying stochastic modeling techniques: the case of Greece. In: Janssen J, Skiadas CH, Zopounidis C (eds) Advances in applying stochastic modeling and data analysis. Kluwer Academic, Dordrecht

Itô K (1944) Stochastic integral. In: Proceedings of the imperial academy of Tokyo, vol 20, pp 519–524

Itô K (1951) On stochastic differential equations. Mem Am Math Soc 4:1–51

Katsamaki A, Skiadas CH (1995) Analytic solution and estimation of parameters on a stochastic exponential model for a technological diffusion process. Appl Stoch Model Data Anal 11(1):59-75
Kloeden PE, Platen E (1999) Numerical solution of stochastic differential equations. Springer, Berlin
Kloeden PE, Platen E, Schurz H (2003) Numerical solution of SDE through computer experiments. Springer, Berlin
Oksendal B (1989) Stochastic differential equations: an introduction with applications, 2nd edn. Springer, New York
Skiadas CH (1985) Two generalized rational models for forecasting innovation diffusion. Technol Forecast Soc Change 27: 39–61
Skiadas CH (1986) Innovation diffusion models expressing asymmetry and/or positively or negatively influencing forces. Technol Forecast Soc Change 30:313–330
Skiadas CH (1987) Two simple models for the early and middle stage prediction of innovation diffusion. IEEE Trans Eng Manag 34:79–84
Skiadas CH (2010) Exact solutions of stochastic differential equations: Gompertz, generalized logistic and revised exponential. Meth Comput Appl Probab 12(2):261–270
Skiadas CH, Giovanis AN (1997) A stochastic bass innovation diffusion model studying the growth of electricity consumption in Greece. Appl Stoch Model Data Anal 13:85–101
Skiadas CH, Giovanis AN, Dimoticalis J (1993) A sigmoid stochastic growth model derived from the revised exponential. In: Janssen J, Skiadas CH (eds) Applied stochastic models and data analysis. World scientific, Singapore, pp 864–870
Skiadas CH, Giovanis AN, Dimoticalis J (1994) Investigation of stochastic differential models: the Gompertzian case. In: Gutierez R, Valderama Bonnet MJ (eds) Selected topics on stochastic modeling. World Scientific, Singapore, pp 296–310
Smoluchowski M (1906) Zur kinetischen theorie der Brownschen molekularbewegung und der suspensionen. Ann D Phys 21: 756–780
Uhlenbeck GE, Ornstein LS (1930) On the theory of Brownian motion. Phys Rev 36:823–41

Stochastic Modeling Analysis and Applications

Anil G. Ladde[1], Gangaram S. Ladde[2]
[1]Chesapeake Capital Corporation, Richmond, VA, USA
[2]Professor
University of South Florida, Tampa, FL, USA

The classical random flow and Newtonian mechanics are two theoretical approaches to analyze dynamic processes in biological, engineering, physical and social sciences under random perturbations. Historically, in the classical approach (Bartlett; 1969, Ross; 1971), one considers a dynamic system as a random flow or process with a certain probabilistic laws such as: diffusion, Markovian, nonmarkovian and etc. From this type consideration, one attempts to determine the state transition probability distributions/density functions (STPDF) of the random process. The determination of the unknown STPDF leads to the study of deterministic problems in the theory of ordinary or partial or integro-differential equations (Lakshmikantham and Leela 1969a, b). For example, a random flow that obeys a Markovian probabilistic law leads to

$$\frac{\partial}{\partial s}P(s,x,t,B) = q(s,x)P(s,x,t,B) - \int_{R^n-\{x\}} P(s,y,t,B)Q(s,x,dy), \tag{1}$$

that is, Kolmogorov's backward equation, where, $P(s,x,t,B)$ is STPDF; $Q(s,x,dy)$ is the state transition intensity function (STIF) and $q(s,x) = -Q(s,x,\{x\})$. In particular, in the case of Markov chain (see ►Markov Chains) with finite number of states r, equation (1) reduces to:

$$\frac{\partial}{\partial s}P(s,t) = Q(s)P(s,t),\ P(t,t) = I, \tag{2}$$

where, $P_{ij}(s,t) = P(s,i,t,\{j\})$; $P(s,t) = (P_{ij}(s,t))_{r\times r}$; an intensity matrix $Q(s)$ and the identity I are $r \times r$ matrices. These types of equations are referred as master equations in the literature (Arnold 1974; Bartlett 1960; Gīhman 1972; Gikhman and Skorokhod 1969; Goel and Richter-Dyn 1974; Kimura and Ohta 1971; Kloeden and Platen 1992; Ladde 1991; Ladde and Sambandham 2004; Ricciardi 1977; Soong 1973). The solution processes of such differential equations are used to find the higher moments and other statistical properties of dynamic processes described by random flows or processes in sciences. We remark that in general, Kolmogorov's backward or forward (master equations) are nonlinear and non stationary deterministic differential equations (Arnold 1974; Gīhman 1972; Gikhman and Skorokhod 1969; Goel and Richter-Dyn 1974; Ricciardi 1977; Soong 1973). As a result of this, the close form STPDF are not feasible.

A modern approach (Arnold 1974; Gīhman 1972; Ito 1951, Kloeden and Platen 1992; Ladde and Ladde 2009; Ladde 1991; Ladde and Lakshmikantham 1980; Ladde and Sambandham 2004; Nelson 1967; Øksendal 1985; Ricciardi 1977; Soong 1973; Wong 1971) of stochastic modeling of dynamic processes in sciences and engineering sciences is based on fundamental theoretical information, a practical experimental setup and basic laws in science and engineering sciences. Depending on the nature of stochastic disturbances, there are several probabilistic models, namely, ►Random walk, Poisson, Brownian motion (see

►Brownian Motion and Diffusions), Colored Noise processes. In the following, we very briefly outline the salient features of Random Walk and Colored Noise dynamic modeling approaches (Kloeden and Platen 1992; Ladde and Ladde 2009; Wong 1971).

Random Walk Modeling Approach (Ladde and Ladde 2009)

Let $x(t)$ be a state of a system at a time t. The state of the system is observed over an interval of $[t, t+\Delta t]$, where Δt is a small increment in t. Without loss in generality, it is assumed that $x(t)$ is 1-dimensional state and Δt is positive. The state is under the influence of random perturbations. We experimentally observe the data-set of the state: $x(t_0) = x(t), x(t_1), x(t_2), \ldots, x(t_i), \ldots, x(t_n) = x(t+\Delta t)$ of a system at $t_0 = t, t_1 = t+\tau, t_2 = t+2\tau, \ldots, t_i = t+i\tau, \ldots, t_n = t+\Delta t = t+n\tau$ over the interval $[t, t+\Delta t]$, where n belongs to $\{1, 2, 3, \ldots\}$ and $\tau = \frac{\Delta t}{n}$. These observations are made under the following conditions:

RWM 1 The system is under the influence of independent and identical random impulses that are taking place at $t_1, t_2, \ldots, t_i, \ldots, t_n$.

RWM 2 The influence of a random impact on the state of the system is observed on every time subinterval of length τ.

RWM 3 For each $i \in I(1,n) = \{1, 2, \ldots, k, \ldots, n\}$, it is assumed that the state is either increased by $\Delta x(t_i)$ ("success"-the positive increment $(\Delta x(t_i) > 0)$) or decreased by $\Delta x(t_i)$ ("failure"-the negative increment $(\Delta x(t_i) < 0)$). We refer $\Delta x(t_i)$ as a microscopic/local experimentally or knowledge-base observed increment to the state of the system at the ith impact on the subinterval of length τ.

RWM 4 It is assumed that $\Delta x(t_i)$ is constant for $i \in I(1,n)$ and is denoted by $\Delta x(t_i) \equiv Z_i = Z$ with $|Z_i| = \Delta x > 0$. Thus, for each $i \in I(1,n)$, there is a constant random increment Z of magnitude Δx to the state of the system per impact on the subinterval of length τ.

RWM 5 For each random impact and any real number p satisfying $0 < p < 1$, it is assumed that

$$P(\{Z_i = \Delta x > 0\}) = p \text{ and } P(\{Z_i = -\Delta x < 0\}) = 1 - p = q. \tag{3}$$

From RWM1, RWM2 and RWM3, under n independent and identical random impacts, the initial state and n experimental or knowledge-base observed random increments Z_i of constant magnitude Δx in the state, the aggregate change of the state of the system $x(t+\Delta t) - x(t)$ under n observations of the system over the given interval $[t, t+\Delta t]$ of length Δt is described by

$$x(t+\Delta t) - x(t) = n\frac{\left[\sum_{i=1}^{n} Z_i\right]}{n} = \frac{\Delta t}{\tau} S_n, \tag{4}$$

where $S_n = \frac{1}{n}\left[\sum_{i=1}^{n} Z_i\right]$ and $Z_i = x(t_i) - x(t_{i-1})$. S_n is the sample average of the state aggregate incremental data. It is clear that $x(t+\Delta t) - x(t) = x(t_n) - x(t)$ is a discrete-time-real-valued stochastic process which is the sum of n independent Bernoulli random variables Z_i $(Z_i = Z)$, $i = 1, 2, \ldots, n$. We also note that for each n, $x(t_n) - x(t_0)$ is a binomial random variable with parameters (n, p). Moreover, the random variable $x(t_n) - x(t)$ takes values from the set $\{-n\Delta x, (2-n)\Delta x, \ldots, (2m-n)\Delta x, \ldots, n\Delta x\}$. The stochastic process $x(t_n) - x(t)$ is referred to as a *Random Walk process*. Let m be a number of positive increments Δx to the state of the system out of total n changes. $(n-m)$ is the number of negative increments $-\Delta x$ to the state of the system out of total n changes. Furthermore, $m \in I(0,n)$, we further note that

$$S_n = \frac{1}{n}\left[(2m-n)S_n^+\right], \tag{5}$$

where $S_n^+ = \frac{1}{n}\left[\sum_{i=1}^{n} |Z_i|\right]$.

Therefore, the *aggregate change* of state, $x(t+\Delta t) - x(t)$ under n identical random impacts on the system over the given interval $[t, t+\Delta t]$ of time is described by

$$x(t+\Delta t) - x(t) = \frac{1}{n}(2m-n)\frac{S_n^+}{\tau}\Delta t. \tag{6}$$

Moreover, from (6), we have:

$$E[x(t+\Delta t) - x(t)] = (p-q)\frac{S_n^+}{\tau}\Delta t \tag{7}$$

and

$$\mathrm{Var}(x(t+\Delta t) - x(t)) = 4pq\frac{\left(S_n^+\right)^2}{\tau}\Delta t. \tag{8}$$

$\frac{S_n^+}{\tau}$ and $\frac{(S_n^+)^2}{\tau}$ are *sample microscopic or local average increment* and *sample microscopic or local average square increment* per unit time over the uniform length of sample subintervals $[t_{k-1}, t_k]$, $k = 1, 2, \ldots, n$ of interval $[t, t+\Delta t]$, respectively.

We note that the physical nature of the problem imposes certain restrictions on Δx and τ. Similarly, the parameter p cannot be taken arbitrary. In fact, the following conditions seem to be natural for sufficiently large n:

For $x(t+\Delta t)-x(t)=n\Delta x$, $\Delta t=n\tau$, $4pq=(p+q)^2-(p-q)^2=1-(p-q)^2$, and

$$\lim_{\tau\to 0}\left[\frac{\left(S_n^+\right)^2}{\tau}\right]=2D,\ \lim_{\Delta x\to 0}\lim_{\tau\to 0}\left[(p-q)\frac{S_n^+}{\tau}\right]=C \text{ and } \lim_{\Delta x\to 0}\lim_{\tau\to 0}4pq=1, \tag{9}$$

where C and D are certain constants, the former is called a *drift* coefficient, and the latter is called a *diffusion* coefficient. Moreover, C can be interpreted as the *average/mean/expected rate of change of state* of the system per unit time, and D can be interpreted as the mean square rate of change of the system per unit time over an interval of length Δt. From (7), (8) and (9), we obtain

$$\lim_{\Delta x\to 0}\lim_{\tau\to 0}E[x(t+\Delta t)-x(t)]=C\Delta t, \tag{10}$$

and

$$\lim_{\Delta x\to 0}\lim_{\tau\to 0}\mathrm{Var}(x(t+\Delta t)-x(t))=2D\Delta t. \tag{11}$$

Now, we define

$$y(t,n,\Delta t)=\frac{x(t+\Delta t)-x(t)-n(p-q)S_n^+}{\sqrt{4npq\left(S_n^+\right)^2}}. \tag{12}$$

By the application of the DeMoivre–Laplace Central Limit Theorem, we conclude that the process $y(t,n,\Delta t)$ is approximated by standard normal random variable for each t (zero mean and variance one). Moreover,

$$\lim_{\Delta x\to 0}\lim_{\tau\to 0}y(t,n,\Delta t)=\frac{x(t+\Delta t)-x(t)-C\Delta t}{\sqrt{2D\Delta t}}. \tag{13}$$

For fixed Δt, the random variable $\lim_{\Delta x\to 0}\lim_{\tau\to 0}y(t,n,\Delta t)$ has standard normal distribution (zero mean and variance one). Now, by rearranging the expressions in (13), we get

$$x(t+\Delta t)-x(t)=C\Delta t+\sqrt{2D}\,\Delta w(t) \tag{14}$$

where $\sqrt{\Delta t}\left[\lim_{\Delta x\to 0}\lim_{\tau\to 0}y(t,n,\Delta t)\right]=\Delta w(t)=w(t+\Delta t)-w(t)$, $w(t)$ is a Wiener process. Thus the aggregate change of state of the system $x(t+\Delta t)-x(t)$ in (14) under independent and identical random impacts over the given interval $[t,t+\Delta t]$ is interpreted as the sum of the average/expected/mean change $(C\Delta t)$ and the mean square change $\left(\sqrt{2D}\,\Delta w(t)\right)$ of state of the system due to the random environmental perturbations.

If Δt is very small, then its differential $dt=\Delta t$, and from (14) the Itô–Doob differential dx is defined by

$$dx(t)=C\,dt+\sqrt{2D}\,dw(t), \tag{15}$$

where C and D are as defined before. The equation in (15) is called the Itô–Doob type stochastic differential equation (Arnold 1974; Gīhman and Skorohod 1972; Ito 1951; Kloeden and Platen 1992; Laddle and Laddle 2009; Laddle and Lakshmikantham 1980; Øksendal 1985; Soong 1973; Wong 1971).

Observation (1) We recall that the experimental or knowledge base observed constant random variables: $x(t_0)=x(t),Z_1,Z_2,\ldots,Z_k,\ldots,Z_n$ in (4) are mutually independent. Therefore, expectations

$$E[x(t+\Delta t)-x(t)] \text{ and } E\left[(x(t+\Delta t)-x(t))^2\right]=\mathrm{Var}(x(t+\Delta t)-x(t))$$

in (7) and (8) can be replaced by the conditional expectations as:

$$E[x(t+\Delta t)-x(t)]=E[x(t+\Delta t)-x(t)\mid x(t)=x] \tag{16}$$

and

$$\mathrm{Var}(x(t+\Delta t)-x(t))=E\left[(x(t+\Delta t)-x(t))^2\mid x(t)=x\right]. \tag{17}$$

(2) We further note that based on experimental observations, information and basic scientific laws/principles in biological, chemical, engineering, medical, physical and social sciences, we infer that in general the magnitude of the microscopic or local increment depends on both the initial time t and the initial state $x(t)\equiv x$ of a system. As a result of this, in general, the drift (C) and the diffusion (D) coefficients defined in (9) need not be absolute constants. They may depend on both the initial time t and the initial state $x(t)\equiv x$ of the system, as long as their dependence on t and x is very smooth. From this discussion, (16) and (17), one can incorporate both time and state dependent random environmental perturbation effects. As a result of this, (14) reduces to:

$$x(t+\Delta t)-x(t)=C(t,x)\Delta t+\sigma(t,x)\Delta w(t), \tag{18}$$

where $C(t,x)$ and $\sigma^2(t,x)=2D(t,x)$ are also referred to as the average/expected/mean rate and the mean square rate of the state of the system on the interval of length Δt. Moreover, the Itô–Doob type stochastic differential equation (15) becomes:

$$dx(t)=C(t,x)\,dt+\sigma(t,x)dw(t). \tag{19}$$

(3) From (16), (17) and (19), we have

$$\frac{d}{dt}E[x(t)\mid x(t)=x]=C(t,x), \tag{20}$$

$$dx=C(t,x)dt+\sigma(t,x)\xi(t)\,dt, \tag{21}$$

$$dx=C(t,x)\,dt \tag{22}$$

where $w(t)$ is the Wiener process and $\xi(t)$ is the white noise process. We further remark that either (19) or (21) is considered as a stochastic perturbation of deterministic

differential equation (22). The random terms $\sigma(t,x)\,dw(t)$ and $\sigma(t,x)\xi(t)$ in the right-hand side of (19) and (21), respectively, can be, normally, interpreted as random perturbations caused by the presence of microscopic and/or the imperfectness of the controlled conditions, either known or unknown and/or either environmental or internal fluctuations in the parameters in $C(t,x)$. It is this idea that motivates us to build a more general and feasible stochastic mathematical model for dynamic processes in biological, chemical, engineering, medical, physical and social sciences.

Sequential Colored Noise Modeling Approach (Ladde and Ladde 2009; Wong 1971)

The idea is to start with a deterministic mathematical model (22) that is based on phenomenological or biological or chemical/medical/physical social laws and the knowledge of system or environmental parameter(s). From Observation (3), one can identify parameter(s) and the source of random internal or environmental perturbations of parameter(s) of the mathematical model (22), and formulate a stochastic mathematical model in general form as:

$$dx = F(t,x,\xi(t))\,dt, \quad x(t_0) = x_0, \tag{23}$$

and, in particular,

$$dx = C(t,x)\,dx + \sigma(t,x)\xi(t)\,dt, \quad x(t_0) = x_0, \tag{24}$$

where ξ is a stochastic process that belongs to $R[[a,b], R[\Omega,R]]$; rate functions F, $C(t,x)$ and $\sigma(t,x)$ are sufficiently smooth, and are defined on $[a,b]\times R$ into R, $x_0 \in R$ and $t_0 \in [a,b]$. If the sample paths $\xi(t,\omega)$ of $\xi(t)$ are smooth functions (sample continuous), then one can utilize the usual deterministic calculus, and can look for the solution process determined by (23) and (24). We note that such a solution process is a random function with all sample paths starting at x_0. In general this is not feasible, for example, if $\xi(t)$ in (23) or (24) is a Gaussian process. The sequential colored noise modeling (CNM) approach alleviates the limitations of a one-shot modeling approach. The basic ideas are as follows:

CNM 1 Let us start with a sequence $\{\xi_n(t)\}_{n=1}^{\infty}$ of sufficiently smooth (sample path wise continuous) Gaussian processes which converges in some sense to a Gaussian white noise process $\xi(t)$ in (24). For each n, we associate a stochastic differential equation with a smooth random process as follows:

$$dx_n = C(t,x_n) + \sigma(t,x_n)\,\xi_n(t)\,dt, \quad x_n(t_0) = x_0 \tag{25}$$

where $C(t,x)$ and $\sigma(t,x)$ are described in (24).

CNM 2 We assume that the IVP (25) has a unique solution process. The IVP (25) generates a sequence $\{x_n(t)\}_{n=1}^{\infty}$ of solution processes corresponding to the chosen Gaussian sequence $\{\xi_n(t)\}_{n=1}^{\infty}$ in CNM1.

CNM 3 Under reasonable conditions on rate functions $C(t,x)$, $\sigma(t,x)$ in (24) and a suitable convergent sequence of Gaussian processes $\{\xi_n(t)\}_{n=1}^{\infty}$ in CNM1, it is shown that the sequence of solution processes $\{x_n(t)\}_{n=1}^{\infty}$ determined by (25) converges in almost surely or in quadratic mean or even in probability to a process $x(t)$. Moreover, $x(t)$ is the solution process of (24).

CNM 4 The above described ideas CNM1, CNM2 and CNM3 make a precise mathematical interpretation of (24). However, we still need to show that (24) can be modeled by an Itô–Doob form of stochastic differential equation (19). Moreover, one needs to highlight on the concept of convergence of $\{\xi_n(t)\}_{n=1}^{\infty}$ to the white noise process in (24). For this purpose, we define

$$w_n(t) - w_n(t_0) = \int_{t_0}^{t} \xi_n(s)\,ds, \tag{26}$$

and rewrite the IVP (25) into its equivalent integral form:

$$\begin{aligned} x_n(t) &= x_n(t_0) + \int_{t_0}^{t} C(s,x_n(s))\,ds \\ &\quad + \int_{t_0}^{t} \sigma(s,x_n(s))\,\xi_n(s)\,ds \\ &= x_n(t_0) + \int_{t_0}^{t} C(s,x_n(s))\,ds \\ &\quad + \int_{t_0}^{t} \sigma(s,x_n(s))\,dw_n(s). \end{aligned} \tag{27}$$

CNM 5 To conclude the convergence of $\{x_n(t)\}_{n=1}^{\infty}$, we need to show the convergence of both terms in the right-hand side of (27). The procedure for showing this convergence generates the following two mathematical steps:

Step 1: This step is to establish the following as in Ladde and Ladde (2009) and Wong (1971):

$$\begin{aligned} \lim_{n\to\infty}[y_n(t)] &= \lim_{n\to\infty}\left[\int_{t_n}^{t} \phi(s,w_n(s))\,dw_n(s)\right] \\ &= \int_{t_0}^{t} \phi(s,w(s))\,dw(s) \\ &\quad + \frac{1}{2}\int_{t_0}^{t} \frac{\partial}{\partial z}\phi(s,w(s))\,ds, \end{aligned} \tag{28}$$

where ϕ is a known smooth function of two variables. This is achieved by considering a deterministic partial indefinite integral of a given smooth deterministic function ϕ:

$$\psi(t,x) = \int_0^x \phi(t,z)\,dz. \tag{29}$$

Step 2: This step deals with the procedure of finding a limit of the sequence of the solution process $\{x_n(t)\}_{n=1}^{\infty}$ determined by (25) or its equivalent stochastic differential equation (27) as in Ladde and Ladde; 2009 and Wong; 1971:

$$dx_n = C(t,x_n)\,dt + \sigma(t,x_n)\,dw_n(t),$$
$$x_n(t_0) = x_0, \tag{30}$$

where $w_n(t)$ is as defined in (26). For this purpose, we assume that $\sigma(t,z)$ in (24) satisfies the conditions: $\sigma(t,z) \neq 0$, and it is continuously differentiable. We set $\phi(t,z) = \frac{1}{\sigma(t,z)}$ in (29). Under the smoothness conditions on rate functions C, σ and imitating the procedure outlined in Step 1, one can conclude that $\{x_n(t)\}_{n=1}^{\infty}$ converges to a process $x(t)$ on $[t_0,b]$. The final conclusion is to show that $x(t)$ satisfies the following Itô–Doob type stochastic differential equation:

$$dx = \left[C(t,x) + \frac{1}{2}\sigma(t,x)\,\frac{\partial}{\partial x}\,\sigma(t,x)\right] dt + \sigma(t,x)\,dw(t), \quad x(t_0) = x_0. \tag{31}$$

This is achieved by the procedure of solving the Itô–Doob type stochastic differential equation in the form of (30). The procedure is to reduce differential equation (30) into the following reduced integrable differential equation as in (Gīhman and Skorohod (1972); Kloeden and Platen (1992); Ladde and Ladde (2009) and Wong (1971)):

$$dm = f(t)\,dt + g(t)\,dw(t), \tag{32}$$

where $f(t)$ and $g(t)$ are suitable stochastic processes determined by rate functions C and σ in (24). The extra term $\frac{1}{2}\sigma(t,x)\frac{\partial}{\partial x}\sigma(t,x)$ in (31) is referred to as the *correction term*.

In summary, it is further detailed as shown in Ladde and Ladde (2009) and Wong (1971) that if we interpret Gaussian white-noise driven differential equation (24) by the limit of a sequence of stochastic differential equations (25) with a sequential colored noise process, then the Gaussian white-noise driven differential equation (24) is equivalent to the Itˆo–Doob type stochastic differential equation (31). Moreover, this material is 1-dimensional state variable, however, it can be easily extended to multi-dimensional state space.

Several dynamic processes are under both internal and external random distributions. The usage of this information coupled with different modes in probabilistic analysis, namely, an approach through sample calculus, L^p-calculus, and Itô–Doob calculus as in (Ladde and Lakshmikantham; 1980, Ladde and Sambandham; 2004, Nelson; 1967, Øksendal; 1985 and Soong; 1973) leads to different dynamic models. The majority of the dynamic models are in the context of Itô–Doob calculus (Arnold; 1974, Gīhman; 1972, Ito; 1951, Kloeden and Platen; 1992, Ladde; 1991, Ladde and Ladde; 2009, Ladde and Lakshmikantham; 1980; Nelson; 1967, Øksendal; 1985, Soong; 1973, Wong; 1971) and are described by systems of stochastic differential equations

$$dx = f(t,x)\,dt + \sigma(t,x)w(t), \quad x(t_0) = x_0, \tag{33}$$

where dx is the Itô–Doob type stochastic differential of x, $x \in R^n$, w is m-dimensional normalized Wiener process defined on a complete probability space $(\Omega, \Im, P)$, $f(t,x)$ is drift rate vector, and $\sigma(t,x)$ is a diffusion rate matrix of size $n \times m$. Various qualitative properties (Arnold; 1974, Ladde; 1991, Ladde and Lakshmikantham; 1980, Ladde and Sambandham; 2004, Soong; 1973, Wong; 1971) have played a very significant role in state estimation and system designing processes since the beginning or middle of the twentieth century.

Acknowledgment

This research was supported by Mathematical Sciences Division, US Army Research Office, Grant No. W911NF-07-1-0283.

About the Author

Dr. Gangaram Ladde is Professor of Mathematics and Statistics, University of South Florida (since 2007). Prior to that he was Professor of Mathematics, University of Texas at Arlington (1980–2007). He received his Ph.D. in Mathematics from University of Rhode Island in 1972. He has published more than 150 papers, has co-authored 4 monographs, and co-edited 6 proceedings of international conferences, including, (1) *Stochastic Versus Deterministic Systems of Differential Equations*, (with M. Sambandham, Marcel Dekker, Inc, New York, 2004) and (2) *Random Differential Inequalities* (with V. Lakshmikantham, Academic Press, New York, 1980). Dr. Ladde is the Founder and Joint Editor-in-Chief (1983–present) of the *Journal of Stochastic Analysis and Applications*. He is also a Member of Editorial Board of several journals in Mathematical Sciences. Dr. Ladde is recipient of several research awards and grants.

Cross References

►Brownian Motion and Diffusions
►Gaussian Processes
►Markov Chains
►Random Walk

►Stochastic Differential Equations
►Stochastic Modeling, Recent Advances in
►Stochastic Models of Transport Processes
►Stochastic Processes: Classification

References and Further Reading

Arnold L (1974) Stochastic differential equations: theory and applications. Wiley-Interscience (Wiley), New York, Translated from the German

Bartlett MS (1960) Stochastic population models in ecology and epidemiology. Methuen's Monographs on Applied Probability and Statistics, Methuen, London

Gĭhman ĬĬ, Skorohod AV (1972) Stochastic differential equations. Springer, New York, Translated from the Russian by Kenneth Wickwire, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 72

Gikhman II, Skorokhod AV (1969) Introduction to the theory of random processes. Translated from the Russian by Scripta Technica, W.B. Saunders, Philadelphia, PA

Goel NS, Richter-Dyn N (1974) Stochastic models in biology. Academic (A subsidiary of Harcourt Brace Jovanovich), New York–London

Ito K (1951) On stochastic differential equations. Mem Am Math Soc 1951(4):51

Kimura M, Ohta T (1971) Theoretical aspects of population genetics. Princeton University Press, Princeton, NJ

Kloeden PE, Platen E (1992) Numerical solution of stochastic differential equations. Applications of mathematics (New York), vol 23, Springer, Berlin

Ladde GS (1991) Stochastic delay differential systems. World Scientific, Hackensack, NJ, pp 204–212

Ladde AG, Ladde GS (2009) An introduction to differential equations: stochastic modeling, methods and analysis, vol II. In Publication Process

Ladde GS, Lakshmikantham V (1980) Random differential inequalities. Mathematics in Science and Engineering, vol 150, Academic (Harcourt Brace Jovanovich), New York

Ladde GS, Sambandham M (2004) Stochastic versus deterministic systems of differential equations. Monographs and textbooks in pure and applied mathematics, vol 260. Marcel Dekker, New York

Lakshmikantham V, Leela S (1969a) Differential and integral inequalities: theory and applications, volume I: ordinary differential equations. Mathematics in science and engineering, vol 55-I. Academic, New York

Lakshmikantham V, Leela S (1969b) Differential and integral inequalities: theory and applications, vol II: functional, partial, abstract, and complex differential equations. Mathematics in science and engineering, vol 55-II. Academic, New York

Nelson E (1967) Dynamical theories of Brownian motion. Princeton University Press, Princeton, NJ

Oksendal B (1985) Stochastic differential equations. An introduction with applications. Universitext, Springer, Berlin

Ricciardi LM (1977) Diffusion processes and related topics in biology. Springer, Berlin, Notes taken by Charles E. Smith, Lecture Notes in Biomathematics, vol 14

Ross SM (1972) Introduction to probability models. Probability and mathematical statistics, vol 10. Academic, New York

Soong TT (1973) Random differential equations in science and engineering. Mathematics in science and engineering, vol 103. Academic (Harcourt Brace Jovanovich), New York

Wong E (1971) Stochastic processes in information and dynamical systems. McGraw-Hill, New York, NY

Stochastic Models of Transport Processes

Alexander D. Kolesnik
Professor
Institute of Mathematics & Computer Science, Academy of Sciences of Moldova, Kishinev, Moldova

The transport process $\mathbf{X}(t) = (X_1(t), \ldots, X_m(t))$ in the Euclidean space, $\mathbb{R}^m$, $m \geq 1$, is generated by the stochastic motion of a particle that, at the time instant $t = 0$, starts from some initial point (e.g., origin) of $\mathbb{R}^m$ and moves with some finite speed c in random direction. The motion is controlled by some stochastic process $x(t)$, $t \geq 0$, causing, at random time instants, the changes of direction chosen randomly according to some distribution on the unit sphere $S_1^m \subset \mathbb{R}^m$. Such stochastic motions, also called random flights, represent the most important type of random evolutions (for limit and asymptotic theorems for general random evolutions see, for instance, Papanicolaou [1975], Pinsky [1991], Korolyuk and Swishchuk [1994] and the bibliographies therein). While the finiteness of the velocity is the basic feature of such motions, the models differ with respect to the way of choosing the new directions (the scattering function), the type of the governing stochastic process $x(t)$, and the dimension of the space $\mathbb{R}^m$. If the new directions are taken on according to the uniform probability law and the phase space $\mathbb{R}^m$ is isotropic and homogeneous, $\mathbf{X}(t)$ is referred to as the *isotropic* transport process. The most studied model is referred to the case when the speed c is constant and $x(t)$ is the homogeneous Poisson process (see ►Poisson Processes).

The simplest one-dimensional isotropic transport process with constant finite speed c driven by a homogeneous Poisson process of rate $\lambda > 0$ was first studied by Goldstein (1951) and Kac (1956). They have shown that the transition density $f = f(x,t)$, $x \in \mathbb{R}^1$, $t > 0$, of the process satisfies the telegraph equation

$$\frac{\partial^2 f}{\partial t^2} + 2\lambda \frac{\partial f}{\partial t} - c^2 \frac{\partial^2 f}{\partial x^2} = 0, \tag{1}$$

and can be found by solving this equation with the initial conditions $f(x,0) = \delta(x)$, $\left.\frac{\partial f}{\partial t}\right|_{t=0} = 0$, where $\delta(x)$ is the one-dimensional Dirac delta-function. The explicit form of

the transition density of the process (i.e., the fundamental solution to (1)) is given by the formula

$$f(x,t) = \frac{e^{-\lambda t}}{2}\left[\delta(ct+x)+\delta(ct-x)\right] + \frac{e^{-\lambda t}}{2c}\left[\lambda I_0\left(\frac{\lambda}{c}\sqrt{c^2t^2-x^2}\right) + \frac{\lambda ct}{\sqrt{c^2t^2-x^2}}\, I_1\left(\frac{\lambda}{c}\sqrt{c^2t^2-x^2}\right)\right]\Theta(ct-|x|),$$
$$x\in\mathbb{R}^1, \qquad |x|\le ct, \qquad t>0, \tag{2}$$

where $I_0(x)$ and $I_1(x)$ are the Bessel functions of zero and first orders, respectively, with imaginary argument and $\Theta(x)$ is the Heaviside function. The first term in (2) represents the density of the singular component of the distribution (which is concentrated in two terminal points $\pm ct$ of the interval $[-ct, ct]$), while the second one represents the density of the absolutely continuous part of the distribution (which is concentrated in the open interval $(-ct, ct)$).

Let $\mathbf{X}(t)$, $t>0$, be the isotropic transport process in the Euclidean plane, $\mathbb{R}^2$, generated by the random motion of a particle moving with constant speed c and choosing new directions at random Poissonian (λ) instants according to the uniform probability law on the unit circumference. Then the transition density $f=f(\mathbf{x},t)$, $\mathbf{x}\in\mathbb{R}^2$, $t>0$, of $\mathbf{X}(t)$ has the form (Stadje 1987; Masoliver et al. 1993; Kolesnik and Orsingher 2005)

$$f(\mathbf{x},t) = \frac{e^{-\lambda t}}{2\pi ct}\,\delta(c^2t^2-\|\mathbf{x}\|^2) + \frac{\lambda}{2\pi c}\,\frac{\exp\left(-\lambda t+\frac{\lambda}{c}\sqrt{c^2t^2-\|\mathbf{x}\|^2}\right)}{\sqrt{c^2t^2-\|\mathbf{x}\|^2}} \times\Theta\left(ct-\|\mathbf{x}\|\right),$$
$$\mathbf{x}=(x_1,x_2)\in\mathbb{R}^2, \qquad \|\mathbf{x}\|=\sqrt{x_1^2+x_2^2}\le ct, \qquad t>0. \tag{3}$$

Similar to the one-dimensional case, the density (3) is the fundamental solution (the Green's function) to the two-dimensional telegraph equation

$$\frac{\partial^2 f}{\partial t^2}+2\lambda\frac{\partial f}{\partial t} = c^2\left\{\frac{\partial^2 f}{\partial x_1^2}+\frac{\partial^2 f}{\partial x_2^2}\right\}. \tag{4}$$

The transition density $f=f(\mathbf{x},t)$, $\mathbf{x}\in\mathbb{R}^3$, $t>0$, of the isotropic transport process $\mathbf{X}(t)$ with unit speed $c=1$ in the three-dimensional Euclidean space, $\mathbb{R}^3$, is given by the formula (Stadje 1989)

$$f(\mathbf{x},t) = \frac{e^{-\lambda t}}{4\pi t^2}\,\delta(t^2-\|\mathbf{x}\|^2) + \frac{\lambda\, e^{-\lambda t}}{4\pi\|\mathbf{x}\|}\left[\lambda\int_{-1}^{-\|\mathbf{x}\|/t}\exp(\lambda(\xi t+\|\mathbf{x}\|)\operatorname{arth}\xi)\,(\operatorname{arth}\xi)^2\,d\xi + \frac{1}{t}\operatorname{arth}\left(\frac{\|\mathbf{x}\|}{t}\right)\right]\Theta\left(t-\|\mathbf{x}\|\right),$$
$$\mathbf{x}=(x_1,x_2,x_3)\in\mathbb{R}^3, \qquad \|\mathbf{x}\|=\sqrt{x_1^2+x_2^2+x_3^2}\le t, \qquad t>0, \tag{5}$$

where $\operatorname{arth}(x)$ is the hyperbolic inverse tangent function.

In the four-dimensional Euclidean space, $\mathbb{R}^4$, the transition density $f=f(\mathbf{x},t)$, $\mathbf{x}\in\mathbb{R}^4$, $t>0$, of the isotropic transport process $\mathbf{X}(t)$ has the following form (Kolesnik 2006)

$$f(\mathbf{x},t) = \frac{e^{-\lambda t}}{2\pi^2(ct)^3}\,\delta(c^2t^2-\|\mathbf{x}\|^2) + \frac{\lambda t}{\pi^2(ct)^4} \times\left[2+\lambda t\left(1-\frac{\|\mathbf{x}\|^2}{c^2t^2}\right)\right]\exp\left(-\frac{\lambda}{c^2t}\|\mathbf{x}\|^2\right) \times\Theta\left(ct-\|\mathbf{x}\|\right),$$
$$\mathbf{x}=(x_1,x_2,x_3,x_4)\in\mathbb{R}^4, \qquad \|\mathbf{x}\|=\sqrt{x_1^2+x_2^2+x_3^2+x_4^2}\le ct, \qquad t>0. \tag{6}$$

We see that in the spaces $\mathbb{R}^2$ and $\mathbb{R}^4$, the transition densities of $\mathbf{X}(t)$ have very simple analytical forms (3) and (6) expressed in terms of elementary functions. In contrast, the three-dimensional density (5) has the fairly complicated form of an integral with variable limits which, apparently, cannot be explicitly evaluated. This fact shows that the behavior of transport processes in the Euclidean spaces $\mathbb{R}^m$ substantially depends on the dimension m. Moreover, while the transition densities of the processes on the line $\mathbb{R}^1$ and in the plane $\mathbb{R}^2$ are the fundamental solutions (i.e., the Green's functions) to the telegraph equations (1) and (4), respectively, the similar results for other spaces have not been obtained so far.

However, for the integral transforms of the distributions of $\mathbf{X}(t)$, one can give the most general formulas that are valid in any dimensions. Let $H(t)=\mathrm{E}\left\{e^{i(\boldsymbol{\alpha},\mathbf{X}(t))}\right\}$ be the characteristic function (Fourier transform) of the isotropic transport process $\mathbf{X}(t)$ in the Euclidean space $\mathbb{R}^m$ of arbitrary dimension $m\ge 2$. Here, $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_m)\in\mathbb{R}^m$ is the real m-dimensional vector of inversion parameters

and $(\boldsymbol{\alpha}, \mathbf{X}(t))$ means the inner product of the vectors $\boldsymbol{\alpha}$ and $\mathbf{X}(t)$. Introduce the function

$$\varphi(t) = 2^{(m-2)/2}\,\Gamma\left(\frac{m}{2}\right)\frac{J_{(m-2)/2}(ct\|\boldsymbol{\alpha}\|)}{(ct\|\boldsymbol{\alpha}\|)^{(m-2)/2}}, \qquad m \geq 2, \tag{7}$$

where $\|\boldsymbol{\alpha}\| = \sqrt{\alpha_1^2 + \cdots + \alpha_m^2}$, $\Gamma(x)$ is the Euler gamma-function and $J_{(m-2)/2}(x)$ is the Bessel function of order $(m-2)/2$ with real argument. Note that (7) is the characteristic function of the uniform distribution on the surface of the sphere of radius ct in the space $\mathbb{R}^m$, $m \geq 2$. Then the characteristic function $H(t)$, $t \geq 0$, satisfies the following convolution-type Volterra integral equation of second kind (Kolesnik 2008):

$$H(t) = e^{-\lambda t}\varphi(t) + \lambda\int_0^t e^{-\lambda(t-\tau)}\varphi(t-\tau)H(\tau)\,d\tau, \quad t \geq 0. \tag{8}$$

In the class of continuous functions, the integral equation (8) has the unique solution given by the uniformly converging series

$$H(t) = e^{-\lambda t}\sum_{n=0}^{\infty}\lambda^n\,[\varphi(t)]^{*(n+1)}, \tag{9}$$

where $[\varphi(t)]^{*(n+1)}$ means the $(n+1)$-multiple convolution of function (7) with itself. The Laplace transform $\mathcal{L}$ of the characteristic function $H(t)$ has the form (Kolesnik 2008)

$$\mathcal{L}[H(t)](s) = \frac{F\left(\frac{1}{2}, \frac{m-2}{2}; \frac{m}{2}; \frac{(c\|\boldsymbol{\alpha}\|)^2}{(s+\lambda)^2+(c\|\boldsymbol{\alpha}\|)^2}\right)}{\sqrt{(s+\lambda)^2 + (c\|\boldsymbol{\alpha}\|)^2} - \lambda\,F\left(\frac{1}{2}, \frac{m-2}{2}; \frac{m}{2}; \frac{(c\|\boldsymbol{\alpha}\|)^2}{(s+\lambda)^2+(c\|\boldsymbol{\alpha}\|)^2}\right)}, \quad m \geq 2, \tag{10}$$

for $\mathrm{Re}\, s > 0$, where $F(\xi, \eta; \zeta; z)$ is the Gauss hypergeometric function.

One of the most remarkable features of the isotropic transport processes in $\mathbb{R}^m$, $m \geq 2$, is their weak convergence to the Brownian motion (see ►Brownian Motion and Diffusions) as both the speed c and the intensity of switchings λ tend to infinity in such a way that the following Kac condition holds:

$$c \to \infty, \quad \lambda \to \infty, \quad \frac{c^2}{\lambda} \to \rho, \quad \rho > 0. \tag{11}$$

Under this condition (11), the transition density $f = f(\mathbf{x}, t)$, $\mathbf{x} \in \mathbb{R}^m$, $m \geq 2$, $t > 0$, of the isotropic transport process $\mathbf{X}(t)$ converges to the transition density of the homogeneous Brownian motion with zero drift and diffusion coefficient $\sigma^2 = 2\rho/m$ (Kolesnik 2008), i.e.,

$$\lim_{\substack{c,\,\lambda\to\infty \\ (c^2/\lambda)\to\rho}} f(\mathbf{x}, t) = \left(\frac{m}{4\rho\pi t}\right)^{m/2} \times \exp\left(-\frac{m}{4\rho t}\,\|\mathbf{x}\|^2\right), \qquad m \geq 2,$$

where $\|\mathbf{x}\|^2 = x_1^2 + \cdots + x_m^2$.

Some of these results are also valid for the transport processes with arbitrary scattering functions. Suppose that both the initial and each new direction are taken on according to some arbitrary distribution on the unit sphere $S_1^m \subset \mathbb{R}^m$, $m \geq 2$. Let $\chi(\mathbf{x})$, $\mathbf{x} \in S_1^m$ denote the density of this distribution, assumed to exist. Introduce the function

$$\psi(t) = \int_{S_1^m} e^{ict(\boldsymbol{\alpha}, \mathbf{x})}\,\chi(\mathbf{x})\,\mu(d\mathbf{x}),$$

where $\mu(d\mathbf{x})$ is the Lebesgue measure on S_1^m. Then the characteristic function of such a transport process satisfies a Volterra integral equation similar to (8), in which the function $\varphi(t)$ is replaced everywhere by the function $\psi(t)$. The unique continuous solution of such an equation is similar to (9) with the same replacement.

About the Author

Alexander Kolesnik, Ph.D. in Probability and Statistics (1991) and Habilitation (2010), is a Leading Scientific Researcher (Professor). He has published more than 40 articles. He is currently preparing a monograph on the statistical theory of transport processes at finite velocity. He was coeditor (1996–2006) of InterStat (Electronic Journal on Probability and Statistics, USA), and external referee of many international journals on probability and statistics.

Cross References

►Brownian Motion and Diffusions
►Poisson Processes
►Stochastic Modeling Analysis and Applications
►Stochastic Modeling, Recent Advances in

References and Further Reading

Goldstein S (1951) On diffusion by discontinuous movements and on the telegraph equation. Q J Mech Appl Math 4:129–156
Kac M (1956) A stochastic model related to the telegrapher's equation. In: Some stochastic problems in physics and mathematics, Magnolia petroleum company colloquium, lectures in the pure and applied science, No. 2 (Reprinted in: Rocky Mount J Math (1974), 4:497–509)
Kolesnik AD (2006) A four-dimensional random motion at finite speed. J Appl Probab 43:1107–1118
Kolesnik AD (2008) Random motions at finite speed in higher dimensions. J Stat Phys 131:1039–1065

Kolesnik AD, Orsingher E (2005) A planar random motion with an infinite number of directions controlled by the damped wave equation. J Appl Probab 42:1168–1182
Korolyuk VS, Swishchuk AV (1994) Semi-Markov random evolutions. Kluwer, Amsterdam
Masoliver J, Porrá JM, Weiss GH (1993). Some two and three-dimensional persistent random walks. Physica A 193:469–482
Papanicolaou G (1975) Asymptotic analysis of transport processes. Bull Am Math Soc 81:330–392
Pinsky M (1991) Lectures on random evolution. World Scientific, River Edge
Stadje W (1987) The exact probability distribution of a two-dimensional random walk. J Stat Phys 46:207–216
Stadje W (1989) Exact probability distributions for non-correlated random walk models. J Stat Phys 56:415–435

Stochastic Processes

Rolando Rebolledo
Professor, Head of the Center for Stochastic Analysis,
Facultad de Matemáticas
Universidad Catolica de Chile, Santiago, Chile

The word "stochastic process" is derived from the Greek noun "stokhos" which means "aim." Another related Greek word "stokhastikos," "the dart game," provides an alternative image for randomness or chance. Although the concept of Probability is often associated with dice games, the dart game seems to be more adapted to the modern approach to both Probability Theory and Stochastic Processes. Indeed, the fundamental difference between a dice game and darts is that while in the first, one cannot control the issue of the game, in the dart game, one tries to attain an objective with different degrees of success, thus, the player increases his knowledge of the game at each trial. As a result, time is crucial in the dart game, the longer you play, the better you increase your skills.

Definition of a Stochastic Process

The mathematical definition of a stochastic process, in the Kolmogorov model of Probability Theory, is given as follows. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, that is, Ω is a non empty set called *sample space*, $\mathcal{F}$ is a sigma field of subsets of Ω, which represents the family of *events*, and $\mathbb{P}$ is a *probability measure* defined on $\mathcal{F}$. T is another non empty set, and $(E, \mathcal{E})$ a measurable space to represent all possible *states*. Then, a *stochastic process with states in E* is a map $X : T \times \Omega \to E$ such that for all $t \in T$, $\omega \mapsto X(t, \omega)$ is a measurable function. In other words, a primary interpretation of a stochastic process X is as a collection of random variables, and as such, notations like $(X_t)_{t\in T}$ are used to refer to X, that is $X_t(\omega) = X(t, \omega)$, for all $(t, \omega) \in T \times \Omega$. If T is an ordered number set, (e.g., $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}^+$, $\mathbb{R}$), it is often referred as the set of *time variables* and taken as a subset of integers or real numbers. For each $\omega \in \Omega$, the map $X(\cdot, \omega) : t \mapsto X(t, \omega)$ is called the *trajectory* of the process. Thus, each trajectory is an element of E^T, the set of all E-valued functions defined on T. Particularly, if T is a countable set, the process is said to be indexed by *discrete times* (the expression *Time Series* is also in use in this case). Discrete time stochastic processes were the first studied in Probability Theory under the name of *chains* (see ▶Markov Chains).

Example 1

1. Consider a sequence $(\xi_n)_{n\geq 1}$ of real random variables. According to the definition, this is a stochastic process. New stochastic processes can be defined on this basis. For instance, take $(S_n)_{n\geq 1}$, defined as, $S_n = \xi_1 + \ldots + \xi_n$, for each $n \geq 1$.

   Suppose now that the random variables $(\xi_n)_{n\geq 1}$ are independent and identically distributed on $\{-1, 1\}$ with $\mathbb{P}(\xi = \pm 1) = 1/2$. Then, $(S_n)_{n\geq 1}$ becomes a *Simple Symmetric Random Walk*.
2. Consider a real function $x : [0, \infty[\to \mathbb{R}$, this is also a stochastic process. It suffices to consider any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and define $X(\omega, t) = x(t)$, for all $\omega \in \Omega$, $t \geq 0$. This is a trivial stochastic process.
3. Consider an initial value problem given by

$$\begin{cases} x' = f(t, x); \\ x(0) = x, \end{cases} \tag{1}$$

   where f is a continuous function on the two variables (t, x). Newtonian Mechanics can be written within this framework, which is usually referred as a mathematical model for a *closed dynamical system* in Physics. That is, the system has no interaction with the environment, and time is reversible. Now define Ω as the set of all continuous functions from $[0, \infty[$ into $\mathbb{R}$. Endow Ω with the topology of uniform convergence on compact subsets of the positive real line and call $\mathcal{F}$ the corresponding Borel σ-field. Thus, any $\omega \in \Omega$ is a function $\omega = (\omega(t); t \geq 0)$. Define the stochastic process $X(\omega, t) = \omega(t)$, known as the *canonical process*. The initial value problem is then written as

$$X(\omega, t) = x + \int_0^t f(s, X(\omega, s))ds. \tag{2}$$

   This can be phrased as an example of a *Stochastic Differential Equation*, without noise term. The solution

is a deterministic process which provides a description of the given closed dynamical system. Apparently, there is no great novelty and one can wonder whether the introduction of Ω is useful. However, this framework includes processes describing open dynamical systems too, embracing the interaction of the main system with the environment, and that is an important merit of the stochastic approach. Typically, the interaction of a given system with the environment is described through the action of so-called noises interfering with the main dynamics. Let us complete our example adding a noise term to the closed dynamics.

To consider the action of a *noise*, take a sequence $(\xi_n)_{n\geq 1}$ of random variables defined on Ω, such that $\xi_n(\omega) \in \{-1,1\}$. Let be given a probability $\mathbb{P}$ on the measurable space $(\Omega,\mathcal{F})$ such that $\mathbb{P}(\xi_n = \pm 1) = 1/2$. Call $S_n = \xi_1 + \ldots + \xi_n$ and denote $[t]$ the greatest integer $\leq t$. The equation

$$X(\omega,t) = x + \int_0^t f(s,X(\omega,s))ds + S_{[t]}(\omega), \qquad (3)$$

is an example of a stochastic differential equation driven by a ▶random walk. The stochastic process obtained as a solution is no longer deterministic and describes an open system dynamics. ▽

Distributions

The space of trajectories E^T is usually endowed with the product σ-field $\mathcal{E}^{\otimes T}$ generated by all projections $\pi_t : E^T \to E$, which associate to each function $x \in E^T$ its value $x(t) \in E$, $t \in T$. Thus, a stochastic process is, equivalently, a random variable $X : \Omega \to E^T$, $\omega \mapsto X(\cdot,\omega)$. The *Law* or *Probability Distribution* P_X of a stochastic process X is the image of the probability $\mathbb{P}$ on the measurable space $(E^T,\mathcal{E}^{\otimes T})$ of all trajectories. Given a probability measure P on the space $(E^T,\mathcal{E}^{\otimes T})$, one may construct a *Canonical Process* X whose distribution P_X coincides with P. Indeed, it suffices to consider $\Omega = E^T$, $\mathcal{F} = \mathcal{E}^{\otimes T}$, $\mathbb{P} = P$, $X(t,\omega) = \omega(t)$, for each $\omega = (\omega(s);\ s \in T) \in E^T$, $t \in T$.

Let a finite set $I = \{t_1,\ldots,t_n\} \subset T$ be given, and denote π_I the canonical projection defined on E^T with values in E^I, such that $x \mapsto (x(t_1),\ldots,x(t_n))$. Call $\mathcal{P}_f(T)$ the family of all finite subsets of T. The *Finite Dimensional Distributions* or *Marginal Probability Distributions* of an E-valued stochastic process is the family $(P_{X,I})_{I\in\mathcal{P}_f(T)}$ of distributions, where $P_{X,I}$ is defined as

$$P_{X,I}(A) = P_X\left(\pi_i^{-1}(A)\right) = \mathbb{P}\left((X(t_1,\cdot),\ldots,X(t_n,\cdot)) \in A\right), \qquad (4)$$

for all $A \in \mathcal{E}^{\otimes I}$.

Example 2

1. A *Poisson Process* $(N_t)_{t\geq 0}$ is defined as a stochastic process with values in $\mathbb{N}$ such that
   (a) $N_0(\omega) = 0$ and $t \mapsto N_t(\omega)$ is increasing, for all $\omega \in \mathbb{N}$.
   (b) For all $0 \leq s \leq t < \infty$, $N_t - N_s$ is independent of $(N_u;\ u \leq s)$.
   (c) For all $0 \leq s \leq t < \infty$, the distribution of $N_t - N_s$ is Poisson with parameter $t - s$, that is

$$\mathbb{P}(N_t - N_s = k) = \frac{(t-s)^k}{k!}e^{-(t-s)}.$$

2. A d-dimensional *Brownian Motion* (see also ▶Brownian Motion and Diffusions) is a stochastic process $(B_t)_{t\geq 0}$, taking values in $\mathbb{R}^d$ such that:
   (a) If $0 \leq s < t < \infty$, then $B_t - B_s$ is independent of $(B_u;\ u \leq s)$.
   (b) If $0 \leq s < t < \infty$, then

$$\mathbb{P}(B_t - B_s \in A) = (2\pi(t-s))^{-d/2}\int_A e^{-|x|^2/2(t-s)}dx,$$

   where dx represents the Lebesgue measure on $\mathbb{R}^d$ and $|x|$ is the euclidian norm in that space.

   The Brownian Motion starts at x if $\mathbb{P}(B_0 = x) = 1$. ▽

Construction of Canonical Processes

An important problem in the construction of a canonical stochastic process given the family of its finite dimensional distributions was solved by Kolmogorov in the case of a countable set T and extended to continuous time later by several authors. At present, a particular case, general enough for applications, is the following version of the Daniell–Kolmogorov Theorem. Suppose that E is a Polish space (complete separable metric space) and let $\mathcal{E}$ be its Borel σ-field. Let T be a subset of $\mathbb{R}^+$. Suppose that for each $I \in \mathcal{P}_f(T)$ a probability P_I is given on the space $(E,\mathcal{E}^{\otimes I})$. Then, there exists a probability P on $(E^T,\mathcal{E}^{\otimes T})$ such that for all $I \in \mathcal{P}_f(T)$,

$$P_I(A) = P \circ \pi_I^{-1}(A) = P\left(\pi_I^{-1}(A)\right), \qquad (5)$$

for all $A \in \mathcal{E}^{\otimes I}$, if and only if the following *Consistency Condition* is satisfied:

$$P_I = P_J \circ \pi_{J,I}^{-1}, \qquad (6)$$

for all $I, J \in \mathcal{P}_f(T)$ such that $I \subset J$, where $\pi_{J,I}$ denotes the canonical projection from the space E^J onto E^I.

Example 3 Consider $J = \{t_1, \ldots, t_n\}$ and let Φ_t be the normal distribution of mean zero and variance $t \geq 0$, that is,

$$\Phi_t(A) = (2\pi t)^{-1/2} \int_A e^{-x^2/2t} dx.$$

Let $P_J = \Phi_{t_1} \otimes \Phi_{t_2 - t_1} \otimes \ldots \Phi_{t_n - t_{n-1}}$, that is for all Borel sets $A_1, \ldots, A_n$,

$$P_J(A_1 \times A_2 \times \ldots \times A_n) = \Phi_{t_1}(A_1)\Phi_{t_2-t_1}(A_2) \ldots \Phi_{t_n - t_{n-1}}(A_n).$$

This is a probability on $\mathbb{R}^n$. Take $I = \{t_1, \ldots, t_{n-1}\}$. Notice that $\pi_{J,I}^{-1}(A_1 \times \ldots \times A_{n-1}) = A_1 \times \ldots \times A_{n-1} \times \mathbb{R}$, thus

$$\begin{aligned} P_I(A_1 \times A_2 \times \ldots \times A_{n-1}) &= \Phi_{t_1}(A_1)\Phi_{t_2-t_1}(A_2) \ldots \\ &\quad \Phi_{t_{n-1}-t_{n-2}}(A_{n-1}) \\ &= P_J(A_1 \times A_2 \times \ldots \times \mathbb{R}). \end{aligned}$$ ▽

Regularity of Trajectories

Another interpretation of a stochastic process is based on regularity properties of trajectories. Indeed, if one knows that each trajectory belongs almost surely to a function space $S \subset E^T$, endowed with a σ-field $\mathcal{S}$, one may provide another characterization of the stochastic process X as an S-valued random variable, $\omega \mapsto X(\cdot, \omega)$ defined on Ω.

Regarding the regularity, Kolmogorov first proved one of the most useful criteria on continuity of trajectories. Suppose that $X = (X(t,\omega);\ t \in [0,1],\ \omega \in \Omega)$ is a real-valued stochastic process and assume that there exist $\alpha, \delta > 0$ and $0 < C < \infty$ such that

$$\mathbb{E}\left(|X(t+h) - X(t)|^\alpha\right) < C|h|^{1+\delta}, \tag{7}$$

for all $t \in [0,1]$ and all sufficiently small $h > 0$, then X has continuous trajectories with probability 1. Therefore, if X satisfies (7), then there exists a random variable $\widetilde{X} : \Omega \to C[0,1]$, where $C[0,1]$ is the metric space of real continuous functions defined on $[0,1]$, endowed with the metric of uniform distance, such that $\mathbb{P}\left(\{\omega \in \Omega :\ X(\cdot,\omega) = \widetilde{X}(\omega)\}\right) = 1$.

Wiener Measure, Brownian Motion

The above result is crucial to construct the *Wiener Measure* on the space $C[0,1]$ or, more generally, on $C(\mathbb{R}^+)$, which is the law of the *Brownian Motion* (see also ►Brownian Motion and Diffusions). Indeed, by means of Kolmogorv's Consistency Theorem, one first constructs a probability measure P on the product space $(\mathbb{R}^{\mathbb{R}^+}, \mathcal{B}(\mathbb{R})^{\otimes \mathbb{R}^+})$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ-field of $\mathbb{R}$, considering the consistent family of probability distributions

$$P_I = \Phi_{t_1} \otimes \Phi_{t_2 - t_1} \otimes \ldots \otimes \Phi_{t_n - t_{n-1}}, \tag{8}$$

where $I = \{t_1, \ldots, t_n\}$, and Φ_t denotes the normal distribution with mean 0 and variance t. Since the family $(P_I)_{I \in \mathcal{P}_f(\mathbb{R}^+)}$ is consistent, there exists a unique P probability measure on $(\mathbb{R}^{\mathbb{R}^+}, \mathcal{B}(\mathbb{R})^{\otimes \mathbb{R}^+})$ such that $P_I = P \circ \pi_I^{-1}$. One can construct the canonical process with law P which should correspond to the Brownian Motion. Unfortunately, the set of real-valued continuous functions defined on $\mathbb{R}^+$ is not an element of $\mathcal{B}(\mathbb{R})^{\otimes \mathbb{R}^+}$. However, thanks to (7) one proves that the exterior probability measure P^* defined by P is concentrated on the subset $C(\mathbb{R}^+)$ of $\mathbb{R}^{\mathbb{R}^+}$ thus, the restriction P_W of P^* to $C(\mathbb{R}^+)$ gives the good definition of Wiener Measure. Thus, a canonical version of the Brownian Motion is given by the canonical process on the space $C(\mathbb{R}^+)$.

Series Expansion in L^2

In the early years of the Theory of Stochastic Processes, a number of authors, among them Karhunen and Loève, explored other regularity properties of trajectories, deriving some useful representations by means of series expansions in an L^2 space. More precisely, let $T \in \mathcal{B}(\mathbb{R}^+)$ be given and call $\mathfrak{h} = L^2(T)$ the Hilbert space of all real-valued Lebesgue-square integrable functions defined on T. Suppose that all trajectories $X(\cdot,\omega)$ belong to $\mathfrak{h}$ for all $\omega \in \Omega$, and denote $(e_n)_{n \in \mathbb{N}}$ an orthonormal basis of $\mathfrak{h}$. Therefore, $x_n(\omega) = \langle X(\cdot,\omega), e_n \rangle$ satisfies $\sum_{n \in \mathbb{N}} |x_n(\omega)|^2 < \infty$, for all $\omega \in \Omega$. And the series

$$\sum_{n \in \mathbb{N}} x_n(\omega) e_n, \tag{9}$$

converges in $\mathfrak{h}$, providing a representation of $X(\cdot,\omega)$. So that, by an abuse of language one can represent $X(t,\omega)$ by $\sum_{n \in \mathbb{N}} x_n(\omega) e_n(t)$.

Example 4 Consider $T = [0,1]$ and the Haar orthonormal basis on the space $\mathfrak{h} = L^2([0,1])$ constructed by induction as follows: $e_1(t) = 1$ for all $t \in [0,1]$;

$$e_{2^m+1} = \begin{cases} 2^{m/2}, & \text{if } 0 \leq t < 2^{-m-1}, \\ -2^{m/2}, & \text{if } 2^{-m-1} \leq t < 2^{-m}, \\ 0, & \text{otherwise.} \end{cases}$$

And finally, define $e_{2^m+j}(t) = e_{2^m+1}\left(t - 2^{-m}(j-1)\right)$, for $j = 1, \ldots, 2^m$, $m = 0, 1, \ldots$. Given a sequence $(b_n)_{n \geq 1}$ of independent standard normal random variables (that is, with distribution $\mathcal{N}(0,1)$), the $L^2(\Omega \times [0,1])$-convergent series $\sum_{n \geq 1} b_n(\omega) f_n(t)$ provides a representation of the Brownian Motion $(B_t)_{t \in [0,1]}$, where $f_n(t) = \int_0^t e_n(s) ds$, $(t \in [0,1],\ n \in \mathbb{N})$. ▽

The General Theory of Processes

The General Theory of Processes emerged in the seventies as a contribution of the Strasbourg School initiated by Paul André Meyer. This Theory uses the concept of a *History* or *Filtration*, which consists of an increasing family of σ-fields $\mathbb{F} = (F_t)_{t\in T}$, where T is an ordered set, $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$ for all $s \le t$. Thus, a stochastic process X is *adapted* to $\mathbb{F}$ if for all $t \in T$, the variable $X(t,\cdot)$ is $\mathcal{F}_t/\mathcal{E}$–measurable. Stronger measurability conditions mixing regularity conditions have been introduced motivated by the construction of stochastic integrals and the modern theory of Stochastic Differential Equations. Let $T = \mathbb{R}^+$ and assume E to be a Polish space endowed with the σ-field of its Borel sets. Denote $C_E = C(\mathbb{R}^+, E)$ (respectively $D_E = D(\mathbb{R}^+, E)$) the space of all E-valued continuous functions defined on $\mathbb{R}^+$ to E (resp. the space of all E-valued functions which have left hand limit at each point $t > 0$ and are right-continuous at $t > 0$, endowed with the Skorokhod's topology). Consider now the family $\mathcal{C}_E$ (resp. $\mathcal{D}_E$) of all $\mathbb{F}$-adapted stochastic processes $X : \mathbb{R}^+ \times \Omega \to E$ such that their trajectories belong to C_E (resp. to D_E). The *Predictable* (resp. *Optional*) σ-field on the product set $\mathbb{R}^+ \times \Omega$ is the one generated by $\mathcal{C}_E$ (resp. $\mathcal{D}_E$), that is $\mathcal{P} = \sigma(\mathcal{C}_E)$, (resp. $\mathcal{O} = \sigma(\mathcal{D}_E)$). Then, a process X is *predictable* (resp. *optional*) if $(t,\omega) \mapsto X(t,\omega)$ is measurable with respect to $\mathcal{P}$, (resp. $\mathcal{O}$). A crucial notion in the development of this theory is that of *Stopping Time*: a function $\tau : \Omega \to [0,\infty]$ is a stopping time if for all $t > 0$, $\{\omega \in \Omega : \tau(\omega) \le t\}$ is an element of the σ-field $\mathcal{F}_t$. This definition is equivalent to say that τ is a stopping time if and only if $(t,\omega) \mapsto 1_{[0,\tau(\omega)[}(t)$ is an optional process, where the notation 1_A is used for the indicator or characteristic function of a set A.

The development of the General Theory of Processes encountered at least two serious difficulties which could not be solved in the framework of Measure Theory and required a use of Capacity Theory. They are the *Section Theorem* and the *Projection Theorem*. The Section Theorem asserts that if the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is complete (that is $\mathcal{F}$ contains all $\mathbb{P}$-null sets) and $A \in \mathcal{O}$, then there exists a stopping time τ such that its graph is included in A. And the Projection Theorem states that given an optional set $A \subset \mathbb{R}^+ \times \Omega$, the projection $\pi(A)$ on Ω belongs to the complete σ-field $\mathcal{F}$. For instance, this result allows to prove that given a Borel set B of the real line, the random variable $\tau_B(\omega) = \inf\{t \ge 0 : X(t,\omega) \in B\}$ ($\inf \emptyset = \infty$), defines a stopping time for an $\mathbb{F}$-adapted process X with trajectories in D almost surely, provided the filtration $\mathbb{F}$ is right-continuous, that is, for all $t \ge 0$, $\mathcal{F}_t = \mathcal{F}_{t+} := \bigcap_{s>t} \mathcal{F}_s$, and in addition each σ-field contains all $\mathbb{P}$-null sets. Within this theory, the system $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}^+}, \mathbb{P})$ is usually called a *Stochastic Basis* and a system $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in T}, E, \mathcal{E}, \mathbb{P}, (X_t)_{t\in T})$ provides the whole structure needed to define an E-valued adapted stochastic process.

Attending to measurability properties only, stochastic processes may be classified as optional or predictable, as mentioned before, for which no probability is needed. However, richer properties of processes strongly depend on the probability considered in the stochastic basis. For instance, the definitions of *martingales*, *submartingales*, *supermartingales*, *semimartingales* depend on a specific probability measure, through the concept of *conditional expectation*. Let us mention that *semimartingales* form the most general class of possible integrands to give a rigorous meaning to *Stochastic Integrals* and *Stochastic Differential Equations*.

Probability is moreover fundamental for introducing concepts as *Markov Process* (see ▶Markov Processes), *Gaussian Process*, *Stationary Sequence and Stationary Process*.

Extensions of the Theory

Extensions to the theory have included changing either the nature of T to consider *Random Fields*, where $t \in T$ may have the meaning of a space label (T is no more a subset of the real line), or the state space E, to deal for instance with measure-valued processes, or random distributions.

Example 5 Let $(T, \mathcal{T}, \nu)$ be a σ-finite measure space, and $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space. Call $\mathcal{T}_\nu$ the family of all sets $A \in \mathcal{T}$ such that $\nu(A) < \infty$. A *Gaussian white noise* based on ν is a random set function W defined on $\mathcal{T}_\nu$ and values in $\mathbb{R}$ such that

(a) $W(A)$ is centered Gaussian and $\mathbb{E}\left(W(A)^2\right) = \nu(A)$, for all $A \in \mathcal{T}_\nu$;
(b) If $A \cap B = \emptyset$, then $W(A)$ and $W(B)$ are independent.

In particular, if $T = \mathbb{R}^{+2}$, $\mathcal{T}$ the corresponding Borel σ-field, and $\nu = \lambda$ the product Lebesgue measure, define $B_{t_1,t_2} = W(]0,t_1] \times]0,t_2])$, for all $(t_1,t_2) \in T$. The process $(B_{t_1,t_2})_{(t_1,t_2)\in T}$ is called the *Brownian sheet*. ▽

Going further, on the state space E consider the algebra $\mathfrak{E}$ of all bounded $\mathcal{E}$-measurable complex-valued functions. Then, to each E-valued stochastic process X one associates a family of maps $j_t : \mathfrak{E} \to L^\infty_{\mathbb{C}}(\Omega, \mathcal{F}, \mathbb{P})$, where $j_t(f)(\omega) = f(X(t,\omega))$, for all $t \ge 0$, $\omega \in \Omega$. The family $(j_t)_{t\in\mathbb{R}^+}$, known as the *Algebraic Flow* can be viewed as a family of complex random measures (each j_t is a Dirac measure supported by $X(t,\omega)$) or, better, as a *-homomorphism between the two *-algebras $\mathfrak{E}$, $L^\infty_{\mathbb{C}}(\Omega, \mathcal{F}, \mathbb{P})$, the * operation being here the

complex conjugation. The stochastic process is completely determined by the algebraic flow $(j_t)_{t\in\mathbb{R}^+}$.

Example 6 Consider a Brownian motion B defined on a stochastic basis $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathbb{R}^+}, \mathbb{P})$, with states in $\mathbb{R}$, and call $\mathfrak{B}$ the algebra of bounded complex valued Borel function defined on the real line. $\mathfrak{B}$ is a *-algebra of functions, that is, there exists an involution * (the conjugation), such that $f \mapsto f^*$ is antilinear and $(fg)^* = g^* f^*$, for all $f, g \in \mathfrak{B}$. The algebraic flow associated to B is given by $j_t(f) = f(B_t)$, for all $t \geq 0$, and any $f \in \mathfrak{B}$, that is $j_t : \mathfrak{B} \to L^\infty_{\mathbb{C}}(\Omega, \mathcal{F}, \mathbb{P})$. If $\mathbb{P}(B_0 = x) = 1$, then $j_0(f) = f(x)$ almost surely. Moreover, notice that Itô's formula implies that for all bounded f of class C^2, it holds

$$j_t(f) = f(x) + \int_0^t j_s\left(\frac{d}{dx}f\right) dB_s + \int_0^t j_s\left(\frac{1}{2}\frac{d^2}{dx^2}f\right) ds. \quad \triangledown$$

Algebraic flows provide a suitable framework to deal with more generalized evolutions, like those arising in the description of *Open Quantum System Dynamics*, where the algebras are non commutative. Thus, given two unital *-algebras (possibly non commutative) $\mathfrak{A}, \mathfrak{B}$, a notion of *Algebraic Stochastic Process* is given by a flow $(j_t)_{t\in\mathbb{R}^+}$, where $j_t : \mathfrak{B} \to \mathfrak{A}$ is a *-homomorphisms, for all $t \geq 0$. That is, each j_t is a linear map, which satisfies $(j_t(b))^* = j_t(b^*)$, $j_t(a^* b) = j_t(a)^* j_t(b)$, for all $a, b \in \mathfrak{B}$, and $j_t(\mathbf{1}_{\mathfrak{B}}) = \mathbf{1}_{\mathfrak{A}}$, where $\mathbf{1}_{\mathfrak{A}}$ (resp. $\mathbf{1}_{\mathfrak{B}}$) is the unit of $\mathfrak{A}$ (resp. $\mathfrak{B}$).

The Dawning of Stochastic Analysis as a Pillar of Modern Mathematics

These days, Stochastic Processes provide the better description of complex evolutionary phenomena in Nature. Coming from our understanding of the macro world, through our everyday life, exploring matter at its smallest component, stochastic modeling has become fundamental. In other words, stochastic processes have become influential in all sciences, namely, in biology (population dynamics, ecology, neurosciences), computer science, engineering (especially electric and operation research), economics (via finance), physics, among others. The new branch of Mathematics, known as Stochastic Analysis, is founded on stochastic processes. Stochastics is invading all branches of Mathematics: Combinatorics, Graph Theory, Partial and Ordinary Differential Equations, Group Theory, Dynamical Systems, Geometry, Functional Analysis, among many other specific subjects. The dawning of Stochastic Analysis era is a fundamental step in the evolution of human understanding of Chance as a natural interconnection and interaction of matter in Nature. This has been a long historical process which started centuries ago with the dart game.

Acknowledgments

The author is gratefully indebted with a number of anonymous referees for heartening support. Their comments were fundamental to improve the first version of this contribution. Also, no symphony orchestra could sound appropriately with no experimented conductor. The hearted conductor of this encyclopedia has been Professor Miodrag Lovric to whom I express my deep gratitude for his efficient and courageous work.

This work received partial support of grant PBCT-ADI13 of the Chilean Science and Technology Bicentennial Foundation.

About the Author

The following bibliography is nothing but a very small sample of references on stochastic processes, which could be termed classic, as well as more recent textbooks. General references as well as specialized books on the field are fast increasing, following the success of stochastic modeling, and one can be involuntarily and easily unfair by omitting outstanding authors.

Rolando Rebolledo obtained his "Doctorat d'État" at the Université Pierre et Marie Curie (Paris VI), France, in 1979. He is Professor at the Faculty of Mathematics, and Head of the Center for Stochastic Analysis, Pontificia Universidad Católica de Chile. He was President of the Sociedad de Matemática de Chile during five periods (1982–1985, 1994–1995, 1995–1998). He was Chairman of the Latin American Regional Committee of the Bernoulli Society (1989–1993), member of the Council of that Society and Scientific Secretary of the Committee for the Year 2000 of the Bernoulli Society. He chaired the Commission on Development and Exchanges of the International Mathematical Union (1994–1998, 1998–2002). Professor Rebolledo is a member of the American Mathematical Society, Bernoulli Society, Fellow of the International Statistics Institute since 1994. He has been twice awarded with the "Presidential Chair" in Chile (1995–1998, 1999–2002), and with the Medal of the Catholic University for outstanding research achievements (1996 and 1999). Dr. Rebolledo has published over 80 research papers, and edited five Proceedings of the International ANESTOC Workshops. Rolando Rebolledo has been Visiting Professor at many universities all over the world, including Denmark, Germany, Brazil, Venezuela, Italy, France, Russia, Australia, USA, and Portugal.

Cross References

►Brownian Motion and Diffusions
►Extremes of Gaussian Processes
►Gaussian Processes

►Lévy Processes
►Markov Chains
►Markov Processes
►Martingales
►Point Processes
►Poisson Processes
►Random Walk
►Renewal Processes
►Sampling Problems for Stochastic Processes
►Statistical Inference for Stochastic Processes
►Stochastic Differential Equations
►Stochastic Processes: Applications in Finance and Insurance
►Stochastic Processes: Classification

References and Further Reading

Accardi L, Lu YG, Volovich I (2002) Quantum theory and its stochastic limit. Springer, Berlin
Bhattacharya R, Waymire EC (2007) A basic course in probability theory. Springer Universitext, New York
Bhattacharya R, Waymire EC (2009) Stochastic processes with applications. SIAM Classics in Applied Mathematics, Philadelphia
Dellacherie C (1972) Capacités et processus stochastiques. Springer, New York
Dellacherie C, Meyer PA (1978–1987) Probabilités et potentiel, vols 1–4. Hermann, Paris
Doob JL (1953) Stochastic processes. Wiley, New York
Dynkin EB (1965) Markov processes. Springer, Berlin (Translated from Russian)
Ethier K, Kurtz TG (1986) Markov processes: characterization and convergence. Wiley, New York
Feller W (1966) An introduction to probability theory and its applications, vol 2. Wiley, New York
Gikhman II, Skorokhod AV (1974–1979) Theory of stochastic processes, vol 1–3. Springer, Berlin (Translated from Russian)
Itô K (2006) Essentials of stochastic processes. American Mathematical Society, Providence
Karatzas I, Shreve SE (1991) Brownian motion and stochastic calculus. Springer, New York
Lévy P (1965) Processus stochastiques et mouvement Brownien. Gauthier-Villars, Paris
Meyer PA (1966) Probability and potentials. Ginn-Blaisdell, Boston
Meyer PA (1993) Quantum probability for probabilists. Lecture notes in mathematics, vol 1538, Springer, Berlin
Neveu J (1975) Discrete-parameter martingales. North-Holland, Amsterdam; American Elsevier, New York
Parthasarathy KR (1992) An introduction to quantum stochastic calculus. Birkhaüser, Basel
Protter P (1990) Stochastic integration and differential equations: a new approach. Springer, Berlin
Rebolledo R (2006) Complete positivity and the Markov structure of open quantum systems, in open quantum systems II. Lecture notes in mathematics, 1882, pp 149–182
Varadhan SRS (2007) Stochastic processes. Courant lectures notes in mathematics, vol 16. American Mathematical Society, New York

Stochastic Processes: Applications in Finance and Insurance

Leda D. Minkova
Associate Professor, Faculty of Mathematics and Informatics
Sofia University "St. Kl. Ohridski", Sofia, Bulgaria

The applications of ►stochastic processes and martingale methods (see ►Martingales) in finance and insurance have attracted much attention in recent years.

Martingales in Finance

Let us consider a continuous time arbitrage free financial market with one risk-free investment (bond) and one risky asset (stock). All processes are assumed to be defined on the complete probability space $(\Omega, \mathcal{F}_T, (\mathcal{F}_t), P)$ and adapted to the filtration $(\mathcal{F}_t)$, $t \le T$. The bond yields a constant rate of return $r \ge 0$ over each time period. The risk-free bond represents an accumulation factor and its price process B equals

$$dB_t = rB_t dt,\ t \in [0, T],\ B_0 = 1, \tag{1}$$

or $B_t = e^{rt}$. The evolution of the stock price S_t is described by the linear stochastic differential equation

$$dS_t = S_t(\mu dt + \sigma dW_t),\ t \in [0, T],\ S_0 = S, \tag{2}$$

where the expected rate of return μ and the volatility coefficient σ are constants. The stochastic process W_t, $t \ge 0$ is a one-dimensional Brownian motion. The solution of Eq. 2 is given by

$$S_t = S \exp\left(\sigma W_t + \left(\mu - \frac{\sigma^2}{2}\right)t\right),\ t \in [0, T]. \tag{3}$$

The process (3) is considered by Samuelson (1965) and is called a geometric Brownian motion. The market with two securities is called a standard diffusion (B, S) market and is suggested by F. Black and M. Scholes (1973). The references are given in Shiryaev (1999) and Rolski et al. (1999).

A European call (put) option, written on risky security gives its holder the right, but not obligation to buy (sell) a given number of shares of a stock for a fixed price at a future date T. The exercise date T is called maturity date and the price K is called a strike price. The problem of option pricing is to determine the value to assign to the option at a time $t \in [0, T]$. The writer of the option has to calculate the fair price as the smallest initial investment that would

allow him to replicate the value of the option throughout the time T. The replication portfolio can be used to hedge the risk inherent in writing the option.

Definition 1 **(Martingale measure)** *A probability measure $\overline{P}$ defined on $(\Omega, \mathcal{F}_T)$ is called a martingale measure if it is equivalent to P ($\overline{P} \sim P$) and the discounted process $\overline{S}_t = S_t B_t^{-1}$ is a $\overline{P}$– local martingale.*

For the Black–Scholes model, the martingale measure is unique and is defined by the following theorem of Girsanov type.

Theorem 1 *The unique martingale measure $\overline{P}$ is given by the Radon–Nikodym derivative*

$$\frac{d\overline{P}}{dP} = \exp\left(-\frac{\mu - r}{\sigma} W_T - \frac{1}{2}\left(\frac{\mu - r}{\sigma}\right)^2 T\right), \; P - a.s.$$

Under the martingale measure, $\overline{P}$, the discounted stock price $\overline{S}_t$ satisfies the equation

$$d\overline{S}_t = \sigma \overline{S}_t d\overline{W}_t, \; t \geq 0,$$

where

$$\overline{W}_t = W_t + \frac{\mu - r}{\sigma} t, \; t \leq T$$

is a standard Brownian motion (see ►Brownian Motion and Diffusions) with respect to the measure $\overline{P}$.

The new probability measure $\overline{P}$ is called also a *risk-neutral measure*. The ratio $\frac{\mu - r}{\sigma}$ is called a *market price of risk*.

Consider a European call option written on a stock S_t, with exercise date T and strike price K. If we assume that the price of a stock is described by (2) and the payoff function is $f_T = \max(S_T - K, 0)$, then the fair price C_t of the European call option at time t is given by the famous Black–Scholes formula Black F, Scholes M (1973).

Theorem 2 **(Black–Scholes formula)** *The value C_t at time t of the European call option is given by*

$$C_t = S_t \Phi(d_1) - K e^{-r(T-t)} \Phi(d_2), \; t \leq T$$

where

$$d_1 = \frac{\log\left(\frac{S_t}{K}\right) + (T-t)\left(r + \frac{\sigma^2}{2}\right)}{\sigma\sqrt{T-t}},$$

$$d_2 = \frac{\log\left(\frac{S_t}{K}\right) + (T-t)\left(r - \frac{\sigma^2}{2}\right)}{\sigma\sqrt{T-t}} = d_1 - \sigma\sqrt{T-t}$$

and Φ is the standard Gaussian cumulative distribution function.

Insurance Risk Model

The standard model of an insurance company, called *risk process* $\{X(t), t \geq 0\}$ is given by

$$X(t) = ct - \sum_{k=1}^{N(t)} Z_k, \quad \left(\sum_{1}^{0} = 0\right). \tag{4}$$

Here c is a positive real constant representing the *risk premium* rate. The sequence $\{Z_k\}_{k=1}^{\infty}$ of mutually independent and identically distributed random variables, with common distribution function F, $F(0) = 0$, and mean value μ, is independent of the *counting process* $N(t)$, $t \geq 0$. The process $N(t)$ is interpreted as the number of claims on the company during the interval $[0, t]$. In the classical risk model, also called the Cramér–Lundberg model, the process $N(t)$ is a homogeneous Poisson process (see ►Poisson Processes), see for instance Grandell (1991). The ruin probability of a company with initial capital $u \geq 0$ is given by

$$\Psi(u) = P\left(u + X(t) < 0 \text{ for some } t > 0\right).$$

The martingale techniques have been introduced by H. Gerber in 1973 (see Gerber 1979). Since then, the martingale approach is a basic tool in risk theory (see the References in Schmidli (1996), Rolski et al. (1999), and Embrechts et al. (1997)).

Under the net profit condition $\theta = \frac{c}{\lambda\mu} - 1 > 0$, the following fundamental result holds (Embrechts et al. 1997).

Theorem 3 **(Cramér–Lundberg theorem)** *Assume that there exists $R > 0$ such that*

$$\int_0^{\infty} e^{Rx} dF_I(x) = 1 + \theta, \tag{5}$$

where $F_I(x) = \int_0^x (1 - F(y))dy$ is the integrated tail distribution of F.

a) *For all $u \geq 0$,*

$$\Psi(u) \leq e^{-Ru}; \tag{6}$$

b) $\lim_{u\to\infty} e^{Ru}\Psi(u) = \left[\frac{R}{\theta\mu}\int_0^{\infty} xe^{Rx}(1-F(x))dx\right]^{-1} < \infty$, *provided that*

$$\int_0^{\infty} xe^{Rx}(1 - F(x))dx < \infty.$$

c)

$$1 - \Psi(u) = \frac{\theta}{1+\theta}\sum_{n=0}^{\infty}\left(\frac{1}{1+\theta}\right)^n F_I^{*n}(u). \tag{7}$$

The condition (5) is known as the *Cramér condition*. Inequality (6) is called the *Lundberg inequality* and the constant R is the adjustment coefficient or *Lundberg exponent* (see Grandell 1991). Formula (7) is known as Pollaczek–Khinchin formula.

Example 1 **(Exponentially Distributed Claims)** *Suppose that the claim sizes are exponentially distributed with parameter* μ, *that is* $F(z) = 1 - e^{-\frac{z}{\mu}}$, $z \geq 0, \mu > 0$.

In this case, $F_I(z)$ *is also an exponential distribution function and the solution of equation (5) is*

$$R = \frac{1}{\mu}\frac{\theta}{1+\theta}.$$

The Pollaczek–Khinchin formula (7) gives the ruin probability

$$\Psi(u) = \frac{1}{1+\theta}e^{-\frac{1}{\mu}\frac{\theta}{1+\theta}u}, \quad u \geq 0.$$

Cross References

►Brownian Motion and Diffusions
►Insurance, Statistics in
►Martingales
►Optimal Statistical Inference in Financial Engineering
►Radon–Nikodým Theorem
►Stochastic Processes
►Stochastic Processes: Classification
►Testing Exponentiality of Distribution

References and Further Reading

Black F, Scholes M (1973) The pricing of options and corporate liabilities. J Polit Econ 81:637–657

Embrechts P, Klüppelberg C, Mikosh T (1997) Modelling extremal events for insurance and finance. Springer, Berlin

Gerber HU (1979) An introduction to mathematical risk theory. S.S. Huebner Foundation, Wharton School, Philadelphia

Grandell J (1991) Aspects of risk theory. Springer, New York

Pliska SR (1997) Introduction to mathematical finance. Blackwell, Oxford

Rolski T, Schmidli H, Schmidt V, Teugels J (1999) Stochastic processes for insurance and finance. Wiley, Chichester

Samuelson PA (1965) Rational theory of warrant pricing. Ind Manag Rev 6:13–31

Schmidli H (1996) Martingales and Insurance Risk. In Eighth International Summer School on Probability Theory and Mathematical Statistics, pp 155–188

Shiryaev AN (1999) Essentials of stochastic finance: facts, models, theory. World Scientific, Singapore

Stochastic Processes: Classification

Venkatarama Krishnan
Professor Emeritus ECE
UMass Lowell, Lowell, MA, USA

Definitions

Let $\{\Omega, \mathcal{F}, P\}$ be a complete probability space where Ω is the *sample space*, $\mathcal{F}$ is the σ-field associated with the sample space containing all the null sets of Ω, and P is the probability measure defined on the field $\mathcal{F}$. Let $\{\mathbb{R}, \mathcal{R}\}$ be a measurable range space called the *state space*, where $\mathbb{R} \equiv (-\infty, \infty)$ is the real line and $\mathcal{R}$ is the σ-field associated with the real line $\mathbb{R}$. A *random variable* X is a function that assigns a rule of correspondence between each $\omega \in \Omega$ and each $x \in \mathbb{R}$. This correspondence will induce a probability measure P_X defined on the field $\mathcal{R}$. Thus, X maps the probability space $\{\Omega, \mathcal{F}, P\}$ to the probability range space $\{\mathbb{R}, \mathcal{R}, P_X\}$

$$X : \{\Omega, \mathcal{F}, P\} \longrightarrow \{\mathbb{R}, \mathcal{R}, P_X\}. \tag{1}$$

The distribution function $F_X(x)$ of X is given by

$$P\{\omega : X(\omega) \leq x\} = P\{X \leq x\} = F_X(x), \; x \in \mathbb{R} \tag{2}$$

and the density function $f_X(x)$, which may include impulse functions of x, is the derivative of $F_X(x)$.

The definition (see, e.g., Gikhman and Skorokhod 1996, p. 1 and 144) of a *stochastic* (or random) process requires a parameter set Θ and an increasing sequence of sub σ-fields $\{\mathcal{F}_\theta \subset \mathcal{F}, \theta \in \Theta\}$ called the *filtration* σ*-field* such that $\mathcal{F}_\zeta \subset \mathcal{F}_\theta$ for each $\{\theta, \zeta \in \Theta, \zeta < \theta\}$. The filtration σ-field is a consequence of the distinction between the uncertainty of the future and the knowledge of the past. The family $\{X(\theta), \mathcal{F}_\theta\}$ of random variables defined on the probability space $\{\Omega, \mathcal{F}, P\}$ will be called a *random function* if the parameter set Θ is arbitrary and a *stochastic process* if the parameter set Θ is the time set $\mathbb{T} \equiv (-\infty, \infty)$, and θ is interpreted as time t. Thus, $X(t) \in \mathcal{F}_t$ is a stochastic process that maps the probability space $\{\Omega, \mathcal{F}, P\}$ to the range space $\{\mathbb{R}, \mathcal{R}, P_X\}$ for every point $\omega \in \Omega$ and $t \in \mathbb{T}$. $X(t)$ is said to be *adapted* to the filtration field $\{\mathcal{F}_t, t \in \mathbb{T}\}$ if $X(t)$ is $\mathcal{F}_t$-measurable in the sense the inverse image set $\{X(t)^{-1}[\mathbb{B}]\} \in \mathcal{F}_t$ for every subset $\mathbb{B}$ of the real line $\mathbb{R} \in \mathcal{R}$.

The important point to emphasize is that a stochastic process is not a single time function but an ensemble of time functions. If the time parameter t belongs to a set of integers $\mathbb{Z} \equiv \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ then $X(n)$ or X_n denotes a *discrete-time* stochastic process.

A non-negative real line will be represented by $\mathbb{R}^+ \equiv [0,\infty)$ and non-negative time set by $\mathbb{T}^+ \equiv [0,\infty)$. A set of non-negative integers will be denoted by $\mathbb{N} \equiv \{0,1,\ldots\}$ and a set of positive integers by $\mathbb{N}^+ \equiv \{1,2,\ldots,N\}$.

Since $X(t)$ is a random variable for every $t \in \mathbb{T}$, the distribution function $F_X(x:t)$ will be given by

$$P\{X(\omega,t) \le x\} = P\{X(t) \le x\} \equiv F_X(x:t),\ x \in \mathbb{R},\ t \in \mathbb{T} \tag{3}$$

and the density function $f_X(x:t)$, which again may include impulse functions of x, is the partial derivative of $F_X(x;t)$ with respect to x.

Autocorrelation and autocovariance functions for a stochastic process $X(t)$ for $\{t_1,t_2 \in \mathbb{T}\}$ are defined by:

$$\begin{aligned} R_X(t_1,t_2) &= \big[X(t_1)X(t_2)\big] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f(x_1,x_2:t_1,t_2)dx_1dx_2, \end{aligned} \tag{4}$$

$$\begin{aligned} C_X(t_1,t_2) &= E\big\{[X(t_1)-\mu_x(t_1)][X(t_2)-\mu_x(t_2)]\big\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[x_1-\mu_x(t_1)][x_2-\mu_x(t_2)] \\ &\quad f(x_1,x_2:t_1,t_2)dx_1dx_2, \end{aligned} \tag{5}$$

where $\mu_x(t_1)$ and $\mu_x(t_2)$ are the mean values of $X(t)$ at times t_1 and t_2 respectively.

Stochastic processes can be classified in different categories but many of them straddle categories.

Stationary and Ergodic Process

A stochastic process $X(t)$ is nth order *stationary* if the nth order distribution function satisfies

$$F_X(x_1,\ldots,x_n:t_1,\ldots,t_n) = F_X(x_1,\ldots,x_n: t_1+\tau,\ldots,t_n+\tau) \text{for any } \tau \in \mathbb{T}. \tag{6}$$

It is *strictly stationary* if Eq. (6) is true for all $n \in \mathbb{Z}$. However, the most useful concepts of stationarity are the first order stationarity defined by

$$F_X(x:t) = F_X(x:t+\tau) = F_X(x), \tag{7}$$

and the second order stationarity called *wide sense stationary* defined by

$$F_X(x_1,x_2:t_1,t_2) = F_X(x_1,x_n:t_1+\tau,t_2+\tau) = F_X(x_1,x_2:\tau). \tag{8}$$

Wide sense stationarity can be determined from the following two criteria:

1. The expected value $E[X(t)] = \mu_X$ = a constant.
2. The autocorrelation function $R_X(t_1,t_2) = R_X(t_2-t_1) = R_X(\tau)$ is a function of the time difference τ.

A stationary process $X(t)$ is *mean ergodic* if the ensemble average is equal to the time average of the sample function $X(t)$.

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}X(t)dt = \int_{-\infty}^{\infty}xf_X(t)dt = \mu_X, \tag{9}$$

or, equivalently the covariance $C_X(\tau)$ satisfies the condition $\int_{-\infty}^{\infty}|C_X(\tau)|d\tau < \infty$.

A stationary process is *correlation ergodic* if

$$\begin{aligned} &\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}X(t)X(t+\tau)dt \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x_1x_2f_X(x_1,x_2:\tau)dx_1dx_2 = R_X(\tau), \end{aligned} \tag{10}$$

which is equivalent to the condition $\int_{-\infty}^{\infty}\big|E\{[X(t)X(t+\tau)]^2\} - E\{[X(t)]^2\}\big|d\tau < \infty$.

State and Time Discretized Process

The stochastic process $X(t)$ can be classified into four broad categories depending upon whether the state space is discretized with $\mathbb{R} \equiv \mathbb{Z}$ or the time is discretized with $\mathbb{T} \equiv \mathbb{Z}$ or both. As mentioned earlier, discrete-time random processes will be denoted by X_n or $X(n)$ where $n \in \mathbb{Z}$.

1. *Discrete State Discrete Time Process* (DSDT)
   At any given time $i > 0$ a particle takes a positive step from $X_0 = 0$ with probability p and a negative step with probability q with $p+q = 1$. The random variable Z_i representing each step is independent and identically distributed. The position X_n of the particle at time n is a stochastic process $X_n = Z_1 + Z_2 + \cdots + Z_n$. It represents a DSDT process with discrete time set $\mathbb{N}^+ = \{1,\ldots,n,\ldots\}$ and discrete state space $\mathbb{R} = \mathbb{Z} = \{\ldots,-1,0,1,\ldots\}$ representing the position of the particle. This process known as a *simple* ►*random walk* (see, e.g., Cox and Miller 1977, p. 25) is nonstationary. If $p = q$ then the process is called a *symmetric* simple random walk.
2. *Discrete State Continuous Time Process* (DSCT)
   A customer arrives at the service counter of a supermarket at a random time $t \ge 0$ at an average rate of λ per unit time interval. If $N(t)$ is the stochastic process representing the number of customers arriving in the time interval $[0,t]$ then $N(t)$ is a DSCT process with time set $\mathbb{T}^+ = \{0 \le t < \infty\}$ and discrete state space $\mathbb{N} = \{0,1,\ldots\}$ representing the number of customers. This process known as *Poisson process* (see ►Poisson Processes) is nonstationary.
3. *Continuous State Discrete Time Process* (CSDT)
   In the DSDT process of (1), each step of the particle at any time $i > 0$ is a continuous random variable Z instead of a discrete one, governed by a distribution function $F_Z(z)$ with mean μ_Z. If X_n is the position of the particle at time $i = n$ then X_n represents a CSDT

process with discrete time set $\mathbb{N}^+ = \{1, \ldots, n, \ldots\}$, and continuous state space $\mathbb{R}^+ = \{0 \le x < \infty\}$ representing the position of the particle. This process is nonstationary.

4. *Continuous State Continuous Time Process* (CSCT)
In the DSDT process of (1) the particle undergoes a positive or negative step of Δx in a time interval Δt. If certain limiting conditions on Δx and Δt are satisfied then as Δx and Δt tend to 0, a CSCT process results, which is called *Wiener process* (see, e.g., Cox and Miller 1977, p. 205) or *Brownian motion* (see ▶Brownian Motion and Diffusions). Extrusion of plastic shopping bags where the thicknesses of the bags vary constantly with respect to time with the statistics being constant over long periods of time is an example of a CSCT process. These processes are nonstationary.

Gaussian Process

A stochastic process $X(t)$ defined on a complete probability space is a *Gaussian stochastic process* if for any collection of times $\{t_0, t_1, \ldots, t_n\} \in \mathbb{T}$, the random variables $X_0 = X(t_0), X_1 = X(t_1), \ldots, X_n = X(t_n)$ are jointly Gaussian distributed for all $n \in \mathbb{Z}$, with joint probability function

$$f_{X_0 X_1 X_2 \ldots X_n}(x) = \frac{1}{(2)^{n/2}|\mathbf{C}_X|} \exp\left(- \frac{(\mathbf{x}-\boldsymbol{\mu}_X)^T \mathbf{C}_X^{-1} (\mathbf{x}-\boldsymbol{\mu}_X)}{2} \right) \tag{11}$$

where $\boldsymbol{\mu}_X$ is the mean vector and $\mathbf{C}_X$ is the covariance matrix of the random variables $\{X_0, X_1, \ldots, X_n\}$. The Wiener process is also an example of a Gaussian process.

Markov Process

Let the σ-field $\mathcal{F}_t$ generated by $\{X(s), s \le t, t \in \mathbb{T}\}$ represent the past history up to the present and the σ-field $\mathcal{F}_t^c$ generated by $\{X(s), s > t, t \in \mathbb{T}\}$ represent the future evolution. Let a random variable Y be $\mathcal{F}_t$-measurable and another random variable Z be $\mathcal{F}_t^c$-measurable. Then the process $\{X(t), t \in \mathbb{T}\}$ is called a *Markov process* (see Markov Processes) if the following hold:

1. Given the present information $X(t)$, the past Y and the future Z are conditionally independent.

$$E[YZ|X(t)] = E[Y|X(t)]E[Z|X(t)]. \tag{12}$$

2. The future Z, conditioned on the past history up to the present $\mathcal{F}_t$, is equal to the future given the present.

$$E[Z|\mathcal{F}_t] = E[Z|X(t)]. \tag{13}$$

3. The future Z, conditioned on the past value $X(s)$ is the future conditioned on the present value $X(t)$ and again conditioned on the past value $X(s)$.

$$E[Z|X(s)] = E\{E[Z|X(t)]|X(s)\} \text{ for } s < t. \tag{14}$$

This is known as the *Chapman-Kolmogorov equation* (see, e.g., Ross 2000, p. 166).

In terms of probability, with $\tau > 0$ and states x_h, x_i, x_j, Eq. (13) is equivalent to:

$$\begin{aligned} P\{X(t+\tau) = x_j | X(t) = x_i, X(u) \\ = x_h,\ 0 \le u < t\} = P\{X(t+\tau) \\ = x_j | X(t) = x_i\}. \end{aligned} \tag{15}$$

Or, for $t_0 < t_1 < \ldots < t_{n-1} < t_n$, and $\{x_k, k = 0, \ldots, n, \ldots\}$ belonging to some discrete-state space

$$\begin{aligned} P\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n, X(t_{n-1}) \\ = x_{n-1}, \ldots, X(t_0) = x_0\} \\ = P\{X(t_{n+1}) = x_{n+1} | X(t_n) = x_n\}. \end{aligned} \tag{16}$$

A Markov process has an important property that the density $f_{\tau_i}(t)$ of the random time τ_i spent in any given state x_i is an exponential and hence it is called *memoryless*.

Markov Chains

Discrete state Markov processes are called *chains*, and if time is continuous they are called *continuous* ▶*Markov chains*, and if time is discrete they are called *discrete Markov Chains*. The Poisson process is an example of a continuous Markov chain.

A stochastic process $\{X(t), t \in \mathbb{T}^+\}$ is a continuous-time Markov chain if for each of the discrete states h, i, j and any time $\tau > 0$

$$\begin{aligned} P\{X(t+\tau) = j | X(t) = i, X(u) = h, 0 \le u < t\} \\ = P\{X(t+\tau) = j | X(t) = i\}. \end{aligned} \tag{17a}$$

The quantity $P\{X(t+\tau) = j | X(t) = i\}$ is the time dependent transition probability defined by $p_{ij}(t, \tau)$, which is generally a function of times t and τ. If the transition from the state i to the state j is dependent only on the time difference $\tau = (t+\tau) - t$ then the transition probability is stationary and the Markov chain is called *homogeneous*. In this case transition probability becomes $p_{ij}(\tau)$.

The probability density function $f_{\tau_i}(t)$ of the random time τ_i spent in any given state i for a continuous Markov chain is exponential and hence it is called *memoryless*.

A stochastic process $\{X(n), n = 0, 1, \ldots\}$ is a discrete-time Markov chain if for each of the discrete states i, j and $\{i_k, k = 0, 1, \ldots, n-1\}$ and any time $m > 0$,

$$\begin{aligned} P\{X(n+m) = j | X(n) = i, X(n-1) = i_{n-1}, \ldots, X(0) = i_0\} \\ = P\{X(n+m) = j | X(n) = i\}. \end{aligned} \tag{17b}$$

The quantity $P\{X(n+m) = j|X(n) = i\}$ is called the m-step transition probability defined by $p_{ij}^{(m)}(n)$, which is generally a function of time n. If the transition from the state i to the state j is dependent only on the time difference $m = (n + m) - n$ then the transition probability is stationary and the Markov chain is *homogeneous*. In this case the m-step transition probability becomes $p_{ij}^{(m)}$.

The one-step probability from state i to state j of a homogeneous discrete Markov chain is given by:

$$P\{X(n+1) = j|X(n) = i\} = p_{ij}. \tag{18}$$

The probability mass function f_{τ_i} of the random time τ_i spent in any given state i for a discrete Markov chain is geometric and hence it is called *memoryless*.

Semi-Markov Process

In a Markov process the distributions of state transition times are exponential for a continuous process, and geometric for a discrete process and hence they are considered memoryless. While the definition of a *semi-Markov process* $X(t)$ defined on a complete probability space is the same as that of a Markov process (Eqs. 15 and 16), the distributions of transition times $\tau_i \in \mathbb{T}$ between states need not be memoryless but can be arbitrary. For a continuous-time semi-Markov process the state transitions can occur at any instant of time $t \in \mathbb{T}$ with an arbitrary density $f_{\tau_i}(t)$ for the time τ_i spent in state x_i and for a discrete-time semi-Markov process the state transitions can occur at time instants $\mathbb{Z} = \{\ldots, -1, 0, 1, \ldots\}$ with an arbitrary probability mass f_{τ_i} for the time τ_i spent in state i. If the amount of time spent in each state is 1 then this semi-Markov process is a Markov chain. Markov processes are a subclass of semi-Markov processes.

Independent Increment Process

A stochastic process $\{X(t), t \in \mathbb{T}\}$ is defined on a complete probability space with a sequence of time variables $\{t_0 < t_1 < \ldots < t_n\} \in \mathbb{T}$. If the increments $X(t_0), [X(t_1) - X(t_0)], \ldots, [X(t_n) - X(t_{n-1})]$ of the process $\{X(t), t \in \mathbb{T}\}$ are a sequence of independent random variables then the process is called an *independent increment* process (see, e.g., Krishnan 2006, p. 507). If the distribution of the increments $X_t - X_s, t > s$ depends only on the time difference $t - s = \tau$, then the process is a *stationary independent increment* process.

If the time set is discrete given by $\mathbb{N}^+ = \{1, 2, \ldots\}$ then the independent increment process is a sequence of independent random variables given by $Z_0 = X_0, \{Z_i = X_i - X_{i-1}, i \in \mathbb{N}^+\}$. Independent increment process is a special case of a Markov process. It is not a stationary process because of the following (see, e.g., Krishnan 2005, p. 61):

$$\begin{aligned} E[X(t)] &= \mu_0 + \mu_1 t, \text{ where } \mu_0 = E[X(t_0)] \text{ and} \\ \mu_1 &= E[X(t_1)] - \mu_0; \\ \mathrm{Var}[X(t)] &= \sigma_0^2 + \sigma_1^2 t, \text{ where } \sigma_0^2 = E[X(t_0) - \mu_1]^2 \text{ and} \\ \sigma_1^2 &= E[X(t_1) - \mu_0]^2 - \sigma_0^2. \end{aligned} \tag{19}$$

Poisson and Wiener processes are examples of stationary independent increment processes.

Uncorrelated and Orthogonal Increment Process

A stochastic process $\{X(t), t \in \mathbb{T}\}$ with $s_1 < t_1, s_2 < t_2$ and $t_1 \le t_2$

1. Has *uncorrelated increments* (see, e.g., Krishnan 2006, p. 508) if

$$E[(X_{t_2} - X_{s_2})(X_{t_1} - X_{s_1})] = E[(X_{t_2} - X_{s_2})]E[(X_{t_1} - X_{s_1})]. \tag{20}$$

2. Has *orthogonal increments* (see, e.g., Krishnan 2006, p. 508) if

$$E[(X_{t_2} - X_{s_2})(X_{t_1} - X_{s_1})] = 0. \tag{21}$$

Clearly, independent increments imply uncorrelated increments but the converse is not true.

General Random Walk Process

The simple random walk discussed earlier can be generalized. Starting from $X_0 = 0$ a particle takes independent identically distributed random steps $Z_1, Z_2, \ldots, Z_n$, whose values are drawn from an arbitrary distribution, which do not change with the state of the process. This distribution may be continuous with density function $f_Z(z)$ or discrete with probability of transition from state i to state j being p_{ij}. In the latter case p_{ij} will be dependent on the difference $j-i$, or, $p_{ij} = p_{j-i}$. The position $X_n = Z_1 + Z_2 + \cdots + Z_n$, $n \in \mathbb{N}^+$ of the particle is a stochastic process where n is the number of state transitions, which is always forward from state x_i to x_{i+1}. Depending upon whether the instants of these transitions are taken from the set $\mathbb{T}^+$ or $\mathbb{N}^+$ the process X_n is either a continuous-time or a discrete-time *general random walk* (see, e.g., Cox and Miller 1977, p. 46). In either case the distribution of the time intervals between these transitions is arbitrary and hence it is a special case of a semi-Markov process.

Birth and Death Process

Let $\{X(t), t \ge 0\}$ be a continuous Markov chain. State transitions can occur only from the state $x_i = i$ to $x_{i+1} = i + 1$, or $x_{i-1} = i - 1$, or stays at $x_i = i$. $X(t)$ is called a *birth and*

death process (see, e.g., Kleinrock 1975, p. 53) if in a small interval Δt

$$P\{X(t+\Delta t) - X(t) = j | X(t) = i\} = \begin{cases} \lambda_i \Delta t + o(\Delta t), & \text{if} \quad j = 1, \\ \mu_i \Delta t + o(\Delta t), & \text{if} \quad j = -1, \\ o(\Delta t), & \text{if} \quad |j| > 1. \end{cases} \tag{22}$$

$$\text{and } P\{X(t+\Delta t) - X(t) = 0 | X(t) = i\} = 1 - (\lambda_i + \mu_i)\Delta t + o(\Delta t), \tag{23}$$

where $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$. λ_i is the rate at which births occur and μ_i is the rate at which deaths occur when the population size is i. The probability of the population size being i at any time $t > 0$ is given by $P\{X(t) = i\} = P_i(t)$. This is a Markov process with independent increments. If $\lambda_i = i\,\lambda$ and $\mu_i = i\,\mu$ then this process is called a *linear* birth and death process.

The *pure birth process* is a sub-class of birth and death process with $\mu_i \equiv 0$ for all i. State transitions can occur only from the state $x_i = i$ to $x_{i+1} = i + 1$ with rate λ_i or stays in the same state $x_i = i$.

The *Poisson process* is a sub-class of pure birth processes with $\lambda_i \equiv \lambda$ a constant for all i. Here the probability of i events in time t is given by $P_i(t, \lambda) = [(\lambda t)^i / i!] e^{-\lambda t}, t > 0$. This process has stationary independent increments.

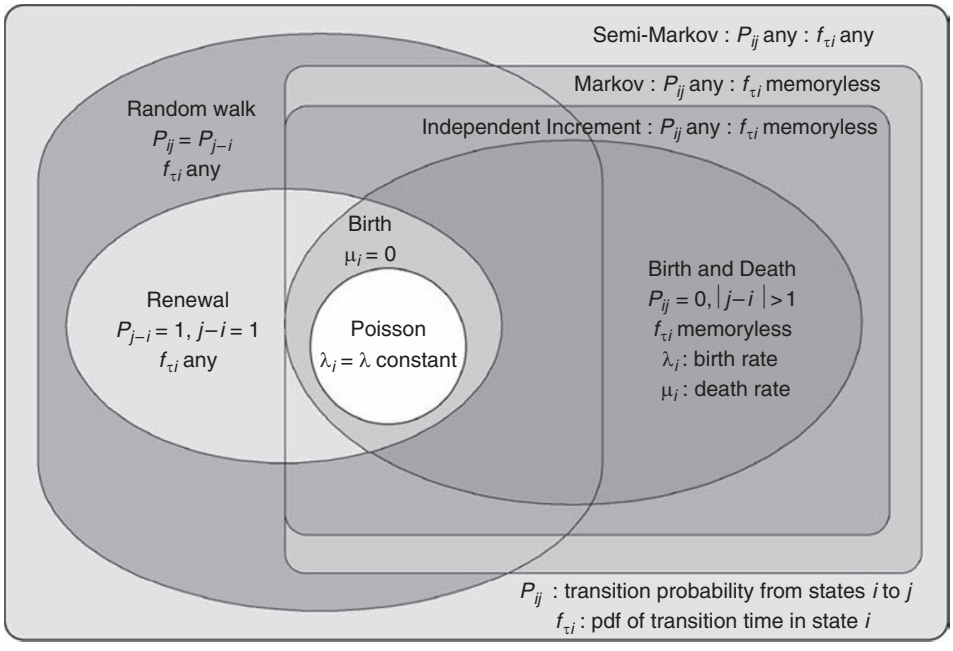
Renewal Process

In the general random walk process X_n discussed in the previous section the interest was in the probability of the state of the particle after n transitions. In ►renewal processes the concern is only in the number of transitions that occur in a time interval $[0, t]$ and not on the state. Starting from $t = 0$ the transitions occur at sequence of times $0 < t_1 < t_2 < \ldots < t_n, n > 0$ with *inter-arrival times* defined by random variables $Y_1 = t_1, Y_2 = (t_2 - t_1), \ldots, Y_n = (t_n - t_{n-1})$. The random variables $Y_i, i \in \mathbb{N}^+$ are independent and identically distributed with an arbitrary density function $f(y)$ with $E[Y_i] = \mu$ for all i.

The stochastic process defined by $X_n = Y_1 + Y_2 + \cdots + Y_n$ is called a *renewal* process (see, e.g., Cox and Miller 1977, p. 340), where a renewal occurs at the epochs at $t_1 < t_2 < \cdots < t_n$. In this process X_n represents the *time* of the *nth* renewal whereas in the random walk X_n represents the *state* of the process at time n. This process is a subclass of semi-Markov processes and also a subclass of random walk processes. If the density function $f(y)$ is either exponential or geometric then this process is Markov. The relationship among the various discrete-state random processes similar to the one in Kleinrock (1975, p. 25) is shown in Fig. 1.

Martingale Process

A *martingale process* (see, e.g., Doob 1990, p. 91 and p. 294; ►Martingales) is a stochastic process where the best estimate of the future value conditioned on the past history



Stochastic Processes: Classification. Fig. 1 Relationships among some discrete state stochastic processes

including the present is the present value. Since there is no trend to the process it is unpredictable. Many problems in engineering and finance can be cast in the martingale framework. Pricing stock options (see, e.g., Ross 2000, p. 556) and bonds has been cast in the martingale framework.

Let $\{\Omega, \mathcal{F}, P\}$ be a complete probability space and let $\{\mathcal{F}_n, n \in \mathbb{N}\}$ be an increasing family of sub σ-fields of $\mathcal{F}$. The real valued sequence of random variables $\{X_n, n \in \mathbb{N}\}$ adapted to the family $\{\mathcal{F}_n, n \in \mathbb{N}\}$ is a discrete $\mathcal{F}_n$-martingale if for all n:

1. $E|X_n| < \infty$
2. $E\{X_n|\mathcal{F}_m\} = X_m$ for $m \le n$

If condition (2) is modified as

3. $E\{X_n|\mathcal{F}_m\} \ge X_m$ for $m \le n$ submartingale
4. $E\{X_n|\mathcal{F}_m\} \le X_m$ for $m \le n$ supermartingale

Analogously, let $\{\mathcal{F}_t, t \in \mathbb{T}^+\}$ be an increasing family of sub σ-fields of $\mathcal{F}$ of a complete probability space. The real valued stochastic process $\{X(t), t \in \mathbb{T}^+\}$ adapted to the family $\{\mathcal{F}_t, t \in \mathbb{T}^+\}$ is a continuous $\mathcal{F}_t$-martingale if for all $t \in \mathbb{T}^+$:

1. $E|X(t)| < \infty$,
2. $E\{X(t)|\mathcal{F}_s\} = X_s$ for $s \le t$.

If condition (2) is modified as

3. $E\{X(t)|\mathcal{F}_s\} \ge X_s$ for $s \le t$ submartingale.
4. $E\{X(t)|\mathcal{F}_s\} \le X_s$ for $s \le t$ supermartingale.

Note that any martingale is both a submartingale and a supermartingale.

In the simple random walk process given in DSDT, if $n(p-q)$ is subtracted from X_n, then $Y_n = [X_n - n(p-q)]$ is an example of a discrete martingale with respect to the sequence $\{Z_k, k = 1, \ldots, n-1\}$ even though X_n is not. The Wiener process $W(t)$ is an example of a continuous $\mathcal{F}_t$-martingale. In the Poisson process $N(t)$, if the mean λt is subtracted then $Y(t) = [N(t) - \lambda t]$ is another example of a continuous $\mathcal{F}_t$-martingale even though $N(t)$ is not. However, both X_n and $N(t)$ are Markov processes leading to the conclusion that a Markov process is not necessarily a martingale. It can also be shown that a martingale is not necessarily a Markov process.

The martingale property captures the notion of a fair game. A fair coin is tossed and a player wins a dollar if the toss is heads and loses a dollar if the toss is tails. At the end of the *mth* toss the player has X_m dollars. The estimated amount of money after the $m+1$st toss is still X_m dollars since the expected value of the $m+1$st toss is zero.

Periodic Process

Let $\{X(t), t \in \mathbb{T}\}$ be a stochastic process defined on a complete probability space taking values in the range space $\{\mathbb{R}, \mathcal{R}\}$. $X(t)$ is *periodic in the wide sense* (see, e.g., Krishnan 2006, p. 558) with period $T_c (T_c > 0)$ if the mean $\mu_X(t)$ and the autocorrelation function $R_X(t,s)$ satisfy

$$\mu_X(t) = \mu_X(t + kT_c) \text{ for all } t \text{ and integer } k \tag{24}$$

$$\begin{aligned} R_X(t,s) &= R_X(t + kT_c, s) \\ &= R_X(t, s + kT_c) \text{ for all } t, s \text{ and integer } k. \end{aligned} \tag{25}$$

Note that $R_X(t,s)$ is periodic in both arguments t and s.

However, for a stationary periodic process $X(t)$ with $\tau = t - s$, Eq. (25) simplifies to

$$R_X(\tau) = R_X(\tau + kT_c) \text{ for all } \tau \text{ and integer } k. \tag{26}$$

Since $R_X(\tau)$ is uniformly continuous, a zero mean stationary periodic stochastic process $X(t)$ with fundamental frequency $\omega_c = 2\pi/T_c$ can be represented in the mean square sense by a Fourier series

$$\begin{aligned} X(t) &= \sum_{n=-\infty}^{\infty} X_n \exp(jn\omega_0 t), X_0 = 0 \\ \text{where } X_n &= \frac{1}{T_c} \int_0^{T_c} X(t) \exp(-jn\omega_0 t) dt. \end{aligned} \tag{27}$$

Cyclostationary process

Allied to the periodic process is the *cyclostationary process* (see, e.g., Krishnan 2006, p. 560). A *strict sense* cyclostationary process $X(t)$ on a complete probability space with period $T_c (T_c > 0)$ is defined by

$$\begin{aligned} &F_X(x_1, \ldots, x_n; t_1, \ldots, t_n) \\ &= F_X(x_1, \ldots, x_n; t_1 + kT_c, \ldots, t_n + kT_c) \end{aligned} \tag{28}$$

for all n and k.

Since the above definition is too restrictive, a *wide sense* cyclostationary $X(t)$ can be defined by

$$\begin{aligned} \mu_X(t) &= \mu_X(t + kT_c) \\ R_X(t_1, t_2) &= R_X(t_1 + kT_c, t_2 + kT_c). \end{aligned} \tag{29}$$

About the Author

Venkatarama Krishnan, Ph D, is Professor Emeritus in the Department of Electrical and Computer Engineering at the University of Massachusetts Lowell. Previously, he has taught at Smith College (2003), the Indian Institute of Science Bangalore (1971–1987), Polytechnic University of New York (1964–1971), University of Pennsylvania (1961–1964), Villanova University (1958–1961), and Princeton

University (1957–1958). In 1956 he was the recipient of an Orson Desaix Munn Scholarship from Princeton University. He was also a co-director (1992–2000) of the Center for Advanced Computation and Telecommunications at University of Massachusetts Lowell. He has taught Probability and Stochastic Processes continuously for over forty years and received the best teaching award from University of Massachusetts Lowell in 2000. He has authored four books in addition to technical papers, the latest book being *Probability and Stochastic Processes* published by Wiley in 2006. Prof. Krishnan is a life senior member of IEEE, and is listed in *Who is Who in America, 2010.*

Cross References

►Brownian Motion and Diffusions
►Gaussian Processes
►Lévy Processes
►Markov Chains
►Markov Processes
►Martingales
►Point Processes
►Poisson Processes
►Random Walk
►Renewal Processes
►Stochastic Processes

References and Further Reading

Doob JL (1990) Stochastic processes, Wiley, New York
Gikhman II, Skorokhod AV (1996) Introduction to the theory of random processes. Dover, New York
Krishnan V (2005) Nonlinear filtering and smoothing. Dover, New York
Krishnan V (2006) Probability and random processes.Wiley, Hoboken, NJ
Cox DR, Miller HD (1977) The theory of stochastic processes. Chapman and Hall/CRC, London
Kleinrock L (1975) Queueing systems, vol 1. Wiley, New York
Ross SM (2000) Introduction to probability models. Harcourt Academic, San Diego

Stratified Sampling

Michael P. Cohen
Adjunct Professor
George Mason University, Fairfax, VA, USA
NORC at the University of Chicago, Washington DC, USA

Stratification refers to dividing a population into groups, called *strata*, such that pairs of population units within the same stratum are deemed more similar (*homogeneous*) than pairs from different strata. The strata are mutually exclusive (non-overlapping) and exhaustive of the population. Clearly sufficient information on each population unit must be available before we can divide the population into strata.

The primary reason for dividing a population into strata is to make use of the strata in drawing a sample. For example, instead of drawing a simple random sample of sample size n from the population, one may draw a ►simple random sample of sample size n_h from stratum h of L strata, where $n = n_1 + \cdots + n_L$. The sample selection for any stratum is done independently of the other strata. The stratum sample sizes n_h are often chosen proportional to the number of population units in stratum h but other allocations of the stratum samples may be preferred in specific situations.

There are two major reasons for drawing a stratified sample instead of an unstratified one:

1. Such samples are generally more efficient (in the sense that estimates have smaller variances) than samples that do not use stratification. There are exceptions, primarily when the strata are far from homogeneous with respect to the variable being estimated.
2. The sample sizes are controlled (rather than random) for the population strata. This means, in particular, that one may guarantee adequate sample size for estimates that depend only on certain strata. For instance, if men and women are in separate strata, one can assure the sample size for estimates for men and for women.

Estimation Under Simple Random Sampling Within Strata

The independence of the sample selection by strata allows for straightforward variance calculation when simple random sampling is employed within strata. Let Y_T denote the population total for a variable Y for which an estimate is sought. Let N_h and n_h denote respectively the population size and sample size for stratum h. Let, moreover, Y_{hj} and y_{hi} denote respectively the Y-value of the jth population element or i^{th} sample element in stratum h. Then, if

$$\overline{Y}_h = \frac{1}{N_h}\sum_{j=1}^{N_h} Y_{hj} \text{ and } \overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi},$$

define

$$S_h^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}(Y_{hj} - \overline{Y}_h)^2 \text{ and } s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \overline{y}_h)^2.$$

We estimate Y_T by $\hat{y}$ where $\hat{y} = \sum_{h=1}^{L} N_h \bar{y}_h$. The variance of $\hat{y}$ is

$$V(\hat{y}) = \sum_{h=1}^{L} \frac{N_h^2}{n_h}(1 - n_h/N_h)S_h^2$$

and the variance is estimated by

$$\hat{V}(\hat{y}) = \sum_{h=1}^{L} \frac{N_h^2}{n_h}(1 - n_h/N_h)s_h^2.$$

Similarly, the population mean $\overline{Y} = Y_T/N$, where $N = \sum_{h=1}^{L} N_h$ is the size of the population, is estimated by $\hat{y}/N$ and its variance by $\hat{V}(\hat{y})/N^2$.

Allocation of Sample Sizes to Strata Under Simple Random Sampling within Strata

For a total sample size of n and given values of S_h, the question arises how should one allocate the sample to the strata; that is, how should one choose the n_h, $h = 1, \ldots, L$, so that $n = n_1 + \cdots + n_L$ and $V(\hat{y})$ is minimized? This is a straightforward constrained minimization problem (solved with Lagrange multipliers) that yields the solution:

$$n_h = \frac{nN_hS_h}{\sum_{k=1}^{L} N_kS_k}$$

Note that, as one would expect, the more variability in a stratum (larger S_h), the larger the relative sample size in that stratum. This method of determining the stratum sample sizes is termed *Neyman allocation* in view of the seminal paper on stratified sampling by Neyman (1934).

Sometimes the strata are not equally costly to sample. For example, there may be additional travel costs in sampling a rural geographically-determined stratum over an urban one. If it costs C_h to sample a unit in stratum h, then the allocation

$$n_h = \frac{nN_hS_h/\sqrt{C_h}}{\sum_{k=1}^{L} N_kS_k/\sqrt{C_k}}$$

is best in two senses: It minimizes $V(\hat{y})$ subject to fixed total cost (a fixed budget) $C_T = C_1 + \cdots + C_L$ and it minimizes C_T subject to fixed $V(\hat{y})$.

These allocations assume that the S_h, $h = 1, \ldots, L$, are known. In practice, rough estimates, perhaps based on a similar previous survey, will serve. The same comment applies to the costs for the cost-based allocation.

In the absence of any prior information, even approximate, the simple *proportional allocation* $n_h = nN_h/N$ is often used. In this case, the estimator $\hat{y}$ has a particularly simple form

$$\hat{y} = \sum_{h=1}^{L} N_h\bar{y}_h = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} = \sum_{h=1}^{L} \frac{N_h}{(nN_h/N)} \sum_{i=1}^{n_h} y_{hi}$$
$$= \frac{N}{n} \sum_{h=1}^{L} \sum_{i=1}^{n_h} y_{hi}.$$

Therefore $\hat{y}$ is just the sum of the sample values expanded by N/n. In many surveys a wide variety of quantities are estimated and their within-stratum variability may differ so proportional allocation may be employed as a compromise.

Unbiased estimation requires at least one sample selection per stratum. Unbiased variance estimation requires at least two selections per stratum.

Stratum Boundaries

Sometimes stratification is based on small discrete categories like gender or race. Other times, one may have data on a variable that can be regarded as continuous closely related to the variable one wants to estimate from the sample. For example, one may want to estimate the output of factories based on strata defined by the number of workers at the factory. One stratum might be all factories with 75–100 workers. In this case, 75 and 100 are said to be the stratum boundaries. How should these boundaries be chosen?

One method that has been shown to be good is the cumulative square root of frequencies method developed by Dalenius and Hodges (1957): Start by assuming (in our example) that the factories have been divided into a rather large number of categories based on the numbers of workers, numbered from fewest workers to the most workers. If f_k is the number of factories in category k, calculate $Q_k = \sqrt{f_1} + \cdots + \sqrt{f_k}$. Divide the factories into strata so that the differences between the at adjacent stratum boundary points are as equal as possible.

More recently, Lavallée and Hidiroglou (1988) developed an iterative procedure especially designed for skewed populations.

Variance Estimation for Stratified Samples

For simple estimators and stratified sampling, direct formulas are available to calculate variance estimates. These formulas are tailored to the specific estimator whose variance is sought. General purpose variance estimators have

been developed, however, that allow one to estimate variances for a wide class of estimators using a single procedure. See Wolter (2007) and Shao and Tu (1995) for a complete discussion of these procedures.

The procedure *balance half-sample replication* (or *balanced repeated replication*) has been developed as a variance estimation procedure when two primary sampling units (PSUs) are selected from each stratum. There may be additional sampling within each PSU so the sample design may be complex. The variance estimation is based on half sample replicates, each replicate consisting of one PSU from each stratum. The pattern that determines which PSU to choose from each stratum for a particular replicate is based on a special kind of matrix, called a Hadamard matrix.

A form of the *jackknife method* (see ►Jackknife) is also widely employed with two PSU per stratum sample designs (although it can be extended to other designs). This jackknife method is based on forming replicates, but the replicate consists of one PSU selected to be in the replicate from a specific stratum, with both PSUs being in the replicate for all other strata.

Various forms of the *bootstrap method* (see ►Bootstrap Methods) have been employed in recent years as general variance estimation methods for stratified sampling.

Although not as generic, the *Taylor series* (or *linearization*) method is a powerful technique for estimating variances in complex samples.

Stratified Sampling with Maximal Overlap (Keyfitzing)

Sometimes it is worthwhile to select a stratified sample in a manner that maximizes overlap with another stratified sample, subject to the constraint that the probabilities of selection are the ones desired. For example, cost savings may arise if a new stratified sample is similar to a previous one, yet births, deaths, and migration in the population may preclude it being exactly the same. Keyfitz (1951) developed a method to deal with this problem, so it is often called *Keyfitzing*. More recent researchers have extended the method to more general situations.

Stratification in Two Phases

It may be that it is clearly desirable to stratify on a certain characteristic, but that characteristic may not be available on the sampling frame (list of units from which the sample is selected). For example, in travel surveys one would likely want to stratify on household type (e.g., single adult head of household or adult couple with children) but this information is usually not provided on an address list. One solution is to first conduct a large, relatively inexpensive first phase of the survey for the sole purpose of obtaining the information needed to stratify. This information is then employed in the stratification of the second stage of the survey. This process is called *two-phase sampling* or *double sampling*.

Let n_h^I be the size of the first stage sample that lies in stratum h and let $n^I = n_1^I + \cdots + n_L^I$ be the first-stage sample size. At the second stage, n_h^{II} units with Y-values $y_{h1}, \ldots, y_{hn_h^{II}}$ are sampled in stratum h. Then one can estimate Y_T by

$$\tilde{y} = N \sum_{h=1}^{L} \frac{n_h^I}{n^I} \sum_{i=1}^{n_h^{II}} \frac{y_{hi}}{n_h^{II}}$$

Approximate variance formulas can also be given. See, e.g., Raj and Chandhok (1998) or Scheaffer et al. (2006). Because the n_h^I are random, the usual (one-phase) variance formulas would underestimate the variance.

Poststratification

After a sample has been selected and the data collected, sometimes the estimation procedures of stratification can be employed even if the sample selection was for an unstratified design. An important requirement is that the population proportions N_h/N must be known, at least approximately. If so, then

$$\hat{y} = N \sum_{h=1}^{L} \frac{N_h}{N} \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h} = N \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h$$

is an improved estimate of the population total. The usual variance estimator $\hat{V}(\hat{y})$, however, is no longer valid as it does not account for the randomness of the n_h. More complicated variance estimators can be developed for this purpose.

Another reason to employ poststratification is to reduce bias due to nonresponse.

Controlled Selection

Controlled selection is a sample selection method that is related to stratified sampling but differs in that independent selections are not made from the cells ("strata"). The method was introduced by Goodman and Kish (1950). For an example of controlled selection, imagine a two-dimensional array of cells of population units, say of industrial classification categories by geographic areas. All population units lie in exactly one cell, analogous to strata. The sample size is not large enough for there to be the two selections per cell needed for unbiased variance estimation if the selections were independent by cell. Under controlled selection, only certain balanced patterns of cell combinations can be selected. When properly carried out, this is a valid probability selection technique.

About the Author

Dr. Michael P. Cohen is Senior Consultant to the National Opinion Research Center and Adjunct Professor, Department of Statistics, George Mason University. He was President of the Washington Statistical Society (2007–2008), and of the Washington Academy of Sciences (2003–2004). He served as Assistant Director for Survey Programs of the U.S. Bureau of Transportation Statistics (2002–2006). He is a Fellow of the American Statistical Association, the American Educational Research Association, and the Washington Academy of Sciences. He is an Elected Member of the International Statistical Institute and Sigma Xi and a Senior Member of the American Society for Quality. Dr. Cohen has over 60 professional publications. He served as an Associate Editor, *Journal of the American Statistical Association*, Applications and Case Studies Section (2004–2006). He has been an Associate Editor of the *Journal of Official Statistics* since 2003. He is the Guest Problem Editor of the *Journal of Recreational Mathematics* for 2009–2010.

Cross References

►Balanced Sampling
►Jackknife
►Multistage Sampling
►Sampling From Finite Populations
►Simple Random Sample

References and Further Reading

Bethlehem J (2009) Applied survey methods: a statistical perspective. Wiley, Hoboken
Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York
Dalenius T, Hodges JL (1957) The choice of stratification points. Skandinavisk Aktuarietidskrift 1–2:203–213
Goodman R, Kish L (1950) Controlled selection – a technique in probability sampling. J Am Stat Assoc 45:350–372
Keyfitz N (1951) Sampling with probabilities proportional to size: adjustment for changes in the probabilities. J Am Stat Assoc 46:105–109
Knottnerus P (2003) Sample survey theory: some Pythagorean perspectives. Springer, New York
Lavallée P, Hidiroglou M (1988) On the stratification of skewed populations. Surv Methodol 14:33–43
Lohr S (1999) Sampling: design and analysis. Brooks/Cole, Pacific Grove
Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J R Stat Soc 97:558–606
Raj D, Chandhok P (1998) Sample survey theory. Narosa Publishing House, New Delhi
Scheaffer RL, Mendenhall W, Ott RL (2006) Elementary survey sampling, 6th edn. Duxbury, Belmont
Shao J, Tu D (1995) The jackknife and the bootstrap. Springer, New York
Wolter KM (2007) Introduction to variance estimation, 2nd edn. Springer, New York

Strong Approximations in Probability and Statistics

Murray D. Burke
Professor
University of Calgary, Calgary, AB, Canada

Strong approximations in Probability and Statistics are results that describe the closeness almost surely of random processes such as partial sums and ►empirical processes to certain ►Gaussian processes. As a result, strong laws such as the law of the iterated logarithm and weak laws such as the central limit theorem (see ►Central Limit Theorems) follow.

Let $X_1, X_2, \ldots$ be a sequence of independent random variables with the same distribution function. Put $S_n = X_1 + \cdots + X_n$. If the mean $m = E(X_1)$ exists (finite), then the strong law of large numbers states that $S_n/n \to m$, almost surely, as $n \to \infty$. One can ask the question, at what rate does this convergence take place? This question is answered, in 1941 by Hartman and Wintner, who proved the law of the iterated logarithm (LIL): If, in addition, the variance σ^2 of X_1 is finite, then

$$\begin{aligned} \limsup_{n\to\infty} \frac{S_n - nm}{\sigma\sqrt{2n\log\log n}} &\to_{a.s.} 1, \\ \liminf_{n\to\infty} \frac{S_n - nm}{\sigma\sqrt{2n\log\log n}} &\to_{a.s.} -1. \end{aligned} \tag{1}$$

To gain further insight about the asymptotic behavior of partial sums, we can consider $S_{\lfloor nt \rfloor}$, $0 \le t \le 1$, as a random process. In 1964, Strassen proved that it can be approximated by a standard Brownian motion process (see ►Brownian Motion and Diffusions). A standard Brownian motion (or Wiener process) is a random process $\{W(t); t \ge 0\}$ that has stationary and independent increments, where the distribution of $W(t)$ is normal with mean 0 and variance t, for any fixed $t > 0$ and $W(0) = 0$.

Strassen showed that if $m = E(X_1)$ and $Var(X_1) = \sigma^2 < \infty$, then there exists a common probability space on which one can define a standard Brownian motion process W and a sequence of independent and identically distributed random variables $Y_1, Y_2, \ldots$ such that $\{S_n = \sum_{i=1}^n X_i : n \ge 1\} =_D \{\tilde{S}_n = \sum_{i=1}^n Y_i : n \ge 1\}$ and, as

$n \to \infty$,

$$\sup_{0\le t\le 1} \frac{|\sigma^{-1}(\tilde{S}_{\lfloor nt\rfloor} - m\lfloor nt\rfloor) - W(nt)|}{\sqrt{n\log\log n}} \to_{a.s.} 0, \qquad (2)$$

where $\lfloor nt\rfloor$ is the largest integer less than or equal nt.

Statement (2) is an example of a strong approximation which gives rise to the *strong invariance principle*. From it one can deduce the law of the iterated logarithm for partial sums (1) from that of standard Brownian motion (Khinchin's LIL). Alternately, one can prove it for a specific sequence of random variables, say simple coin tossing, and then, via (2), it is inherited by any independent sequence with a common distribution having finite variance.

If one assumes further conditions on the moments of the random variables (beyond finite variance) then the rate of convergence in (2) can be improved. In particular, if one assumes that X_1 has a finite moment generating function in an open interval containing the origin, then Komlós et al. (1975) have proven a Theorem 1-type result with convergence statement:

$$\limsup_{n\to\infty} \sup_{0\le t\le 1} \frac{|\sigma^{-1}(\tilde{S}_{\lfloor nt\rfloor} - m\lfloor nt\rfloor) - B(nt)|}{\log n} \le C, \ a.s. \quad (3)$$

for some constant $C > 0$.

Many almost-sure results including (3) are proven by first establishing an inequality for the maximal deviations and then applying a Borel-Cantelli lemma (see ▶Borel–Cantelli Lemma and Its Generalizations). The Komlós et al. inequality is:

$$P\left\{\max_{1\le k\le n} |\sigma^{-1}(\tilde{S}_k - mk) - B(k)| > c_1\log n + x\right\} < c_2 e^{-c_3 x},$$

where c_1, c_2, c_3 are positive constants depending only on the distribution of X_1. The Borel-Cantelli lemma to be used is: for any sequence of events $A_n, n \ge 1$, if $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n$, infinitely often$) = 0$. Massart (1989) proved a multivariate version of (3).

The rate $\mathcal{O}(\log n)$ in (3) is the best rate possible. This is a consequence of the Erdös- Rényi laws of large numbers:

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with mean $E(X_1) = m$ and where the ▶moment generating function $M(t) = E(e^{t(X_1-m)})$ of $X_1 - m$ is finite in an interval containing $t = 0$. Then, for any $c > 0$,

$$\max_{0\le k\le n-\lfloor c\log n\rfloor} \frac{S_{k+\lfloor c\log n\rfloor} - S_k - m\lfloor c\log n\rfloor}{\lfloor c\log n\rfloor} \to_{a.s.} \alpha(c),$$

where $\alpha(c) = \sup\{x : \varrho(x) \ge e^{-1/c}\}$, with $\varrho(x) = \inf_t e^{-tx}M(t)$, the Chernoff function of $X_1 - m$.

If the left side of (3) converged to 0, almost surely, then $\sigma^{-1}(X_i - m)$ and $B(i) - B(i-1)$ would share the same function α. Since α uniquely determines the distribution function of a random variable, $\sigma^{-1}(X_i - m) = {}_D B(i) - B(i-1)$, a standard normal distribution.

Empirical process are important in many areas of statistics. If $X_1, X_2, \ldots$ is a sequence of independent k-dimensional random vectors with distribution function F, let $F_n(x) = n^{-1}\sum_{i=1}^{n} I[X_i \le x]$, $x \in R$, is the proportion of $X_1, X_2, \ldots, X_n$ that are less than or equal to the real vector $x = (x_1, \ldots, x_k)$ in the usual partial ordering of R^k. The empirical process is defined as

$$\alpha_n(x) = \sqrt{n}[F_n(x) - F(x)], \quad x \in R^k.$$

Strong approximation results are available for the empirical process which describe its behavior in terms of both $x \in R^k$ and the sample size. A Kiefer process $K_F(x,y)$ is a Gaussian process defined on $R^k \times [0,\infty)$ that has mean zero and covariance function $E(K(x_1,y_1)K(x',y')) = (\min\{y_1,y'\})(F(x\wedge x') - F(x)F(x'))$, where $x \wedge x' = \left(\min\left\{x_1,x_1'\right\},\ldots,\min\left\{x_k,x_k'\right\}\right)$.

In 1988, Csörgő and Horváth proved that there exists a common probability space on which one can define a Kiefer process K and a sequence of independent and identically distributed random variables $Y_1, Y_2, \ldots$ such that its empirical process $\{\tilde{\alpha}_n(x);\ x \in R^k, n = 1,2,\ldots\} = {}_D\{\alpha_n(x);\ x \in R^k, n = 1,2,\ldots\}$, the empirical process of the original sequence of X_i, and

$$\limsup_{n\to\infty} \max_{1\le j\le n} \sup_{x\in R^k} \frac{|\tilde{\alpha}_j(x) - j^{-1/2}K(x,j)|}{n^{-1/(4k)}(\log n)^{3/2}} \le C, \ a.s. \qquad (4)$$

When the dimension $k = 1$, the denominator in (4) can be improved to $n^{-1/2}(\log n)^2$. Similar to partial sums, the law of the iterated logarithm for the empirical process can be deduced from that of the Kiefer process, that is

$$\limsup_{n\to\infty} \sup_{x\in R^k} \frac{|\alpha_n(x)|}{\sqrt{\frac{1}{2}\log\log n}} =_{a.s.} 1.$$

Other results involve the strong approximation of the empirical process by a sequence of Brownian bridges B_n, where each is a Gaussian process defined on R^k and each has mean zero and covariance function $EB_n(x)B_n(x') = F(x\wedge x') - F(x)F(x')$. For general F, Borisov proved an approximation with rate $O(n^{-1/(2(2k-1))}logn)$, a.s. When F has a density, Rio obtained a rate of $O(n^{-1/12}(logn)^{(5k+1)/6})$, a.s. Here the exponent of n is independent of the dimension. When F is the uniform distribution on $[0,1]^k$, Massart, in 1989, proved (4) with a rate of $O(n^{-1/(2(k+1))}(\log n)^2)$ and obtained an approximation in terms of sequences of Brownian bridges with a rate of $O(n^{-1/(2k)}(\log n)^{1/2})$.

About the Author

Murray David Burke is Professor of Mathematics and Statistics at the University of Calgary. He was Chair of the Division of Statistics and Actuarial Science (1988–1991, 1997–2001) and President of the Probability Section of the Statistical Society of Canada (2009–2010). He is an elected member of the International Statistical Institute.

Cross References

►Approximations to Distributions
►Borel–Cantelli Lemma and Its Generalizations
►Brownian Motion and Diffusions
►Convergence of Random Variables
►Empirical Processes
►Laws of Large Numbers
►Limit Theorems of Probability Theory

References and Further Reading

Borisov IS (1982) An approximation of empirical fields. In: Nonparametric statistical inference. Coll Math Soc János Bolyai, Budapest, Hungary, 1980, vol 32. North Holland, Amsterdam, 1982, pp 77–87

Csörgő M, Horváth L (1988) A note on strong approximations of multivariate empirical processes. Stoch Proc Appl 27:101–109

Csörgő M, Révész P (1981) Strong approximations in probability and statistics. Academic, New York

DasGupta A (2008) Asymptotic theory of statistics and probability. Springer, New York

Hartman P, Wintner A (1941) On the law of the iterated logarithm. Am J Math 63:169–176

Komlós J, Major P, Tusnády G (1975) An approximation of partial sums of independent r.v.'s and the sample df. I. Z Wahrcheinlichkeitstheorie verw Gebiete 32:111–131

Massart P (1989) Strong approximations for multivariate empirical and related processes, via KMT constructions. Ann Probab 17:266–291

Rio E (1996) Vitesses de convergence dans le principe d'invariance faible pour la fonction de répartition empirique multivarée. CR Acad Sci Paris t 322(1):169–172

Strassen V (1964) An invariance principle for the law of the iterated logarithm. Z Wahrcheinlichkeitstheorie verw Gebiete 3:211–226

Structural Equation Models

Scott L. Hershberger
Global Director of Survey Design
Harris Interactive, New York, NY, USA

Introduction

A *structural equation model* is a representation of a series of hypothesized relationships between observed variables and *latent variables* into a composite hypothesis concerning patterns of statistical dependencies. The hypothesized relationships are described by parameters that indicate the magnitude of the relationship (direct or indirect) that independent (*exogenous*) variables (either observed or latent) have on dependent (*endogenous*) variables (either observed or latent). By enabling the representation of hypothesized relationships into testable mathematical models, a structural equation model offers a comprehensive method for the quantification and testing of theoretical models. Once a theory has been proposed, it can be tested against empirical data.

The term *structural equation model* was first coined by econometricians and is probably the most appropriate name for the process just briefly sketched. *Path analysis*, developed by Sewall Wright (1921), is an early form of SEM that is restricted to observed variables. The exogenous observed variables are assumed to have been measured without error and have unidirectional (*recursive*) relations with one another. As it turns out, *path analysis rules* are still used today to identify the structural equations underlying the models. Using the path analysis approach, models are presented in the form of a drawing (often called a *path diagram*), and the structural equations of the model are inferred by reading the diagram correctly. However, the term *path analysis* implies too many restrictions on the form of the model. *Structural equation modeling* (SEM), on the other hand, has grown to incorporate latent and observed variables that can be measured with and without error and have bidirectional (*nonrecursive*) relationships among variables. Another term used frequently is *causal analysis*. Unfortunately, this is also a misleading term. Although SEM may appear to imply causality, the structural equations are not causal relations but functional relations. *Covariance structure modeling* is another popular term that is used mostly by psychologists. Unfortunately, it too is restrictive. Although the covariance structure of observed data is the most commonly modeled, SEM can be used to model other moments of the data. For example, mean structures are occasionally modeled, and facilities are provided for this in a number of SEM software programs. Modeling the third (skew) and fourth (kurtosis) moments of the data is also possible.

Mathematical Representation

To date, several mathematical models for SEM have been proposed. Although these mathematical models can translate data equally well into the model parameters, they differ in how parsimoniously this translation process is conducted. Perhaps the most well known of these mathematical models, the Keesling–Wiley–Jöreskog (*LISREL*) model, can require up to nine symbols in order to represent a model. In contrast, the *COSAN* model can generally represent the same model using only two symbols. Striking

a compromise between the LISREL and COSAN models, the Benter-Weeks (*EQS*) model can represent any model using only four symbols. Mathematically, the EQS model is represented by

$$\boldsymbol{\eta} = \boldsymbol{\beta}\boldsymbol{\eta} + \boldsymbol{\gamma}\boldsymbol{\xi}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are coefficient matrices, and $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are vectors of random variables. The random variables within $\boldsymbol{\eta}$ are endogenous variables and the variables within $\boldsymbol{\xi}$ are exogenous variables. Endogenous and exogenous variables can be either latent or observed. The matrix $\boldsymbol{\beta}$ consists of coefficients (parameters) that describe the relations among the endogenous variables. The matrix $\boldsymbol{\xi}$ consists of coefficients (parameters) that describe the relations between exogenous and endogenous variables.

It is important to note that the primary interest in SEM centers on describing the network of relations among the variables (implying that one is generally interested in the covariance structure among the variables). Although the structural equation model is written in terms of equations lining the variables, the data used to solve the model parameters are actually covariances or correlations. In fact, this approach is no different from how many other multivariate statistical models are evaluated. For example, multiple regression uses a series of equations that link dependent to independent variables, but it is the correlational structure of the data that is used to solve for the regression coefficients. Similarly, in the EQS model, the sample covariance structure (C) among a set of variables x, y is defined as

$$C = (x + y)(x + y)' = \mathbf{J}(\mathbf{I} - \boldsymbol{\beta})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'(\mathbf{I} - \boldsymbol{\beta})^{-1}\mathbf{J}'$$

where $\boldsymbol{\Gamma}$ is a matrix of coefficients linking exogenous ξ with endogenous η variables, $\boldsymbol{\beta}$ is a matrix of coefficients linking endogenous variables, and $\boldsymbol{\Phi}$ represents the covariances among the exogenous variables. The $\mathbf{J}$ matrix serves as a "filter" for selecting the observed variables from the total number of variables to be included in the model.

The Confirmatory Factor Analysis Model

A popular type of structural equation model is the *confirmatory factor analysis model*. In contrast to *exploratory factor analysis* (EFA), where all loadings are free to vary, confirmatory factor analysis (CFA) allows for the explicit constraint of certain loadings to be zero. As traditionally given, the confirmatory factor model in matrix notation is

$$\mathbf{Y} = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is a vector of scores on the observed variables, $\boldsymbol{\Lambda}$ is a *factor pattern loading matrix*, $\boldsymbol{\xi}$ is a matrix of *common factors*, and $\boldsymbol{\varepsilon}$ is a matrix of measurement errors in the observed variables. As such, the covariance structure implied by the confirmatory factor model is defined as

$$C = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$$

where C is the sample variance-covariance matrix, $\boldsymbol{\Phi}$ is a matrix of the factor variance-covariances, and $\boldsymbol{\Psi}$ is a variance-covariance matrix among the measurement errors.

In the EQS representation, the confirmatory factor model is generally expressed as

$$\boldsymbol{\eta} = \boldsymbol{\beta}\boldsymbol{\eta} + \boldsymbol{\gamma}\boldsymbol{\xi} \text{ with } \boldsymbol{\beta} = 0$$

and the covariance structure implied by the model is given as

$$C(\eta\eta') = (0\eta + \gamma\xi)(0\eta + \gamma\xi)' = \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}'$$

where the asymmetric relations in the model (the effects of the common and error factors on the observed variables) are in $\boldsymbol{\Gamma}$ and the symmetric relations (the factor and error variances and covariances) are in $\boldsymbol{\Phi}$. Note that for the confirmatory factor model the matrix $\boldsymbol{\beta}$ is dropped from the EQS model because in CFA there are no regression relations between endogenous variables.

Model Estimation

Model estimation proceeds by rewriting the structural equations so that each of the parameters of the equations is a function of the elements of the sample covariance matrix C. Subsequently, after obtaining values for the parameters, it one were to substitute these values back into the expression for the covariance structure implied by the model, the resulting sample matrix C can be represented as $\widehat{C}$. Clearly, $\widehat{C}$ should be very close to C because it was the elements of C that assisted in solving for the model parameters: The difference should be small if the model is consistent with the data.

The evaluation of $C - \widehat{C}$ depends on the estimation method used to solve for the model parameters. The most commonly used estimation methods for solving the parameters are *unweighted least squares* (ULS), *generalized* (*weighted*) *least squares* (GLS), and *maximum likelihood* (ML). With each estimation method, the structural equations are solved iteratively, until optimal estimates of the parameters are obtained. Optimal parameter values are values that imply covariances $\left(\widehat{C}\right)$ close to the observed covariances (C). The difference $C - \widehat{C}$ is known as a *discrepancy function* (F). In order to minimize this discrepancy function, the partial derivatives of F are taken with respect to the elements of $C - \widehat{C}$. The form of the discrepancy function varies across the different estimation methods. However, the general form of this discrepancy function is

$$F = \sum_{ij} (C - \widehat{C})' W (C - \widehat{C})$$

in which a weighted sum of differences between the IJ elements of C and $\widehat{C}$ is calculated. As C and $\widehat{C}$ become more different, the discrepancy function becomes larger implying less correspondence between the model-implied covariances and the observed covariances. Most currently available SEM programs (e.g., SPSS' AMOS, EQS, LISREL, Mplus, Mx, the SEM package in R, SAS PROC CALIS) include ULS, GLS, and ML as standard estimation methods.

Model Assessment and Fit

For a model with positive *df* degrees of freedom, it is very unlikely that the discrepancy function will equal 0, implying a model with perfect fit to the data. Thus, there must be some measure of how large the discrepancy function must be in order to determine that the model does not fit the data. If multivariate normality is present, a *chi-square goodness-of-fit test* for the model is available using the sample size and the value of the discrepancy function

$$\chi^2 = (N-1)(F)$$

with *df* = (the number of unique elements of C) − (the number of parameters solved). If chi-square is not significant, then no significant discrepancy exists between the model-implied and observed covariance matrices. As such, the model fits the data and is confirmed. However, the chi-square test suffers from several weaknesses, including a dependence on sample size, and vulnerability to departures from multivariate normality. Thus, it is recommended that other descriptive fit criteria (e.g., ratio of χ^2 to *df*) and fit indices (e.g., the *comparative fit index*, the *root mean square error of approximation*) be examined in addition to the χ^2 value to assess the fit of the proposed model. Quite a few fit criteria and indices have been developed, each with its own strengths and weaknesses, and is it usually advisable to report a range of them.

Model Identification

Only identified models should be estimated. The process of *model identification* involves confirming that a unique numerical solution exists for each of the parameters of the model. Model identification should be distinguished from *empirical identification*, which involves assessing whether the rank of the *information matrix* is not deficit. Most SEM programs automatically check for empirical identification. On the other, model identification is not as easily or automatically assessed. For structural equation models in general, the most frequently invoked identification rules are the t-rule and the rank and order conditions. The *t*-rule is a simple rule to apply, but is only a necessary not a sufficient condition of identification. The *t*-rule is that the number of nonredundant elements in the covariance matrix of the observed variables (p) must be greater than or equal to the number of unknown parameters in the proposed model. Thus, if $t \leq p(p+1)/2$ the necessary condition of identification is met. Unfortunately, although the *t*-rule is simple to apply, it is only good for determining *underidentified* models. The order condition requires that for the model to be identified, the number of p variables excluded from each structural equation must equal $p - 1$. Unfortunately, the order condition is also a necessary bit not sufficient condition for identification. Only the rank condition is a necessary and sufficient condition for identification; however, it is not easy to apply. In general terms, the rank condition requires that the rank of any model matrices (e.g., $\mathbf{\Phi}, \boldsymbol{\beta}, \mathbf{\Gamma}$) be of at least rank $p-1$ for all submatrices formed by removing the parameter of interest. However, the usefulness of these criteria is doubtful because a failure to meet them does not necessarily mean the model is not identified. As it turns out, the only sure way to assess the identification status of a model prior to model fitting is to show through algebraic manipulation that each of the model parameters can be solved in terms of the p variances and $p(p-1)/2$ covariances.

Equivalent Structural Equation Models

Equivalent structural equation models may be defined as the set of models that, regardless of the data, yield identical (a) implied covariance, correlation, and other moment matrices when fit to the same data, which in turn imply identical (b) residuals and fitted moment matrices, (c) fit functions and chi-square values, and (d) goodness-of-fit indices based on fit functions and chi-square. One most frequently thinks of equivalent models as described in (a) above. To be precise, consider two alternative models, denoted $M1$ and $M2$, each of which is associated with a set of estimated parameters and a covariance implied by those parameter estimates (denoted as $\widehat{C}_{M1}$ and $\widehat{C}_{M2}$). Models $M1$ and $M2$ are considered equivalent if, for any sample covariance matrix C, the implied matrices $\widehat{C}_{M1} = \widehat{C}_{M2}$ or alternatively, $(C - \widehat{C}_{M1}) = (C - \widehat{C}_{M2})$. Because of this equivalence, the values of statistical tests of fit that are based on the discrepancy between the sample covariance matrix and the model-implied covariance matrix will be identical. Thus, even when a hypothesized model fits well according to multiple fit indices, there may be equivalent models with identical fit – even if the theoretical implications of those models are very different. However, model equivalence is not unique to SEM. For example, in *exploratory factor analysis*, without the arbitrary constraint of extracting orthogonal factors in decreasing order of magnitude,

there would potentially be an infinite number of equivalent initial solutions.

About the Author

Scott L. Hershberger, Ph.D. is formerly Quantitative Professor of Psychology at the California State University, Long Beach and is now Global Director of Survey Design at Harris Interactive. He is a past Associate editor of the journal, *Structural Equation Modeling*, and is an elected member of the Royal Statistical Society and the International Statistical Institute. He has authored or co-authored numerous articles and several books on multivariate analysis and psychometrics.

Cross References

►Causal Diagrams
►Causation and Causal Inference
►Chi-Square Goodness-of-Fit Tests: Drawbacks and Improvements
►Chi-Square Tests
►Factor Analysis and Latent Variable Modelling
►Multivariate Data Analysis: An Overview
►Multivariate Statistical Analysis
►Psychiatry, Statistics in
►Sociology, Statistics in

References and Further Reading

Bentler P (1995) EQS program manual. Multivariate Software, Encino
Bollen KA (1989) Structural equation models with latent variables. Wiley, New York
Bollen KA, Long JS (eds) (1993) Testing structural equation models. Sage, Newbury Park
Hoyle RH (ed) (1995) Structural equation modeling: concepts, issues, and applications. Sage, Thousand Oaks
Raykov T, Marcoulides GA(2006) A first course in structural equation modeling, 2nd edn. Lawrence Erlbaum Associates, Mahwah
Wright S (1921) Correlation and causation. J Agr Res 20:557–585

Structural Time Series Models

Juan Carlos Abril
President of the Argentinean Statistical Society, Professor Universidad Nacional de Tucumán, San Miguel de Tucumán, Argentina

Introduction

The basic idea of structural time series models is that they are set up as regression models in which the explanatory variables are functions of time with coefficients which change over time. Thus within a regression framework a simple trend would be modeled in terms of a constant and a time with a random disturbance added on, that is

$$y_t = \alpha + \beta t + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{1}$$

This model is easy to estimate using ordinary ►least squares, but suffers from the disadvantage that the trend is deterministic. In general, this is too restrictive, however, the necessary flexibility is introduced by letting the coefficients α and β evolve over time as stochastic processes. In this way the trend can adapt to underlying changes. The current, or *filtered*, estimate of the trend is estimated by putting the model in state space form and applying the Kalman filter. Related algorithms are used for making *predictions* and for *smoothing*, which means computing the best estimate of the trend at all points in the sample using the full set of observations. The extent to which the parameters are allowed to change is governed by *hyperparameters*. These can be estimated by maximum likelihood but, again, the key to this is the state space form and the Kalman filter. The STAMP package of Koopman et al. (2000) carries out all the calculations and is set up so as to leave the user free to concentrate on choosing a suitable model.

An excellent general presentation of the Kalman filter is given in this Encyclopedia by M. S. Grewal under the title *Kalman Filtering*. We give below a set of particular results about the filter that are for application within the areas covered by Time Series and Econometric. Similarly, a general presentation of smoothing is given as well in this Encyclopedia by A.W. Bowman under the title *Smoothing Techniques*. We recall that in our context smoothing means computing the best estimates based on the full sample, therefore we give below a set of particular results that are for application within the areas covered by Time Series and Econometric.

The classical approach to time series modeling is based on the fact that a general model for any indeterministic stationary series is the autoregressive-moving average of order (p, q). This is usually referred to as ARMA(p, q). The modeling strategy consists of first specifying suitable values of p and q on the basis of an analysis of the correlogram and other relevant statistics. The model is then estimated, usually under the assumption that the disturbance is Gaussian. The residuals are then examined to see if they appear to be random, and various test statistics are computed. In particular, the Box–Ljung Q-statistic, which is based on the first P residual autocorrelations, is used to test for residual serial correlation. Box and Jenkins (1976) refer to these stages as identification, estimation and diagnostic checking. If the diagnostic checks are satisfactory, the model is ready to be used for forecasting. If they are not, another specification must be tried. Box and Jenkins stress the role

of parsimony in selecting p and q to be small. However, it is sometimes argued, particularly in econometrics, that a less parsimonious pure autoregressive (AR) model is often to be preferred as it is easier to handle.

Many series are not stationary. In order to handle such situations Box and Jenkins proposed that a series be differenced to make it stationary. After fitting an ARMA model to the differenced series, the corresponding integrated model is used for forecasting. If the series is differenced d times, the overall model is called ARIMA(p, d, q). Seasonal effects can be captured by seasonal differencing.

The model selection methodology for structural models is somewhat different in that there is less emphasis on looking at the correlograms of various transformations of the series in order to get an initial specification. This is not to say that correlograms should never be examined, but the experience is that they can be difficult to interpret without prior knowledge of the nature of the series and in small samples and/or with messy data they can be misleading. Instead the emphasis is on formulating the model in terms of components which knowledge of the application or an inspection of the graph suggests might be present. For example, with monthly observations, one would probably wish to build a seasonal pattern into the model at the outset and only drop it if it proved to be insignificant. Once a model has been estimated, the same type of diagnostics tests as are used for ARIMA models can be performed on the residuals. In particular the Box–Ljung statistic can be computed, with the number of relative hyperparameters subtracted from the number of residual autocorrelations to allow for the loss of degrees of freedom. Standard tests for non-normality and heteroscedasticity can also be carried out, as can tests of predictive performance in a post-sample period. Plots of residuals should be examined, a point which Box and Jenkins stress for ARIMA model building. In a structural time series model, such plots can be augmented by graphs of the smoothed components. These can often be very informative since it enables the model builder to check whether the movements in the components correspond to what might be expected on the basis of prior knowledge.

State Space Form, Kalman Filtering and Smoothing

As we say before, a structural time series model is one in which the trend, seasonal and error terms in the basic model, plus other relevant components, are modeled explicitly. This is in sharp contrast to the philosophy underlying ARIMA models where trend and seasonal are removed by differencing prior to detailed analysis.

The statistical treatment of the structural time series models is based on the state space form, the Kalman filter and the associated smoother. The likelihood is constructed from the Kalman filter in terms of the one-step ahead prediction errors and maximized with respect to the hyperparameters by numerical optimization. The score vector for the parameters can be obtained via a smoothing algorithm which is associated with the Kalman filter. Once the hyperparameters have been estimated, the filter is used to produced one-step ahead predictions residuals which enables us to compute diagnostic statistics for normality, serial correlation and goodness of fit. The smoother is used to estimate unobserved components, such as trends and seasonals, and to compute diagnostic statistics for detecting ►outliers and structural breaks. ARIMA models can also be handled using the Kalman filter. The state space approach becomes particularly attractive when the data are subject to missing values or temporal aggregation.

State Space Form

All linear time series have a state space representation. This representation relates the disturbance vector $\{\boldsymbol{\varepsilon}_t\}$ to the observation vector $\{\mathbf{y}_t\}$ via a Markov process (see ►Markov Processes) $\{\boldsymbol{\alpha}_t\}$. A convenient expression of the state space form is

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t &\sim N(\mathbf{0}, \mathbf{H}_t), \\ \boldsymbol{\alpha}_t &= \mathbf{T}_t\boldsymbol{\alpha}_{t-1} + \mathbf{R}_t\boldsymbol{\eta}_t, & \boldsymbol{\eta}_t &\sim N(\mathbf{0}, \mathbf{Q}_t), \qquad t = 1, \ldots n, \end{aligned} \tag{2}$$

where $\mathbf{y}_t$ is a $p \times 1$ vector of observations and $\boldsymbol{\alpha}_t$ is an unobserved $m \times 1$ vector called the *state vector*. The idea underlying the model is that the development of the system over time is determined by $\boldsymbol{\alpha}_t$ according to the second equation of (2), but because $\boldsymbol{\alpha}_t$ cannot be observed directly we must base the analysis on observations $\mathbf{y}_t$. The first equation of (2) is called the *measurement equation*, and the second one, the *transition equation*. The system matrices $\mathbf{Z}_t$, $\mathbf{T}_t$ and $\mathbf{R}_t$ have dimensions $p \times m$, $m \times m$ and $m \times g$ respectively. The disturbance terms $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are assume to be serially independent and independent of each other at all time points. The matrix $\mathbf{H}_t$ has dimension $p \times p$ with rank p, and the matrix $\mathbf{Q}_t$ has dimension $g \times g$ with rank $g \leq m$. The matrices $\mathbf{Z}_t, \mathbf{T}_t, \mathbf{R}_t, \mathbf{H}_t$ and $\mathbf{Q}_t$ are fixed and their unknown elements, if any, are place in the hyperparameter vector $\boldsymbol{\psi}$ which can be estimated by maximum likelihood. In univariate time series $p = 1$, so $\mathbf{Z}_t$ is a row vector.

The initial state vector $\boldsymbol{\alpha}_0$ is assumed to be $N(\mathbf{a}_0, \mathbf{P}_0)$ where $\mathbf{a}_0$ and $\mathbf{P}_0$ are known. When $\mathbf{a}_0$ and $\mathbf{P}_0$ are unknown, $\boldsymbol{\alpha}_0$ is taken as diffuse. An adequate approximation can

often be achieved numerically by taking $\mathbf{a}_0 = \mathbf{0}$ and $\mathbf{P}_0 = \kappa \mathbf{I}_m$, where κ is a scalar which tends to infinity.

Kalman Filter

In the Gaussian state space model (2), the Kalman filter evaluate the minimum mean squared error estimator of the state vector $\boldsymbol{\alpha}_{t+1}$ using the set of observations $\mathbf{Y}_t = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$, denoted $\mathbf{a}_{t+1} = E(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_t)$, and the corresponding variance matrix $\mathbf{P}_{t+1} = \mathrm{Var}(\boldsymbol{\alpha}_{t+1}|\mathbf{Y}_t)$, for all t. This means that the Kalman filter allows to continuously update the estimation of the state vector whenever a new observation is available. Since all distributions are normal, conditional distributions are also normal. Let $\mathbf{v}_t = \mathbf{y}_t - \mathbf{Z}_t\mathbf{a}_t$, then $\mathbf{v}_t$ is the one-step ahead forecast error $\mathbf{y}_t - E(\mathbf{y}_t|\mathbf{Y}_{t-1})$. Demote its variance matrix by $\mathbf{F}_t$. Then

$$\mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_t\mathbf{Z}_t' + \mathbf{H}_t, \quad t = 1, \ldots n. \tag{3}$$

It is possible to show that the updating recursion is given by

$$\mathbf{a}_{t+1} = \mathbf{T}_{t+1}\mathbf{a}_t + \mathbf{K}_t\mathbf{v}_t, \tag{4}$$

where

$$\mathbf{K}_t = \mathbf{T}_{t+1}\mathbf{P}_t\mathbf{Z}_t'\mathbf{F}_t^{-1}, \tag{5}$$

and

$$\mathbf{P}_{t+1} = \mathbf{T}_{t+1}\mathbf{P}_t\left(\mathbf{T}_{t+1}' - \mathbf{Z}_t'\mathbf{K}_t'\right) + \mathbf{R}_{t+1}\mathbf{Q}_{t+1}\mathbf{R}_{t+1}', \tag{6}$$

for $t = 0, 1, \ldots, n-1$, with $\mathbf{K}_0 = \mathbf{0}$.

The set (3) to (6) constitute the Kalman filter for model (2). The derivation of the Kalman recursions can be found in Anderson and Moore (1979), Harvey (1989), Abril (1999) and Durbin and Koopman (2001).

The output of the Kalman filter is used to compute the log-likelihood function $\log L(\mathbf{y}_t, \boldsymbol{\psi})$, conditional on the hyperparameter vector $\boldsymbol{\psi}$, as given by

$$\log L(\mathbf{y}_t, \boldsymbol{\psi}) = -\frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\log|\mathbf{F}_t| - \frac{1}{2}\sum_{t=1}^{n}\mathbf{v}_t'\mathbf{F}_t^{-1}\mathbf{v}_t, \tag{7}$$

apart from a possible constant. Numerical maximization of (7) with respect to the hyperparameter vector $\boldsymbol{\psi}$ yields the maximum likelihood estimator $\widetilde{\boldsymbol{\psi}}$. Usually (7) is called the *prediction error decomposition* of the likelihood.

Smoothing

The work of de Jong (1988, 1989), Kohn and Ansley (1989) and Koopman (1993) leads to a smoothing algorithm from which different estimators can be computed based on the full sample $\mathbf{Y}_n$. Smoothing takes the form of a backwards recursion

$$\mathbf{u}_t = \mathbf{F}_t^{-1}\mathbf{v}_t - \mathbf{K}_t'\mathbf{r}_t, \qquad \mathbf{M}_t = \mathbf{F}_t^{-1} + \mathbf{K}_t'\mathbf{N}_t\mathbf{K}_t,$$
$$\mathbf{r}_{t-1} = \mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{v}_t + \mathbf{L}_t'\mathbf{r}_t, \quad \mathbf{N}_{t-1} = \mathbf{Z}_t'\mathbf{F}_t^{-1}\mathbf{Z}_t + \mathbf{L}_t'\mathbf{N}_t\mathbf{L}_t, \tag{8}$$

for $t = n, n-1, \ldots, 1$, where $\mathbf{L}_t = \mathbf{T}_{t+1} - \mathbf{K}_t\mathbf{Z}_t$, $\mathbf{r}_n = \mathbf{0}$ and $\mathbf{N}_n = \mathbf{0}$. The recursions require memory space for storing the Kalman output $\mathbf{v}_t$, $\mathbf{F}_t$ and $\mathbf{K}_t$ for $t = 1, \ldots, n$. The series $\{\mathbf{u}_t\}$ will be referred to as *smoothing errors*. The smoothing quantities $\mathbf{u}_t$ and $\mathbf{r}_t$ play a pivotal role in the construction of diagnostic tests for outliers and structural breaks. The smoother can be used to compute the smoothed estimator of the disturbance vector $\widehat{\boldsymbol{\varepsilon}}_t = E(\boldsymbol{\varepsilon}_t|\mathbf{Y}_n)$. The smoothed estimator of the state vector $\widehat{\boldsymbol{\alpha}}_t = E(\boldsymbol{\alpha}_t|\mathbf{Y}_n)$ is constructed as follows

$$\widehat{\boldsymbol{\alpha}}_t = \mathbf{a}_t + \mathbf{P}_t\mathbf{r}_{t-1}, \tag{9}$$

for $t = 1, \ldots, n$, where $\mathbf{r}_t$ satisfies the backwards recursions given in (8).

About the Author

Professor Abril is co-editor of the *Revista de la Sociedad Argentina de Estadística* (Journal of the Argentinean Statistical Society).

Cross References

- ▶Autocorrelation in Regression
- ▶Box–Jenkins Time Series Models
- ▶Forecasting with ARIMA Processes
- ▶Kalman Filtering
- ▶Markov Processes
- ▶Model Selection
- ▶Residuals
- ▶Time Series
- ▶Trend Estimation

References and Further Reading

Abril JC (1999) Análisis de Series de Tiempo Basado en Modelos de Espacio de Estado. EUDEBA, Buenos Aires

Anderson BDO, Moore JB (1979) Optimal filtering. Prentice-Hall, Englewood Cliffs, New Jersey

Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control (revised edition), Holden-Day, San Francisco

de Jong P (1988) A cross-validation filter for time series models. Biometrika 75:594–600

de Jong P (1989) Smoothing and interpolation with the state-space model. J Am Stat Assoc 84:1085–1088

Durbin J, Koopman SJ (2001) Time series analysis by state space methods. Oxford University Press, Oxford

Harvey AC (1989) Forecasting, structural time series models and the kalman filter. Cambridge University Press, Cambridge

Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans ASME, J Basic Eng 83D:35–45

Kalman RE, Bucy RS (1961) New results in linear filtering and prediction problems. Trans ASME, J Basic Eng 83D:95–108
Kohn R, Ansley CF (1989) A fast algorithm for signal extraction, influence and cross-validation in state space models. Biometrika 76:65–79
Koopman SJ (1993) Disturbance smoother for state space models. Biometrika 80:117–126
Koopman SJ, Harvey AC, Doornik JA, Shephard N (2000) STAMP: structural time series analyser, modeller and predictor. Timberlake Consultant Ltd, London

Student's *t*-Distribution

Bronius Grigelionis
Professor, Head of the Mathematical Statistics Department
Institute of Mathematics and Informatics, Vilnius, Lithuania

We say that a random variable X has a Student t distribution with $\nu > 0$ degrees of freedom, a scaling parameter $\delta > 0$ and a location parameter $\mu \in R^1$, denoted $T(\nu, \delta, \mu)$, if its probability density function (pdf) is

$$f_X(x) = \frac{\Gamma\left(\frac{1}{2}(\nu+1)\right)}{\sqrt{\pi}\delta\Gamma\left(\frac{1}{2}\nu\right)}\left[1+\left(\frac{x-\mu}{\delta}\right)^2\right]^{-\frac{\nu+1}{2}}, \quad x \in R^1,$$

where $\Gamma(z)$ is the Euler's gamma function. $T(1, \delta, \mu)$ is the Cauchy distribution. $T(\nu, \delta, \mu)$ is heavy tailed and for an integer r

$$E(X-\mu)^{2r} = \begin{cases} \frac{\delta^{2r-1}\nu^r\Gamma\left(\frac{r}{2}+1\right)\Gamma\left(\frac{\nu}{2}-r\right)}{\sqrt{\pi}\Gamma\left(\frac{1}{2}\nu\right)}, & \text{if } 2r < \nu, \\ +\infty, & \text{if } 2r \geq \nu. \end{cases}$$

Because

$$f_X(x) = \int_0^\infty \frac{1}{\sqrt{2\pi y}} e^{-\frac{(x-\mu)^2}{2y}} g(y)\,dy, \quad x \in R^1,$$

where

$$g(y) = \frac{\left(\frac{1}{2}\delta^2\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{1}{2}\nu\right)} y^{-\frac{\nu}{2}-1} e^{-\frac{\delta^2}{2y}}\,dy, \quad y > 0$$

is pdf of the inverse (reciprocal) gamma distribution, which is a member of the Thorin class, the Student t distribution is a marginal distribution of a Thorin subordinated Gaussian Lévy process (see, e.g., Grigelionis, 2007 and references therein). This property implies that $T(\nu, \delta, \mu)$ is self-decomposable, i.e., for every $c \in (0,1)$, there exists a random variable X_c, independent of X, such that $X \overset{\text{law}}{=} cX + X_c$, and therefore $T(\nu, \delta, \mu)$ is infinitely divisible. Self-decomposability of $T(\nu, \delta, \mu)$ permits to construct several classes of stationary stochastic processes with marginal Student t distributions and various types of dependence structure, relevant for modeling of economic and financial time series. In the fields of finance Lévy processes with marginal Student t distributions can often be fitted extremely well to model distributions of logarithmic asset returns (see Heyde and Leonenko, 2005).

The classical Student t distribution was introduced in 1908 by W.S. Gosset ("Student"), proving that the distribution law $\mathcal{L}(t_n) = T\left(n-1, \sqrt{n-1}, 0\right)$, where

$$t_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n}, \quad n \geq 2,$$

$X_1, \ldots, X_n$ are independent normally distributed random variables, $\mathcal{L}(X_i) = N(\mu, \sigma^2)$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Properties of the classical Student t distributions are surveyed in Johnson, Kotz, 1970.

During last century the theory of Student t statistics has evolved into the theory of general Studentized statistics and self-normalized processes, and the Student t distribution was generalized to the multivariate case, leading to multivariate processes with matrix self-normalization (see de la Peña et al., 2009).

We say that a random d-dimensional vector X has a Student t distribution with $\nu > 0$ degrees of freedom, a symmetric positive definite scaling $d \times d$ matrix Σ and a location vector $\mu \in R^d$, denoted $T_d(\nu, \Sigma, \mu)$, if its pdf is

$$f_X(x) = \frac{\Gamma\left(\frac{1}{2}(\nu+d)\right)}{(\nu\pi)^{d/2}\Gamma\left(\frac{1}{2}\nu\right)|\Sigma|^{1/2}} \times \left(1+\frac{\left((x-\mu)\Sigma^{-1}, x-\mu\right)}{\nu}\right)^{-\frac{\nu+d}{2}}, \quad x \in R^d,$$

where $(x, y) = \sum_{i=1}^d x_i y_i$, $x, y \in R^d$, $|\Sigma| := \det\Sigma$ (see Johnson and Kotz, 1972).

We have that

$$Ee^{i(z,X)} = \frac{e^{i(\mu,z)}}{2^{\frac{\nu}{2}-1}\Gamma\left(\frac{1}{2}\nu\right)} \times \left(\nu(z\Sigma, z)\right)^{\frac{\nu}{4}} K_{\frac{\nu}{2}}\left(\sqrt{\nu(z\Sigma, z)}\right), \quad z \in R^d,$$

where K_ν is the modified Bessel function of the third kind, i.e.,

$$K_\nu(x) = \frac{1}{2}\int_0^\infty u^{-\nu-1}\exp\left\{-\frac{1}{2}x(u+u^{-1})\right\}du, \quad x > 0,\ \nu \in R^1,$$

implying that for $c \in R^d$, $c \neq 0$, $\mathcal{L}((c,X)) = T\left(\nu, \sqrt{\nu\,(c\Sigma,c)},\ (c,\mu)\right)$, which means that $T_d(\nu,\Sigma,\mu)$ is marginal self-decomposable (see, Barndorff–Nielsen and Pérez-Abreu, 2002).

If $\nu > d+1$, $EX = \mu$ and $E(c_1, X-\mu)(c_2, X-\mu) = \nu(c_1\Sigma^{-1}, c_2)\Gamma\left(\frac{\nu-d-1}{2}\right)$, $c_1, c_2 \in R^d$.

As $\nu \to \infty$, $T_d(\nu,\Sigma,\mu) \Rightarrow N_d(\mu,\Sigma)$ and, in particular, $T\left(\nu, \sqrt{\nu}\sigma, \mu\right) \Rightarrow N(\mu,\sigma^2)$, where "$\Rightarrow$" means weak convergence of probability laws.

Let M_d be an Euclidean space of symmetric $d\times d$ matrices with the scalar product $\langle A_1, A_2\rangle := \mathrm{tr}(A_1A_2)$, $A_1, A_2 \in M_d$, $M_d^+ \subset M_d$ be the cone of non-negative definite matrices, $\mathcal{P}\left(M_d^+\right)$ be the class of probability distributions on M_d^+.

Since

$$Ee^{i(z,X)} = e^{i(z,\mu)} \int\limits_{M_d^+} e^{-\frac{1}{2}(zA,z)} U(\,dA),$$

where

$$\phi_U(\Theta) := \int\limits_{M_d^+} e^{-\mathrm{tr}(\Theta A)} U(\,dA)$$

$$= \frac{[2\nu\mathrm{tr}(\Sigma\Theta)]^{\frac{\nu}{4}}}{2^{\frac{\nu}{2}-1}\Gamma\left(\frac{1}{2}\nu\right)} K_{\frac{\nu}{2}}\left(\sqrt{2\nu\mathrm{tr}(\Sigma\Theta)}\right),$$

$$\Theta \in M_d^+,\ U \in \mathcal{P}\left(M_d^+\right),$$

$\mathcal{L}(X-\mu)$ is a U-mixture of centered Gaussian distributions (see Grigelionis, 2009).

If $\nu \geq d$ is an integer, $U = \mathcal{L}\left(\nu W_\nu^{-1}\right)$, where $W_\nu = \sum_{i=1}^{\nu} Y_i^T Y_i$, $Y_1, \ldots, Y_\nu$ are independent d-dimensional centered Gaussian vectors with the covariance matrix Σ, z^T is the transposed vector z, i.e., U is the inverse Wishart distribution.

About the Author

Bronius Grigelionis graduated from the Department of Physics and Mathematics, Vilnius University in 1959. He was a postgraduate student at the Kiev University (1959–1960) and Moscow University (1960–1962) supervised by Prof. B.V. Gnedenko. He earned a doctor's (Ph.D.) in 1963 and the degree of Doctor habilius in 1969 at the Vilnius University. He was a senior Research Fellow at the Institute of Physics and Mathematics and a lecturer of the Vilnius University in 1963–1970. Since 1970 he has been Head of the Mathematical Statistics Department at the Institute of Mathematics and Informatics and Professor of Vilnius University. He is a member of the Lithuanian Mathematics Society, Lithuanian Academy of Sciences, International Statistical Institute, Bernoulli Society, Lithuanian Catholic Academy of Sciences. He has supervised 19 Ph.D. students.

Cross References

►Confidence Interval
►Correlation Coefficient
►Financial Return Distributions
►Heteroscedastic Time Series
►Hotelling's T^2 Statistic
►Multivariate Statistical Distributions
►Regression Models with Symmetrical Errors
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Statistical Distributions: An Overview
►Student's *t*-Tests

References and Further Reading

Barndorff-Nielsen OE (2002) Pérez-Abreu V (2002). Extensions of type G and marginal infinite divisibility. Teor Veroyatnost i Primenen 47(2):301–319

de la Peña VH, Lai TL, Shao QM (2009) Self-normalized processes: limit theory and statistical applications. Springer, Berlin

Grigelionis B (2007) On subordinated multivariate Gaussian Lévy processes. Acta Appl Math 96:233–246

Grigelionis B (2009) On the Wick theorem for mixtures of centered Gaussian distributions. Lith Math J 49(4):372–380

Heyde CC, Leonenko NN (2005) Student processes. Adv Appl Prob 37:342–365

Johnson NL, Kotz S (1970) Distributions in statistics: continuous univariate distributions vol 2. Wiley, New York

Johnson NL, Kotz S (1972) Distributions in statistics: continuous multivariate distributions. Wiley, New York

Student (1908) On the probable error of mean. Biometrika 6:1–25

Student's *t*-Tests

Damir Kalpić[1], Nikica Hlupić[2], Miodrag Lovrić[3]
[1]Professor and Head, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia
[2]Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, Croatia
[3]Professor, Faculty of Economics
University of Kragujevac, Kragujevac, Serbia

Introduction

Student's *t*-tests are parametric tests based on the Student's or *t*-distribution. Student's distribution is named in honor of William Sealy Gosset (1876–1937), who first determined it in 1908. Gosset, "one of the most original minds in contemporary science" (Fisher 1939), was one of the best Oxford graduates in chemistry and mathematics in his generation. In 1899, he took up a job as a brewer

at Arthur Guinness Son & Co, Ltd in Dublin, Ireland. Working for the Guinness brewery, he was interested in quality control based on small samples in various stages of the production process. Since Guinness prohibited its employees from publishing any papers to prevent disclosure of confidential information, Gosset had published his work under the pseudonym "Student" (the other possible pseudonym he was offered by the managing director La Touche was "Pupil," see Box 1987, p. 49), and his identity was not known for some time after the publication of his most famous achievements, so the distribution was named Student's or *t*-distribution, leaving his name less well known than his important results in statistics. His, now, famous paper "The Probable Error of a Mean" published in *Biometrika* in 1908, where he introduced the *t*-test (initially he called it the *z*-test), was essentially ignored by most statisticians for more than 2 decades, since the "statistical community" was not interested in small samples ("only naughty brewers take *n* so small," Karl Pearson writing to Gosset, September 17, 1912, quoted by E.S. Pearson 1939, p. 218). It was only R. Fisher who appreciated the importance of Gosset's small-sample work, and who reconfigured and extended it to two independent samples, correlation and regression, and provided correct number of degrees of freedom. "It took the genius and drive of a Fisher to give Student's work general currency" (Zabel 2008, p. 6); "The importance of 1908 article is due to what Fisher found there, not what Gosset placed there" (Aldrich 2008, p. 11).

One-Sample *t*-Test

In the simplest form, also called the one-sample *t*-test, Student's *t*-test is used for testing a statistical hypothesis (Miller and Miller 1999) about the mean μ of a normal population whose variance σ^2 is unknown and sample size n is relatively small ($n \leq 30$). For a comparison of means of two independent univariate normal populations with equal (but unknown) variances we use two-sample *t*-test, and both of these tests have their multivariate counterparts based on multivariate extension of the *t*-variable called Hotelling's T^2 statistic ▶Hotelling's T^2 statistic (Johnson and Wichern 2007). Student's *t*-test also serves as the basis for the analysis of dependent samples (populations) in paired difference *t*-test or repeated measures design, in both univariate (Bhattacharyya and Johnson 1977) and multivariate cases (Johnson and Wichern 2007).

To understand the motivation for Student's *t*-test, suppose that we have at our disposal a relatively large sample of size $n > 30$ from a normal population with unknown mean μ and known variance σ^2. What we want is to determine the mean μ, i.e., to test our supposition (null hypothesis) $H_0 : \mu = \mu_0$ against one of the alternative hypotheses $\mu \neq \mu_0$ or $\mu > \mu_0$ or $\mu < \mu_0$. Maximum likelihood principle (method) (Hogg et al. 2005, or Anderson 2003) leads to the sample mean $\overline{X}$ as the test statistic, and it is known that $\overline{X}$ has Gaussian or normal distribution with mean μ and variance σ^2/n. Hence, we might calculate (provided σ^2) the probability of observing $\overline{x}$ in a certain range under the assumption of the supposed distribution $N(\mu_0, \sigma^2/n)$ and thereby assess our supposition about the unknown μ. Yet, this would require (numerical) evaluation of the integral of normal density for every particular pair (μ_0, σ^2) and, therefore, we construct the universal standard normal variable or *z*-score

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}, \tag{1}$$

which, in our example, represents the distance from the observed $\overline{X}$ to the hypothesized population mean μ_0, expressed in terms (units) of standard deviation $\sigma/\sqrt{n}$ of $\overline{X}$. Thus, variable Z is an independent parameter and it has a standard normal distribution that has been extensively tabulated and is readily available in statistical books and software. The test itself is now based on Z as the test statistic and the rationale behind the test is that if the null hypothesis is true, then the larger the distance from $\overline{x}$ to μ_0 (larger $|z|$-value), the smaller the probability of observing such an $\overline{x}$. Therefore, given a level of significance α, we reject H_0 if $|z| \geq z_{\alpha/2}$, $z \geq z_\alpha$ or $z \leq -z_\alpha$, respectively, where z_α is the Z-value corresponding to the probability α for a random variable having standard normal distribution to take a value greater than z_α, i.e., $P(z \geq z_\alpha) = \alpha$. By virtue of the central limit theorem (Anderson 2003) and provided that the sample is large enough ($n > 30$), we apply the same test even though the population distribution cannot be assumed to be normal, the only precondition being that the variance is known. Of course, in real applications we rarely know exact population variance σ^2, so we substitute sample variance S^2

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2}$$

for σ^2 and likelihood ratio test statistic (1) becomes Students's *t*-variable

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}. \tag{3}$$

Having at our disposal a sufficiently large sample ($n > 30$), we consider s to be a "faithful" estimate of σ and we might still apply the same test, i.e., compare t with z_α values. This would then be only an approximate large-sample test, but

its result would likely correspond to the real truth. However, when population variance σ^2 is not known and the sample size is relatively small ($n \leq 30$), the test we have been discussing is not reliable anymore because t in (3) is not a faithful approximation of z in (1), as a direct consequence of the fact that sample variance S^2 determined from too small a sample does not approximate σ^2 well. Construction of a reliable test under such conditions requires knowledge of the exact distribution of variable T in (3), and due to Gosset, we know that it is a t-distribution with $n-1$ degrees of freedom. The same as with z-test, the rationale behind the t-test is that if the null hypothesis is true, then observing $\bar{x}$ too much distant from μ_0 is not likely. Specifically, for a given level of significance α and one of the alternatives $\mu \neq \mu_0$ or $\mu < \mu_0$ or $\mu > \mu_0$, following the Neyman–Pearson approach, we calculate the critical value $t_{n-1}(\alpha/2)$ or $t_{n-1}(\alpha)$ defined by $P(t \geq t_{n-1}(\alpha)) = \alpha$, i.e., $t_{n-1}(\alpha)$ is the value corresponding to probability α for a random variable having t-distribution to take a value greater than $t_{n-1}(\alpha)$, and

$$\begin{aligned}\text{reject } H_0 \text{ if } |t| \geq t_{n-1}(\alpha/2) \quad & \text{with the alternative} \\ & \text{hypothesis } \mu \neq \mu_0, \\ t \geq t_{n-1}(\alpha) \quad & \text{with the alternative} \\ & \text{hypothesis } \mu > \mu_0, \\ t \leq -t_{n-1}(\alpha) \quad & \text{with the alternative} \\ & \text{hypothesis } \mu < \mu_0. \end{aligned} \tag{4}$$

Statistical tests imply *reject–do not reject* results, but it is usually more informative to express conclusions in the form of confidence intervals. In the case of the two-sided t-test ($H_1 : \mu \neq \mu_0$) constructed from a random sample of size n, $(1-\alpha)100\%$ confidence interval for the mean of a normal population is

$$\bar{x} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}. \tag{5}$$

Two-Sample *t*-Test

When we compare parameters of two populations (means, variances, or proportions), we need to distinguish two cases: samples may be independent or dependent according to how they were selected. Two random samples are *independent* if the sample selected from one population is not related in any way to the sample from the other population. However, if the random samples are chosen in such a way that each measurement in one sample can be naturally or by design paired or matched with a measurement in the other sample, then the samples are called *dependent*. Dependent samples occur in two situations:

(a) *Repeated measures design*, when the same subject or unit is measured twice, *before and after* a treatment (e.g., the blood pressure of each subject in the study is recorded twice, before and after a drug is administered)

(b) *Matched pairs design*, when subjects are *matched* as closely as possible, and then one of each pair is randomly assigned to each of the treatment group and control group (see ▶Research Designs).

Two Independent Samples

(a) *Equal variances* $\sigma_1^2 = \sigma_2^2 = \sigma^2$

This is a simpler situation because variances of considered populations, though unknown, are equal. With the respective sample sizes being n_1 and n_2, maximum likelihood principle yields a test based on test statistic

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{6}$$

where S_p^2 is the pooled estimator of common variance σ^2 given by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{7}$$

The pooled t-test is based on the fact that variable T in (6) has Student's distribution with $n_1 + n_2 - 2$ degrees of freedom, i.e., $P(t \geq t_{n_1+n_2-2}(\alpha)) = \alpha$. Hence, for instance, we reject the null hypothesis that both population means are equal ($H_0 : \mu_1 = \mu_2$) if $|t| \geq t_{n_1+n_2-2}(\alpha/2)$.

(b) *Unequal variances* $\sigma_1^2 \neq \sigma_2^2$

When the assumption of equal variances is untenable, we are confronted with what is known as ▶Behrens–Fisher problem, which is still an open challenge. There are, however, approximate solutions and a commonly accepted technique is Welch's t-test, also referred to as Welch–Aspin, Welch–Satterthwaite, or Smith–Satterthwaite test (Winer 1971; Johnson 2005). The test statistic is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{8}$$

and it has approximately t-distribution with degrees of freedom estimated as

$$\nu = \frac{(g_1 + g_2)^2}{g_1^2/(n_1 - 1) + g_2^2/(n_2 - 1)}; \quad g_i = \frac{s_i^2}{n_i}. \tag{9}$$

The difference between the denominators in (6) and (8) should be noticed; in (6) we have the *estimate of the common variance*, while in (8) we have the *estimate of variance of the difference.*

The test procedure is to calculate the value t of the test statistics given by (8) and degrees of freedom v according to (9) (if v is not an integer we round it down rather than up in order to take a conservative approach). Then, given the level of significance α, we use the obtained v and Student's distribution to calculate critical value $t_v(\alpha)$ and draw conclusions comparing t and $t_v(\alpha)$ like in an ordinary one-sample t-test.

Two Dependent Samples

The test procedure is essentially the same as for one-sample t-test, the only difference being that we enter (3) with the mean and standard deviation of paired differences instead of with the original data. Number of degrees of freedom is $n - 1$, where n is the number of the observed differences (number of pairs). This test is based on the assumption that the population of paired differences follows normal distribution.

Robustness of *t*-Test

Since the t-test requires certain assumptions in order to be exact, it is of interest to know how strongly the underlying assumptions can be violated without degrading the test results considerably. In general, a test is said to be robust if it is relatively insensitive to violation of its underlying assumptions. That is, a robust test is one in which the actual value of significance is unaffected by failure to meet assumptions (i.e., it is near the nominal level of significance), and at the same time the test maintains high power.

The one-sample t-test is widely considered reasonably robust against the violation of the normality assumption for large sample sizes, except for extremely skewed populations (see Bartlett 1935 or Bradley 1980). Departure from normality is most severe when sample sizes are small and becomes less serious as sample sizes increase (since the sampling distribution of the mean approaches a normal distribution; see ►Central Limit Theorems). However, for extremely skewed distribution even for quite large samples (e.g., 500), t-test may not be robust (Pocock 1982).

Numerous studies have dealt with the adequacy of the two-sample t-test if at least one assumption is violated. In case of unequal variances, it has been shown that the t-test is only robust if sample sizes are equal (e.g., Scheffé 1970; Posten et al. 1982; Zimmerman 2004). However, if two equal sample sizes are very small, the t-test may not be robust (see Huck 2008, pp. 205–207). If both sample size and variances are unequal, the Welch t-test is preferred to as a better procedure.

If the normality assumption is not met, a researcher can select one of the nonparametric alternatives of the t-test – in one-sample scenario ►Wilcoxon–signed–rank test, in two independent samples case ►Wilcoxon–Mann–Whitney test, and if the samples are dependent Wilcoxon-matched pair rank test (for the asymptotic efficiency comparison, see ►Asymptotic Relative Efficiency in Testing).

Extension to comparison of an arbitrary number of independent samples ends up in a technique called ►analysis of variance, abbreviated ANOVA. Multivariate counterparts of one-sample and two-sample t-tests are based on Hotelling's T^2 statistic (Johnson and Wichern 2007), and ANOVA generalizes to multivariate analysis of variance, abbreviated MANOVA (see ►Multivariate Analysis of Variance (MANOVA)).

Cross References

►Behrens–Fisher Problem
►Chernoff-Savage Theorem
►Density Ratio Model
►Effect Size
►Hotelling's T^2 Statistic
►Parametric Versus Nonparametric Tests
►Presentation of Statistical Testimony
►Rank Transformations
►Research Designs
►Robust Inference
►Scales of Measurement and Choice of Statistical Methods
►Significance Testing: An Overview
►Statistical Analysis of Drug Release Data Within the Pharmaceutical Sciences
►Student's t-Distribution
►Validity of Scales
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Aldrich J (2008) Comment on S. L. Zabell's paper: on Student's 1908 paper. The probable error of a mean. J Am Stat Assoc 103(481): 8–11

Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley, Hoboken

Bartlett MS (1935) The effect of non-normality on the t distribution. Proc Cambridge Philos Soc 31:223–231

Bhattacharyya GK, Johnson RA (1977) Statistical concepts and methods. Wiley, New York

Box JF (1987) Guinness, Gosset, Fisher, and small samples. Stat Sci 2(1):45–52

Bradley JV (1980) Nonrobustness in Z; t; and F tests at large sample sizes. Bull Psychonom Soc 16(5):333–336

Fay MP, Proschan MA (2010) Wilcoxon-Mann-Whitney or *t*-Test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv 4:1–39
Fisher RA (1939) Student. Ann Eugenic 9:1–9
Hogg RV, McKean JW, Craig AT (2005) Introduction to mathematical statistics, 6th edn. Prentice-Hall, Pearson
Huck SW (2008) Statistical misconceptions. Routledge Academic, New York
Johnson RA (2005) Miller and Freund's probability and statistics for engineers, 7th edn. Prentice-Hall, Pearson
Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis, 6th edn. Prentice-Hall, Pearson
Miller I, Miller M (1999) John E. Freund's mathematical statistics, 6th edn. Prentice-Hall, Pearson
Pearson ES (1939) Student as a statistician. Biometrika 30:210–250
Pocock SJ (1982) When not to rely on the central limit theorem - an example from absentee data. Commun Stat Part A - Theory Meth 11(19):2169–2179
Posten HO, Yeh HC, Owen DB (1982) Robustness of the two-sample t-test under violations of the homogeneity of variance assumptions. Commun Stat - Theory Meth 11:109–126
Scheffé H (1970) Practical solutions of the Behrens-Fisher problem. J Am Stat Assoc 65(332):1501–1508
Student (1908) The probable error of a mean. Biometrika 6:1–25
Winer BJ (1971) Statistical principles in experimental design. McGraw-Hill, New York
Zabel SL (2008) On Student's 1908 article. The probable error of a mean. J Am Stat Assoc 103(481):1–7
Zimmerman DW (2004) Inflation of type I error rates by unequal variances associated with parametric, nonparametric, and rank transformation tests. Psicológica 25:103–133

Sturges' and Scott's Rules

David W. Scott
Noah Harding Professor, Associate Chairman
Rice University, Houston, TX, USA

Introduction

The fundamental object of modern statistics is the random variable X and its associated probability law. The probability law may be given by the cumulative probability distribution $F(x)$, or equivalently by the probability density function $f(x) = F'(x)$, assuming the continuous case. In practice, estimation of the probability density may approached either parametrically or nonparametrically. If a parametric model $f(x|\theta)$ is assumed, then the unknown parameter θ may be estimated from a random sample using maximum likelihood methods, for example. If no parametric model is available, then a nonparametric estimator such as the histogram may be chosen. This article describes two different methods of specifying the construction of a histogram from a random sample.

Histogram as Density Estimator

The histogram is a convenient graphical object for representing the shape of an unknown density function. We begin by reviewing the stem-and-leaf diagram, introduced by Tukey (1977). Tukey reanalyzed Lord Rayleigh's 15 measurements of the weight of nitrogen. Using the ▶R language, the stem-and-leaf diagram of the weights is given in Fig. 1. One of the 15 raw numbers is $x_1 = 2.30143$. Where does x_1 appear in the diagram? The three digits to the left of "|" are called the *stem*. The stems correspond to the *bins* of a histogram. Here there are four stems, defined by the five cut points $(2.295, 2.300, 2.305, 2.310, 2.315)$. The bin counts are $(6, 2, 0, 7)$, with x_1 falling in the second bin. Rounding x_1 to 2.301 and removing the stem "230," leaves the leaf value of "1," which is what appears to the right of the second stem in Figure 1. In the fourth stem, all seven measurements rounded to 2.310. This sample was measured to high accuracy to estimate the atomic weight of nitrogen, but instead its highly non-normal shape led to the discovery of the noble gas argon.

The ordinary histogram depicts only the bin counts, which we denote by $\{\nu_k\}$, where the integer k indicates the bin number. Then $\sum_k \nu_k = n$, where n denotes the sample size. Given an ordered set of cut points $\{t_k\}$, the *kth* bin B_k is the half-open interval $[t_k, t_{k+1})$. If all of the bins have the same width, then plotting the bin counts gives an indication of the shape of the underlying density; see the left frame of Figure 2 for an example.

The left frame of Fig. 2 depicts a frequency histogram, since the bin counts $\{\nu_k\}$ are plotted. The density histogram is defined by the formula

$$\hat{f}(x) = \frac{\nu_k}{nh} \quad x \in B_k. \tag{1}$$

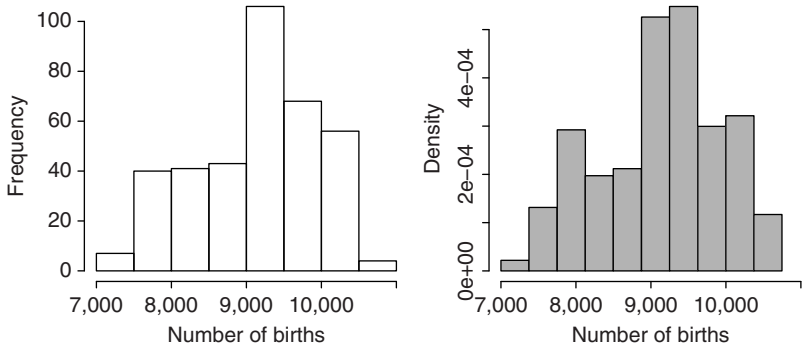
The density histogram estimator is nonnegative and integrates to 1. The right frame of Fig. 2 shows a density histogram with a narrower bin width.

```
> stem (wts)

  The decimal point is 2 digit (s) to the left of the |

229 | 889999
230 | 12
230 |
231 | 0000000
```

Sturges' and Scott's Rules. Fig. 1 Tukey's stem-and-leaf plot of the Raleigh data ($n = 15$)



Sturges' and Scott's Rules. Fig. 2 Histograms of the number of births daily in the USA in 1978. The bin widths are 500 and 375, respectively

Sturges' Rule

The origins of the histogram may be traced back to 1662 and the invention of actuarial tables by John Graunt (1662). But the first practical rule for the construction of histograms took another 260 years. Sturges (1926) essentially developed a normal reference rule, that is, a formula for the number of bins appropriate for normal data. Sturges sought a discrete distribution that was approximately normal to develop his formula. While several come to mind, clearly a binomial random variable $Y \sim B(m,p)$ with $p = \frac{1}{2}$ is suitable. If we imagine appropriately re-scaled normal data, which are continuous, rounded to integer values $(0,1,\ldots,m)$ in the $m+1$ bins (each of width $h=1$)

$$B_0 = \left(-\frac{1}{2},\frac{1}{2}\right] \quad B_1 = \left(\frac{1}{2},\frac{3}{2}\right] \quad \ldots \quad B_m = \left(m-\frac{1}{2},m+\frac{1}{2}\right], \tag{2}$$

then the Binomial probability in the *kth* bin is given by

$$P(Y=k) = \binom{m}{k} p^k (1-p)^{m-k} = \binom{m}{k}\left(\frac{1}{2}\right)^m = \frac{\binom{m}{k}}{2^m \cdot 1}. \tag{3}$$

Comparing the density formulae in Eqs. 1 and 3, we have

$$v_k = \binom{m}{k}, \quad n = 2^m, \quad \text{and} \quad h = 1. \tag{4}$$

If we let K denote the number of bins, then $K = m+1$ for the binomial density, as well as for the appropriately re-scaled normal data. From Eq. 4, we compute

$$n = 2^m = 2^{K-1}; \quad \text{hence} \quad K = 1 + \log_2(n). \tag{5}$$

The formula for K in Eq. 5 is called *Sturges' Rule*.

Scott's Rule

The density histogram $\hat{f}(x) = v_k/nh$ is not difficult to analyze for a random sample of size n from a density $f(x)$. Given a set of equal-width bins, the bin counts $\{v_k\}$ are individually a Binomial random variable $B(n,p_k)$, with probability

$$p_k = \int_{B_k} f(t)\,dt = \int_{t_k}^{t_{k+1}} f(t)\,dt = \int_{t_k}^{t_k+h} f(t)\,dt.$$

So $Ev_k = np_k$. Thus for a fixed point x, the expected value of the density histogram $\hat{f}(x)$ is $(np_k)/nh = p_k/h$. Scott (1979) shows that this is close to the unknown true value $f(x)$ when the bin width h is small.

On the other hand, the variance of v_k is $np_k(1-p_k)$, so that the variance of $\hat{f}(x)$ is $np_k(1-p_k)/(nh)^2 \sim p_k/nh^2$. This variance will be small if h is large. Since h cannot be both small and large, and using the integrated mean squared error as the criterion, Scott (1979) derived the asymptotically optimal bin width to be

$$h_S^* = \left(\frac{6}{n\int f'(t)^2\,dt}\right)^{1/3}. \tag{6}$$

While the formula for h^* in Eq. 6 seems to require knowledge of the unknown density, it is perfectly suitable for deriving Scott's normal-reference bin-width rule. If $f \sim N(\mu,\sigma^2)$, then

$$\int_{-\infty}^{\infty} f'(t)^2\,dt = \frac{1}{4\sqrt{\pi}\sigma^3} \quad \text{and}$$
$$h_S^* = \left(\frac{24\sqrt{\pi}\sigma^3}{n}\right)^{1/3} \approx 3.5\sigma n^{-1/3}. \tag{7}$$

Scott's rule $\hat{h}_S$ replaces σ in the formula for h_S^* by the usual maximum likelihood estimate of the standard deviation.

The Rules in Practice

For the birth count data used in Fig. 2, $n = 365$, $\hat{\sigma} = 817.9$, and the sample range is $(7135, 10711)$; hence, Sturges' and

Scott's rules give

$$K = 9.51\left(\text{or } \hat{h} = \frac{10711 - 7135}{9.51} = 376.0\right) \text{ and } \quad \hat{h}_S = 400.6.$$

Note the density histogram in the right frame of Fig. 2 uses $h = 375$, which has ten bins. Interestingly, the left frame shows the default histogram in R, which implements Sturges' rule as well. However, instead of finding ten bins exactly, R uses the function *pretty* to pick approximately ten bins with "convenient" values for $\{t_k\}$. The result in this case is 8 bins, and $h = 500$. Scott's rule (not shown) is close to $h = 375$.

The Rules with Massive Datasets

While the two rules often give similar results for sample sizes less than a couple hundred, they diverge for larger values of n for any density, including the normal. To see this, let us reconsider the binomial/normal construction at Eq. 2 we used to find Sturges' rule. (The data are basically rounded to one of the $m+1$ integer values $0, 1, \ldots, m$.) Thus we have $K = m+1$ bins, $n = 2^{K-1}$, $\mu = mp = m/2$, and $\sigma^2 = mp(1-p) = m/4$. Note that the variance of this density increases with the sample size in such a way that Sturges' rule always gives $h = 1$ for any sample size.

By way of contrast, Scott's rule from Eq. 7 is given by

$$h_S^* = 3.5\sqrt{\frac{m}{4}}\, n^{-1/3} = 1.75\sqrt{K-1}\, n^{-1/3}$$
$$= 1.75\sqrt{\log_2(n)}\, n^{-1/3}. \qquad (8)$$

Observe that $h_S^* \to 0$ as the sample size $n \to \infty$. In fact, $h_S^* < 1$ for all $n > 87$ for these data. When $n = 200$, $h_S^* = 0.83$, only 17% less than Sturges' $h = 1$. However, when $n = 10^6$, $h_S^* = 0.0781$. Thus the optimal histogram would have nearly 13 (1/0.0781) times as many bins as when using Sturges' rule.

The bin width given in Eq. 8 is also the *ratio* of Scott's rule to the Sturges bin width (since $h = 1$). If the normal data have any other scale, then the ratio is the same. The trick of using the Binomial model facilitates the conversion of bin counts to bin widths. Otherwise, a more careful analysis of the sample range of normal data would be necessary.

Discussion

Both Sturges' and Scott's rules use the normal-reference principle. However, Sturges makes a deterministic calculation, whereas Scott's rule is based upon a balancing of the global variance and squared bias of the histogram estimator. For normal data, we have seen that Sturges' rule greatly understates the optimal number of bins (according to integrated mean squared error). Thus we say that Sturges' rule tends to oversmooth the resulting histogram. Sturges' rule wastes a large fraction of the information available in large samples.

Why are these rules useful in practice? Terrell and Scott; 1985 show that there exists an "easiest" smooth density, whose optimal bin width is only 1.069 times as wide as Scott's normal reference rule. Terrell concludes that for any other density, the (unknown) optimal bin width will be narrower still. Thus, the normal reference rule is always useful as a first look at the data. Narrower bin widths can be investigated if the sample size is large enough and there is obvious non-normal structure.

Hyndman; 1995 cautions that since both v_k and n in Eqs. 3 and 4 could be multiplied by a constant factor, that K could take the general form $c + \log_2(n)$. The fact that Sturges' rule ($c = 1$) continues to be used is probably due to its simple form and its closeness to the optimal number of bins for textbook-sized problems ($n < 200$). Of course, if you impose the boundary condition that with one sample ($n = 1$) you should choose one bin ($K = 1$), then you would conclude that $c = 1$ is appropriate.

A variation of Scott's rule was independently proposed by Freedman and Diaconis; 1981, who suggested using a multiple of the interquartile range rather than $\hat{\sigma}$ in the normal reference rule. Of course, there are more advanced methods of cross-validation for histograms introduced by Rudemo; 1982. Surveys of these and other ideas may be found in Scott; 1992, Wand; 1997, and Doane 1976. Finally, we note that if the bin widths are not of equal width, then the shape of the frequency histogram can be grossly misleading. The appropriate density histogram has the form v_k/nh_k, but more research is required to successfully construct these generalized histograms in practice.

Acknowledgments

This work was partially supported by NSF award DMS-09-07491, and ONR contract N00014-06-1-0060.

About the Author

Professor Scott was awarded the Founders Award, American Statistical Association (2008), for "superb contributions and leadership in statistical research, particularly in multivariate density estimation and visualization, and in editorship of the Journal of Computational and Graphical Statistics." He has also received the U.S. Army Wilks Award (2004), and was named the Texas Statistician of the Year (1993). He was Editor, *Journal of Computational and Graphical Statistics* (2000–2004).

Cross References

►Exploratory Data Analysis
►Nonparametric Density Estimation
►Nonparametric Estimation
►Stem-and-Leaf Plot

References and Further Reading

Doane DP (1976) Aesthetic frequency classifications. Am Stat 30:181–183
Freedman D, Diaconis P (1981) On the histogram as a density estimator: 12 theory. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 57:453–476
Graunt J (1662) Natural and political observations made upon the bills of mortality. Martyn, London
Hyndman RJ (1995) The problem with sturges rule for constructing histograms. Unpublished note, 1995
Rudemo M (1982) Empirical choice of histograms and kernel density estimators. Scand J Stat 9:65–78
Scott DW (1979) On optimal and data-based histograms. Biometrika 66:605–610
Scott DW (1992) Multivariate density estimation: theory, practice, and visualization. Wiley, New York
Sturges HA The choice of a class interval. J Am Stat Assoc 21:65–66
Terrell GR, Scott DW (1985) Oversmoothed nonparametric density estimates. J Am Stat Assoc 80:209–214
Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading, MA
Wand MP (1997) Data-based choice of histogram bin width. Am Stat 51:59–64

Sufficient Statistical Information

Nitis Mukhopadhyay
Professor
University of Connecticut-Storrs, Storrs, CT, USA

Introduction

In the entry ►Sufficient statistics, it was mentioned that we wished to work with a sufficient or minimal sufficient statistic **T** because such a statistic will summarize data, but preserve all "information" about an unknown parameter $\boldsymbol{\theta}$ contained in the original data. Here, $\boldsymbol{\theta}$ may be real or vector valued. But, how much *(Fisher-)information* do we have in the original data which we attempt to preserve through data summary? Our present concern is to quantify Fisher-information content within some data.

The notion of the information about $\boldsymbol{\theta}$ contained in data was introduced by F. Y. Edgeworth in a series of papers, published in the *J. Roy. Statist. Soc.*, during 1908–1909. Fisher (1922) articulated the systematic development of this concept. The reader is referred to Efron's (1998, p. 101) commentaries on Fisher-information.

Section "►One Parameter Case" introduces a one-parameter situation. Section "►Multi-Parameter Case" discusses the two-parameter case which easily extends to a multi-parameter situation. When one is forced to utilize some less than full information data summary, we discuss in section "►Role in the Recovery of Full Information" how the lost information may be recovered by conditioning on ancillary statistics. Mukhopadhyay (2000, Chap. 6) includes in-depth discussions.

One-Parameter Case

Suppose that X is an observable real valued random variable with the pmf or pdf $f(x;\theta)$ where the unknown parameter $\theta \in \Theta$, an open subinterval of $\Re$, while the $\mathcal{X}$ space is *assumed* not to depend upon θ. We *assume* throughout that the partial derivative $\frac{\partial}{\partial\theta} f(x;\theta)$ is finite for all $x \in \mathcal{X}$, $\theta \in \Theta$. We also *assume* that we can interchange the derivative (with respect to θ) and the integral (with respect to x).

Definition 1 *The Fisher-information or simply the information about θ, contained in the data, is given by*

$$\mathcal{I}_X(\theta) = E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(X;\theta)\right\}^2\right].$$

The information $\mathcal{I}_X(\theta)$ measures the square of the sensitivity of $f(x;\theta)$ on an average due to an infinitesimal subtle change in the true parameter value θ. This concept may be understood as follows: Consider

$$\lim_{\Delta\theta\to 0} \frac{f(x;\theta+\Delta\theta) - f(x;\theta)}{\Delta\theta} \div f(x;\theta)$$

which is $\frac{\partial}{\partial\theta}\log f(x;\theta)$. Obviously, $E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] \equiv 0$, and hence one goes on to define $\mathcal{I}_X(\theta) = E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(X;\theta)\right\}^2\right]$.

Example 1 Let X be Poisson(λ), $\lambda > 0$. One verifies that $\mathcal{I}_X(\lambda) = \lambda^{-1}$. That is, as we contemplate having larger and larger values of λ, the variability built in X increases, and hence it seems natural that the information about the unknown parameter λ contained in the data X will go down further and further. ▲

Example 2 Let X be $N(\mu,\sigma^2)$ where $\mu \in (-\infty,\infty)$ is an unknown parameter. Here, $\sigma \in (0,\infty)$ is assumed known. One verifies that $\mathcal{I}_X(\mu) = \sigma^{-2}$. Again, as we contemplate having larger and larger values of σ, the variability built in

X increases, and hence it seems natural that the information about the unknown parameter μ contained in the data X will go down further and further. ▲

The following result quantifies the information about an unknown parameter θ contained in a random sample $X_1, \ldots, X_n$ of size n.

Theorem 1 *Let $X_1, \ldots, X_n$ be iid with a common pmf or pdf given by $f(x;\theta)$. We denote $E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f(X_1;\theta)\right\}^2\right] = \mathcal{I}_{X_1}(\theta)$, the information contained in the observation X_1. Then, the information $\mathcal{I}_X(\theta)$, contained in the random sample $X = (X_1, \ldots, X_n)$, is given by*

$$\mathcal{I}_X(\theta) = n\mathcal{I}_{X_1}(\theta) \text{ for all } \theta \in \Theta.$$

Next, suppose that we have collected random samples $X_1, \ldots, X_n$ from a population and we have somehow evaluated the information $\mathcal{I}_X(\theta)$ contained in $\mathbf{X} = (X_1, \ldots, X_n)$. Also, suppose that we have a summary *statistic* $T = T(\mathbf{X})$ in mind for which we have evaluated the information $\mathcal{I}_T(\theta)$ contained in T. If it turns out that $\mathcal{I}_T(\theta) = \mathcal{I}_X(\theta)$, can we then claim that the statistic T is indeed sufficient for θ? The answer is yes, we certainly can.

We state the following result by referring to Rao (1973, result (iii), p. 330) for details. In an exchange of personal communications, C.R. Rao had provided a simple way to look at the next Theorem 2. In Mukhopadhyay (2000), the Exercise 6.4.15 gives an outline of Rao's elegant proof whereas in the Examples 6.4.3–6.4.4 of Mukhopadhyay (2000), one finds opportunities to apply this theorem.

Theorem 2 *Suppose that $\mathbf{X}$ is the whole data and $T = T(\mathbf{X})$ is some statistic. Then, $\mathcal{I}_{\mathbf{X}}(\theta) \geq \mathcal{I}_T(\theta)$ for all $\theta \in \Theta$. The two information measures will be equal for all θ if and only if T is a sufficient statistic for θ.*

Multi-Parameter Case

When the unknown parameter $\boldsymbol{\theta}$ is multidimensional, the definition of the Fisher information measure gets more involved. To keep the presentation simple, we only discuss the case of a two-dimensional parameter.

Suppose that X is an observable real valued random variable with the pmf or pdf $f(x;\boldsymbol{\theta})$ where the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta$, an open rectangle $\subseteq \Re^2$, and the $\mathcal{X}$ space does not depend upon $\boldsymbol{\theta}$. We *assume* throughout that $\frac{\partial}{\partial\theta_i}f(x;\boldsymbol{\theta})$ exists, $i = 1, 2$, for all $x \in \mathcal{X}$, $\boldsymbol{\theta} \in \Theta$, and that we can also interchange the partial derivative (with respect to θ_1, θ_2) and the integral (with respect to x).

Definition 2 *Denote $I_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\left\{\frac{\partial}{\partial\theta_i}\log f(X;\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial\theta_j}\log f(X;\boldsymbol{\theta})\right\}\right]$, for $i, j = 1, 2$. The Fisher-information matrix or simply the information matrix about $\boldsymbol{\theta}$ is given by*

$$\mathcal{I}_X(\boldsymbol{\theta}) = \begin{pmatrix} I_{11}(\boldsymbol{\theta}) & I_{12}(\boldsymbol{\theta}) \\ I_{21}(\boldsymbol{\theta}) & I_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

In situations where $\frac{\partial^2}{\partial\theta_i\partial\theta_j}f(x;\boldsymbol{\theta})$ exists for all $x \in \mathcal{X}$, for all $i, j = 1, 2$, and for all $\boldsymbol{\theta} \in \Theta$, we can alternatively express

$$I_{ij}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\theta_j}\log f(X;\boldsymbol{\theta})\right] \text{ for } i, j = 1, 2,$$

and rewrite $\mathcal{I}_X(\boldsymbol{\theta})$ accordingly.

Having a statistic $T = T(X_1, \ldots, X_n)$, however, the associated information matrix about $\boldsymbol{\theta}$ will simply be calculated as $\mathcal{I}_T(\boldsymbol{\theta})$ where one would replace the original pmf or pdf $f(x;\boldsymbol{\theta})$ by that of T, namely $g(t;\boldsymbol{\theta})$, $t \in \mathcal{T}$. In order to compare two summary statistics T_1 and T_2, we have to consider their individual two-dimensional *information matrices* $\mathcal{I}_{T_1}(\boldsymbol{\theta})$ and $\mathcal{I}_{T_2}(\boldsymbol{\theta})$. It would be tempting to say that T_1 *is more informative* about $\boldsymbol{\theta}$ than T_2 provided that

the matrix $\mathcal{I}_{T_1}(\theta) - \mathcal{I}_{T_2}(\theta)$ is positive semi definite.

A version of Theorem 1. holds in the multiparameter case. One may refer to Rao (1973, Sect. 5a.3).

Example 3 Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ where $\mu \in (-\infty, \infty)$ and $\sigma^2 \in (0, \infty)$ are both unknown parameters. Denote $\boldsymbol{\theta} = (\mu, \sigma^2), \mathbf{X} = (X_1, \ldots, X_n)$. One can verify that the information matrix is given by

$$\mathcal{I}_X(\boldsymbol{\theta}) = n\mathcal{I}_{X_1}(\boldsymbol{\theta}) = \begin{pmatrix} n\sigma^{-2} & 0 \\ 0 & \frac{1}{2}n\sigma^{-4} \end{pmatrix},$$

for the whole data $\mathbf{X}$. ▲

Example 4 (Example 3. Continued) Let $\overline{X} = n^{-1}\Sigma_{i=1}^n X_i$, the sample mean and $S^2 = (n-1)^{-1}\Sigma_{i=1}^n (X_i - \overline{X})^2$, the sample variance, $n \geq 2$. One can check that

$$\mathcal{I}_{\overline{X}}(\boldsymbol{\theta}) = \begin{pmatrix} n\sigma^{-2} & 0 \\ 0 & \frac{1}{2}\sigma^{-4} \end{pmatrix},$$

$$\mathcal{I}_{S^2}(\boldsymbol{\theta}) = \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2}(n-1)\sigma^{-4} \end{pmatrix}.$$

Surely, $\overline{X}$ and S^2 are independent, and hence

$$\mathcal{I}_{\overline{X},S^2}(\boldsymbol{\theta}) = \mathcal{I}_{\overline{X}}(\boldsymbol{\theta}) + \mathcal{I}_{S^2}(\boldsymbol{\theta}) = \begin{pmatrix} n\sigma^{-2} & 0 \\ 0 & \frac{1}{2}n\sigma^{-4} \end{pmatrix},$$

which coincides with $\mathcal{I}_X(\boldsymbol{\theta})$. ▲

Role in the Recovery of Full Information

In the entry ►Sufficient statistics, we had seen how ancillary statistics could play significant roles in conjunction with non-sufficient statistics. Suppose that T_1 is a non-sufficient statistic for θ and T_2 is ancillary for θ. In other words, in terms of the information content, $\mathcal{I}_{T_1}(\theta) < \mathcal{I}_X(\theta)$ where $\mathbf{X}$ is the whole data and $\mathcal{I}_{T_2}(\theta) = 0$ for all $\theta \in \Theta$. Can we recover all the information contained in $\mathbf{X}$ by reporting T_1 while conditioning on the observed value of T_2? The answer is: we can do so and it is a fairly simple process.

Such a process of conditioning has far reaching implications as emphasized by Fisher (1934, 1956) in his famous "Nile" example. One may also refer to Basu (1964), Hinkley (1980), Ghosh (1988) and Reid (1995) for fuller discussions of *conditional inference*. Also, refer to Mukhopadhyay (2000, Sect. 6.5).

The approach goes through the following steps. One first finds the conditional pdf of T_1 when $T_1 = u$ given that $T_2 = v$, denoted by $g_{T_1|\,v}(u;\theta)$. Using this conditional pdf, one can obtain the information content:

$$\mathcal{I}_{T_1|\,v}(\theta) = E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log\{g_{T_1|\,v}(T_1;\theta)\}\right\}^2\right].$$

In general, the expression of $\mathcal{I}_{T_1|\,v}(\theta)$ would depend on v, that is, the fixed value of T_2. Next, one averages $\mathcal{I}_{T_1|\,v}(\theta)$ over all possible values v, that is, evaluates $E_{T_2}[\mathcal{I}_{T_1|T_2}(\theta)]$. Once this last bit of averaging is done, it will coincide with the information content in the joint statistic (T_1, T_2), that is, one can claim:

$$\mathcal{I}_{T_1,T_2}(\theta) = E_{T_2}\left[\mathcal{I}_{T_1|T_2}(\theta)\right].$$

This analysis provides a way to recover the lost information due to reporting T_1 alone via conditioning on an ancillary statistic T_2. Two examples follow that are taken from Mukhopadhyay (2000, pp. 316–318).

Example 5 Let X_1, X_2 be iid $N(\theta, 1)$ where $\theta \in (-\infty, \infty)$ is an unknown parameter. We know that $\overline{X}$ is sufficient for θ. Now, $\overline{X}$ is distributed as $N\left(\theta, \frac{1}{2}\right)$ so that we can immediately write $\mathcal{I}_{\overline{X}}(\theta) = 2$. Now, $T_1 = X_1$ is not sufficient for θ since $\mathcal{I}_{X_1}(\theta) = 1 < \mathcal{I}_{\overline{X}}(\theta)$. That is, if we report only X_1 after the data (X_1, X_2) has been collected, there will be some loss of information. Next, consider an ancillary statistic, $T_2 = X_1 - X_2$ and now the joint distribution of (T_1, T_2) is $N_2\left(\theta, 0, 1, 2, \rho = \frac{1}{\sqrt{2}}\right)$. Hence, we find that the conditional distribution of T_1 given $T_2 = v$ is $N\left(\theta + \frac{1}{2}v, \frac{1}{2}\right), v \in (-\infty, \infty)$. Thus, we first have $\mathcal{I}_{T_1|v}(\theta) = E_{T_1|\,v}\left[4\left(T_1 - \theta - \frac{1}{2}v\right)^2\right] = 2$ and since this expression does not involve v, we then have $E_{T_2}\left[\mathcal{I}_{T_1|T_2}(\theta)\right] = 2$ which equals $\mathcal{I}_{\overline{X}}(\theta)$. In other words, by conditioning on the ancillary statistic T_2, we have recovered the full information which is $\mathcal{I}_{\overline{X}}(\theta)$. ▲

Example 6 Suppose that (X, Y) is distributed as $N_2(0, 0, 1, 1, \rho)$ where the unknown parameter is the correlation coefficient $\rho \in (-1, 1)$. Now consider the two individual statistics X and Y. Individually, both $T_1 = X$ and $T_2 = Y$ are ancillary for ρ. We note that the conditional distribution of X given $Y = y$ is $N(\rho y, 1-\rho^2)$ for $y \in (-\infty, \infty)$ and accordingly have,

$$\frac{\partial}{\partial\rho}\log f_{X|\,Y=y}(x;\rho) = \frac{\rho}{1-\rho^2} - \left[\frac{\rho(x-\rho y)^2}{(1-\rho^2)^2} - \frac{y(x-\rho y)}{(1-\rho^2)}\right].$$

In other words, the information about ρ contained in the conditional distribution of $T_1 \mid T_2 = v, v \in \Re$, is given by

$$\frac{2\rho^2}{(1-\rho^2)^2} + \frac{v^2}{(1-\rho^2)},$$

which depends on the value v unlike what we had in Example 5. Then, the information contained in (X, Y) will be given by

$$\mathcal{I}_{X,Y}(\rho) = E_{T_2}\left[E_{T_1|\,T_2=v}\left(\left\{\frac{\partial}{\partial\rho}\log f_{T_1|\,T_2=v}(T_1;\rho)\right\}^2\right)\right]$$

$$= E_{T_2}\left[\frac{2\rho^2}{(1-\rho^2)^2} + \frac{{T_2}^2}{(1-\rho^2)}\right]$$

$$= \frac{2\rho^2}{(1-\rho^2)^2} + \frac{1}{(1-\rho^2)} = \frac{1+\rho^2}{(1-\rho^2)^2}.$$

In other words, even though the statistic X tells us nothing about ρ, by averaging the conditional (on the statistic Y) information in X, we have recovered the full information about ρ contained in the whole data (X, Y). ▲

About the Author

For biography *see* the entry ►Sequential Sampling.

Cross References

►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Cramér–Rao Inequality
►Estimation
►Statistical Design of Experiments (DOE)
►Statistical Inference for Stochastic Processes
►Statistical View of Information Theory
►Sufficient Statistics

References and Further Reading

Basu D (1964) Recovery of ancillary information. Contributions to statistics, the 70th birthday festschrift volume presented to P. C. Mahalanobis. Pergamon, Oxford

Efron BF (1998) R. A. Fisher in the 21st century (with discussions by Cox DR, Kass R, Barndorff-Nielsen O, Hinkley DV, Fraser DAS, Dempster AP) Stat Sci 13:95–122

Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc A222:309–368

Fisher RA (1934) Two new properties of mathematical likelihood. Proc R Soc A 144:285–307

Fisher RA (1956) Statistical methods and scientific inference. Oliver and Boyd, Edinburgh/London

Ghosh JK (ed) (1988) Statistical information and likelihood: a collection of critical essays by Dr. D. Basu. Lecture notes in statistics No 45. Springer, New York

Hinkley DV (1980) Fisher's development of conditional inference. In: Fienberg SE, Hinkley DV (eds) R. A. Fisher: an appreciation. Springer, New York, pp 101–108

Mukhopadhyay N (2000) Probability and statistical inference. Marcel Dekker, New York

Rao CR (1973) Linear statistical inference and its applications, 2 edn. Wiley, New York

Reid N (1995) The roles of conditioning in inference (with discussions by Casella G, Dawid AP, DiCiccio TJ, Godambe VP, Goutis C, Li B, Lindsay BC, McCullagh P, Ryan LA, Severini TA, Wells MT) Stat Sci 10:138–157

Sufficient Statistics

Nitis Mukhopadhyay
Professor
University of Connecticut-Storrs, Storrs, CT, USA

Introduction

Many fundamental concepts and principles of statistical inference originated in Fisher's work. Perhaps the deepest of all statistical concepts and principles is *sufficiency*. It originated from Fisher (1920) and blossomed further in Fisher (1922). We introduce the notion of sufficiency which helps in summarizing data without any loss of *information*.

Section "▶Sufficiency" introduces sufficiency and *Neyman factorization*. Section "▶Minimal Sufficiency" discusses *minimal sufficiency*, the *Lehmann-Scheffé approach*, and *completeness*. Section "▶Neyman Factorization" shows the importance of *ancillary* statistics including Basu's theorem. Mukhopadhyay (2000, Chap. 6) provides many more details.

Sufficiency

Let $X_1, \ldots, X_n$ be independent real-valued observations having a common probability mass function (pmf) or probability density function (pdf) $f(x;\theta), x \in \mathcal{X}$, the domain space for x. Here, n is known, but $\theta \in \Theta(\subseteq \Re)$ is unknown. In general, however, the X's and θ are allowed to be vector valued. This should be clear from the context. A summary from data $\mathbf{X} \equiv (X_1, \ldots, X_n)$ is provided by some appropriate statistic, $T \equiv T(\mathbf{X})$ which may be vector valued.

Definition 1 *A real valued statistic T is called sufficient for parameter θ if and only if the conditional distribution of the random sample $\mathbf{X} = (X_1, \ldots, X_n)$ given $T = t$ does not involve θ, for all $t \in \mathcal{T}$, the domain space for T.*

In other words, *given* the value t of a *sufficient statistic T*, *conditionally* there is no more information left in the original data regarding θ. That is, once a sufficient summary T becomes available, the original data $\mathbf{X}$ becomes redundant.

Definition 2 *A statistic $\mathbf{T} \equiv (T_1, \ldots, T_k)$ where $T_i \equiv T_i(X_1, \ldots, X_n), i = 1, \ldots, k$, is called jointly sufficient for parameter θ if and only if the conditional distribution of $\mathbf{X} = (X_1, \ldots, X_n)$ given $\mathbf{T} = \mathbf{t}$ does not involve θ, for all $\mathbf{t} \in \mathcal{T} \subseteq \Re^k$.*

Example 1 Suppose that $X_1, \ldots, X_n$ are independent and identically distributed (iid) Poisson(λ) where λ is unknown, $0 < \lambda < \infty$. Here, $\mathcal{X} = \{0, 1, 2, \ldots\}$, $\theta = \lambda$, and $\Theta = (0, \infty)$. Then, $T = \Sigma_{i=1}^{n} X_i$ is a sufficient statistic for λ.

Neyman Factorization

Suppose that we have observable real valued iid observations $X_1, \ldots, X_n$ from a population with a common pmf or pdf $f(x;\boldsymbol{\theta})$. Then, the likelihood function is given by $L(\boldsymbol{\theta}) = \Pi_{i=1}^{n} f(x_i;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. Fisher (1922) discovered the fundamental idea of factorization whereas Neyman (1935) rediscovered a refined approach to factorize a likelihood function. Halmos and Savage (1949) and Bahadur (1954) introduced measure-theoretic treatments.

Theorem 1 (Neyman Factorization Theorem). *A vector valued statistic $\mathbf{T} = \mathbf{T}(X_1, \ldots, X_n)$ is jointly sufficient for $\boldsymbol{\theta}$ if and only if the following factorization holds:*

$$L(\boldsymbol{\theta}) = g(\mathbf{T}(x_1, \ldots, x_n);\boldsymbol{\theta})\, h(x_1, \ldots, x_n),$$
$$\text{for all } x_1, \ldots, x_n \in \mathcal{X},$$

where the functions $g(\mathbf{T};\boldsymbol{\theta})$ and $h(.)$ are both nonnegative, $h(x_1, \ldots, x_n)$ is free from $\boldsymbol{\theta}$, and $g(\mathbf{T};\boldsymbol{\theta})$ depends on $x_1, \ldots, x_n$ only through the observed value $\mathbf{T}(x_1, \ldots, x_n)$ of $\mathbf{T}$.

Example 2 Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ where $\boldsymbol{\theta} = (\mu, \sigma^2) \in \Re \times \Re^+$ is an unknown parameter vector. Let

$\overline{X}, S^2$ respectively be the sample mean and variance. Then, $\mathbf{T} = (\overline{X}, S^2)$ is jointly sufficient for $\boldsymbol{\theta}$. However, this does not imply component-wise sufficiency. To appreciate this fine line, pretend for a moment that one could claim component-wise sufficiency. But, since $(\overline{X}, S^2)$, and hence $(S^2, \overline{X})$, is jointly sufficient for (μ, σ^2). Now, how many would be willing to push an idea that component-wise, S^2 is sufficient for μ or $\overline{X}$ is sufficient for σ^2!

Theorem 2 (Sufficiency in an Exponential Family). *Suppose that $X_1, \ldots, X_n$ are iid with a common pmf or the pdf belonging to a regular k-parameter exponential family, namely*

$$f(x;\boldsymbol{\theta}) = a(\boldsymbol{\theta})g(x)exp\left\{\Sigma_{i=1}^{k} b_i(\boldsymbol{\theta})R_i(x)\right\}$$

with appropriate forms for $g(x) \geq 0$, $a(\boldsymbol{\theta}) \geq 0, b_i(\boldsymbol{\theta})$ and $R_i(x)$, $i = 1, \ldots, k$. Denote $T_j = \Sigma_{i=1}^{n} R_j(X_i), j = 1, \ldots, k$. Then, the statistic $\mathbf{T} = (T_1, \ldots, T_k)$ is jointly sufficient for $\boldsymbol{\theta}$.

Minimal Sufficiency

From the factorization Theorems 1–2, it should be clear that the whole data $\mathbf{X}$ must always be sufficient for the unknown parameter $\boldsymbol{\theta}$. But, we ought to reduce the data by means of summary statistics in lieu of considering $\mathbf{X}$ itself. What is a natural way to define the notion of a "shortest sufficient" or "best sufficient" summary statistic? The other concern should be to get hold of such a summary, if there is one.

Lehmann and Scheffé (1950) gave a mathematical formulation of the concept known as *minimal sufficiency* and proposed a technique to locate minimal sufficient statistics. Lehmann and Scheffé (1955, 1956) included crucial follow-ups.

Definition 3 *A statistic $\mathbf{T}$ is called minimal sufficient for the unknown parameter $\boldsymbol{\theta}$ if and only if*

1. *$\mathbf{T}$ is sufficient for $\boldsymbol{\theta}$, and*
2. *$\mathbf{T}$ is minimal or "shortest" in the sense that $\mathbf{T}$ is a function of any other sufficient statistic.*

Lehmann–Scheffé Approach

The following result was proved in Lehmann and Scheffé (1950). Its proof requires some understanding of the correspondence between a statistic and so called *partitions* it induces on a sample space.

Theorem 3 (Minimal Sufficient Statistics). *Let us denote $h(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \Pi_{i=1}^{n} f(x_i;\boldsymbol{\theta}) / \Pi_{i=1}^{n} f(y_i;\boldsymbol{\theta})$, the ratio of the likelihood functions at $\mathbf{x}$ and $\mathbf{y}$, for $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$. Let $\mathbf{T} \equiv \mathbf{T}(X_1, \ldots, X_n) = (T_1, \ldots, T_k)$ be a statistic such that the following holds:*

- *With any two arbitrary but fixed data points $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$ from $\mathcal{X}^n$, $h(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ does not involve $\boldsymbol{\theta}$ if and only if $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$.*

Then, $\mathbf{T}$ is minimal sufficient for $\boldsymbol{\theta}$.

In Examples 1–2 and Theorem 2, the reported sufficient statistics also happen to be the minimal sufficient statistics. It should be noted, however, that a minimal sufficient statistic may exist for some distributions from outside a regular exponential family. For example, let $X_1, \ldots, X_n$ be iid Uniform$(0, \theta)$ where $\theta \in \Re^+$ is an unknown parameter. Here, $X_{n:n}$, the largest order statistic, is a minimal sufficient statistic for θ.

Theorem 4 (Distribution of a Minimal Sufficient Statistic in an Exponential Family). *Under the conditions of Theorem 2, the pmf or the pdf of the minimal sufficient statistic $(T_1, \ldots, T_k)$ also belongs to a k-parameter exponential family.*

In the case of population distributions not belonging to a regular exponential family, however, sometimes one may not achieve any substantial data reduction by invoking the concept of minimal sufficiency. For example, suppose that we have iid observations $X_1, \ldots, X_n$ having the following Cauchy pdf:

$$\frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, -\infty < x, \theta < \infty.$$

Here, $-\infty < \theta < \infty$ is an unknown location parameter. Now, let $\mathbf{T} = (X_{n:1}, \ldots, X_{n:n})$ where $X_{n:1} \leq \ldots \leq X_{n:n}$ are the sample order statistics. One can verify that $\mathbf{T}$ is a minimal sufficient statistic for θ.

A Complete Sufficient Statistic

Consider a real valued random variable X whose pmf or pdf is $f(x;\boldsymbol{\theta})$ for $x \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$. Let $T = T(X)$ be a statistic and suppose that its pmf or pdf is denoted by $g(t;\boldsymbol{\theta})$ for $t \in \mathcal{T}$ and $\boldsymbol{\theta} \in \Theta$. Then, $\{g(t;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta\}$ is called the family of distributions induced by T.

Definition 4 *The family $\{g(t;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta\}$ is called complete if and only if the following condition holds. Consider any real valued function $h(t)$ defined for $t \in \mathcal{T}$, having a finite expectation, such that*

$$E_{\boldsymbol{\theta}}[h(T)] = 0 \text{ for all } \boldsymbol{\theta} \in \Theta \text{ implies } h(t) \equiv 0 \text{ w.p.1.}$$

A statistic T is said to be complete if and only if $\{g(t;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \Theta\}$ is complete. A statistic $\mathbf{T}$ is called complete sufficient for $\boldsymbol{\theta}$ if and only if (1) *$\mathbf{T}$ is sufficient for $\boldsymbol{\theta}$ and* (2) *$\mathbf{T}$ is complete.*

A complete sufficient statistic, if it exists, is also a minimal sufficient statistic. For example, let $X_1, \ldots, X_n$ be iid

Uniform$(0, \theta)$ where $\theta \in \Re^+$ is unknown. Here, $X_{n:n}$, the largest order statistic, is a complete sufficient statistic for θ. Hence, $X_{n:n}$ is also minimal sufficient for θ. This proof bypasses Theorem 3. Now, we state a remarkably general result (Theorem 5) in the case of a regular exponential family of distributions. One may refer to Lehmann (1986, pp. 142–143) for a proof of this result.

Theorem 5 (**Completeness of a Minimal Sufficient Statistic in an Exponential Family**). *Under the conditions of Theorem 2, the minimal sufficient statistic* $(T_1, \ldots, T_k)$ *is complete.*

Ancillary Statistics

The concept called *ancillarity* of a statistic is perhaps the furthest away from the notion of sufficiency. A sufficient statistic $\mathbf{T}$ preserves all the information about $\boldsymbol{\theta}$ contained in the data $\mathbf{X}$. In contrast, an ancillary statistic $\mathbf{T}$ *by itself* provides *no information* about $\boldsymbol{\theta}$. This concept evolved from Fisher (1925) and later it blossomed into the vast area of *conditional inference*. In his 1956 book, Fisher emphasized many positive aspects of ancillarity in analyzing real data. For fuller discussions of *conditional inference* one may look at Basu (1964), Hinkley (1980) and Ghosh (1988). Reid (1995) provides an assessment of conditional inference procedures.

Consider the real valued observable random variables $X_1, \ldots, X_n$ from some population having the common pmf or pdf $f(x; \boldsymbol{\theta})$, where the unknown parameter vector $\boldsymbol{\theta} \in \Theta \subseteq \Re^p$. Let us continue writing $\mathbf{X}$ for the full data and $\mathbf{T} = \mathbf{T}(\mathbf{X})$ for a vector valued statistic.

Definition 5 *A statistic* $\mathbf{T}$ *is called ancillary for* $\boldsymbol{\theta}$ *or simply ancillary provided that the pmf or the pdf of* $\mathbf{T}$ *does not involve* $\boldsymbol{\theta}$.

Here is an important result that ties the notions of complete sufficiency and ancillarity. Basu (1955) came up with this elegant result which we state here under full generality.

Theorem 6 (**Basu's Theorem**). *Suppose that we have two vector valued statistics,* $\mathbf{U} = \mathbf{U}(\mathbf{X})$ *which is complete sufficient for* $\boldsymbol{\theta}$ *and* $\mathbf{W} = \mathbf{W}(\mathbf{X})$ *which is ancillary for* $\boldsymbol{\theta}$. *Then,* $\mathbf{U}$ *and* $\mathbf{W}$ *are independently distributed.*

An ancillary statistic by itself tells one nothing about $\boldsymbol{\theta}$! Hence, one may think that an ancillary statistic may not play a role to come up with a sufficient summary statistic. But, that may not be the case. The following examples will highlight the fundamental importance of ancillary statistics.

Example 3 Suppose that (X, Y) has a curved exponential family of distributions with the joint pdf given by

$$f(x, y; \theta) = \begin{cases} exp\{-\theta x - \theta^{-1} y\} & \text{if } 0 < x, y < \infty \\ 0 & \text{elsewhere,} \end{cases}$$

where $\theta (> 0)$ is an unknown parameter. This distribution was discussed by Fisher (1934, 1956) in the context of his famous "Nile" example. Denote $U = XY, V = X/Y$. One can show that U is ancillary for θ, V does not provide the full information about θ, but (U, V) is minimal sufficient for θ. Note that $V^{1/2}$ is the maximum likelihood estimator of θ, but it is not minimal sufficient for θ.

Example 4 This example was due to D. Basu. Let (X, Y) be distributed as bivariate normal with zero means, unit variances, and an unknown correlation coefficient ρ, $-1 < \rho < 1$. Then, marginally, both X and Y are distributed as standard normal variables. Clearly, X by itself is an ancillary statistic, Y by itself is an ancillary statistic, but X and Y combined has all the information about ρ.

About the Author

For biography *see* the entry ►Sequential Sampling.

Cross References

►Approximations for Densities of Sufficient Estimators
►Exponential Family Models
►Minimum Variance Unbiased
►Optimal Shrinkage Estimation
►Properties of Estimators
►Rao–Blackwell Theorem
►Statistical View of Information Theory
►Sufficient Statistical Information
►Unbiased Estimators and Their Applications

References and Further Reading

Bahadur RR (1954) Sufficiency and statistical decision functions. Ann Math Stat 25:423–462
Basu D (1955) On statistics independent of a complete sufficient statistic. Sankhyā 15:377–380
Basu D (1964) Recovery of ancillary information. Contributions to statistics, the 70th birthday festschrift volume presented to P. C. Mahalanobis. Pergamon, Oxford
Fisher RA (1920) A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error. Mon Not R Astron Soc 80:758–770
Fisher RA (1922) On the mathematical foundations of theoretical statistics. Philos Trans R Soc A222:309–368
Fisher RA (1925) Theory of statistical estimation. Proc Camb Philos Soc 22:700–725
Fisher RA (1934) Two new properties of mathematical likelihood. Proc R Soc A 144:285–307
Fisher RA (1956) Statistical methods and scientific inference. Oliver and Boyd, Edinburgh/London

Ghosh JK (ed) (1988) Statistical information and likelihood: a collection of critical essays by Dr. D. Basu. Lecture notes in statistics No. 45. Springer, New York
Halmos PR, Savage LJ (1949) Application of the Radon-Nikodym theorem to the theory of sufficient statistics. Ann Math Stat 20:225–241
Hinkley DV (1980) Fisher's development of conditional inference. In: Fienberg SE, Hinkley DV (eds) R. A. Fisher: an appreciation. Springer, New York, pp 101–108
Lehmann EL (1986) Testing statistical hypotheses, 2nd edn. Wiley, New York
Lehmann EL, Scheffé H (1950) Completeness, similar regions and unbiased estimation-Part I. Sankhyā 10:305–340
Lehmann EL, Scheffé H (1955) Completeness, similar regions and unbiased estimation-Part II. Sankhyā 15:219–236
Lehmann EL, Scheffé H (1956) Corrigenda: completeness, similar regions and unbiased estimation-Part I. Sankhyā 17:250
Mukhopadhyay N (2000) Probability and statistical inference. Marcel Dekker, New York
Neyman J (1935) Sur un teorema concernente le cosidette statistiche sufficienti. Giorn 1st Ital Att 6:320–334
Reid N (1995) The roles of conditioning in inference (with discussions by Casella G, Dawid AP, DiCiccio TJ, Godambe VP, Goutis C, Li B, Lindsay BC, McCullagh P, Ryan LA, Severini TA, Wells MT) Stat Sci 10:138–157

Summarizing Data with Boxplots

Boris Iglewicz
Professor
Temple University, Philadelphia, PA, USA

Introduction

Statisticians have created a variety of techniques for summarizing data graphically. For continuous univariate data the most commonly used graphical display is the histogram. Once the interval width is carefully determined, the histogram provides a visual summary of the data center, spread, ▶skewness, and unusual observations, which may be ▶outliers. While these features are visible, there are no specific numeric summary measures that are part of the histogram display.

Tukey (1977) introduced a simple alternative to the histogram that contains similar features as the histogram, is easier to graph, and includes measures of location, spread, skewness, and a rule for flagging outliers. He called this graphic summary the boxplot. The key components of the boxplot consist of Tukey's five number summary. These are: the median = Q_2; upper quartile = Q_3; lower quartile = Q_1; largest value = $X_{(n)}$; and the smallest value = $X_{(1)}$. This information is all that is needed to graph the simplest version of the boxplot, called the box-and-whisker plot. Such a plot is illustrated as the left plot of Fig. 1. The data consists of daily percent changes in the Dow Jones industrial average closing values for days when the market is open. Thus, if Y_t is the closing Dow Jones Industrial Average at day t, then the data for the boxplots in Fig. 1 consists of $X_t = 100(Y_t - Y_{t-1})/Y_{t-1}$.

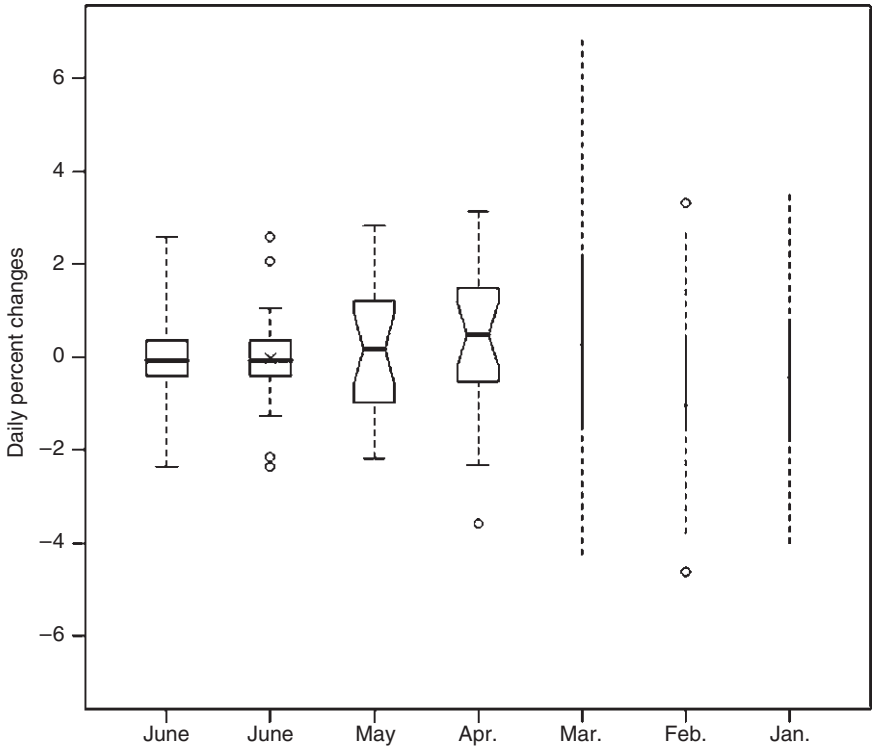
The box-and-whisker plot has a box at the center that contains approximately 50% of the middle observations. The horizontal line inside the plot is the median, Q_2, which provides a nice summary measure for the data center. The upper and lower horizontal lines enclosing the box are the values of Q_3, and Q_1, respectively. From these one can obtain the interquartile range, $IQR = Q_3 - Q_1$, which is a common robust measure of spread. Skewness can also be observed by comparing $Q_3 - Q_2$ with $Q_2 - Q_1$ or $X_{(n)} - Q_2$ with $Q_2 - X_{(1)}$.

Tukey (1977) also added a simple rule for flagging observations as potential outliers. That rule flags observations as outliers if they fall outside the interval $(Q_1 - k(IQR),\ Q_3 + k(IQR))$. Tukey suggested using $k = 1.5$ for a liberal interval with out values so designated. He also suggested using $k = 3$ to designate far out values. The box-and-whisker plot that incorporates the rule for flagging outliers is called a boxplot. The second from left plot in Fig. 1 illustrates such a boxplot for the June 2009 data. In addition, this boxplot contains an X in the middle designating the location of the sample mean. The inclusion of the sample mean is a useful added feature that some statistical computer packages incorporate.

Although the boxplot is a simple graphic summary procedure, a number of modifications have been suggested and properties studied. In section "▶Varied Versions of the Basic Boxplot" we will briefly review other variants of the basic boxplot. In section "▶Outlier Rule" we will consider further the properties of the outlier identification rule and suggest modified versions. In section "▶Quartiles" we will consider the computation of quartiles. A brief summary will be provided in section "▶Summary".

Varied Versions of the Basic Boxplot

A fair number of alternative versions of the basic boxplot have been introduced and used. Tufte (1983) suggested a slight modification that is useful in summarizing a large number of parallel boxplots that can be especially useful when dealing with data collected over many time periods. Tufte suggested removing the box, as in the three right most graphs in Fig. 1, representing the data for January, February, and March 2009, respectively. The box can now be represented by either a solid line or empty space.



Summarizing Data with Boxplots. Fig. 1 Graph contains several versions of boxplot construction based on daily percent changes of the Dow Jones industrial average grouped by month. The two left hand boxplots consist of June 2009 data with the right one including potential outliers. The next two to the right represent notched boxplots. The three right side boxplots are based on a version suggested by Tufte

The point inside the solid line designates the location of the median. The dashed lines go towards the largest and smallest observations excluding flagged outliers, which are individually plotted on the boxplot graph.

Another avenue of innovation is the thickness of the box. The simplest suggestion, given by McGill et al. (1978), is to make the width proportional to the square root of the sample size, thus showing precision. This is further refined by Benjamini (1988), who suggested replacing the two outer vertical lines of the box by density plots. Such density plots depend on the kernel and window width and are thus not unique. He called these plots histplots. Benjamini also introduced density plots for the entire vertical length of the boxplots. These plots he called vaseplots. Both the hisplots and vaseplots consist of lines. The vaseplot is further refined by Hintze and Nelson (1998) who used a curved density plot as a replacement. As the resulting plot often looks like a violin, they called their modification a violin plot.

A further refinement is the notched boxplot introduced by McGill et al. (1978). The goal is to provide a visual significance test comparing the medians of two adjacent boxplots. If the two medians lie within the two notches, then we can say that the two population medians are not significantly different. Two notched boxplots are shown as the April and May data in Fig. 1, where we can see that the two population medians are not significantly different. The intervals are based on asymptotic results from the normal distribution. These are refined in common statistical packages by using sign test type intervals. Benjamini (1988) suggested using the standard boxplot, but represent the notches by a shaded horizontal region.

Outlier Rule

Tukey's simple outlier labeling rule is heavily used, typically with $k = 1.5$, where observations are labeled as outliers if they lie outside the interval $(Q_1 - k(IQR),\ Q_3 + k(IQR))$. Hoaglin et al. (1986) studied the performance of this rule for random normal data. They found that the rule with $k = 1.5$ is very liberal for moderate to large data sets. For example, for $n = 300$ random normal observations there is an 85% chance that at least one observation will be falsely

labeled as an outlier. Even with $n = 100$ that percentage stays at 53%. For the conservative $k = 3.0$ rule, these out probabilities drop drastically to 0.2 percent for $n = 100$. The problem is that this $k = 1.5$ rule does not take sample-size into account. Consequently, the chances of labeling regular observations as outliers increase with increasing sample-size.

Let $B(k, n)$ = probability that all observations of a random normal sample lie inside $(Q_1 - k(IQR),\ Q_3 + k(IQR))$. Hoaglin and Iglewicz (1987) obtained values of k as functions of n to keep $B(k, n) = 0.95$ or $B(k, n) = 0.90$. That is, all n observations are inside the outlier labeling interval. Thus, for $n = 100$, $B(k, n) = 0.95$, they obtained $k = 2.2$, while for $n = 300, k = 2.4$. Iglewicz and Banerjee (2001) extended this procedure to random samples from a variety of both symmetric and skewed distributions in addition to the normal. Their work was further extended by Sim et al. (2005) and Banerjee and Iglewicz (2007).

Quartiles

Although the computation of quartiles seems to be quite simple on the surface, there are actually a number of choices for computing quartiles. As an example, Frigge et al. (1989) discuss eight options for computing quartiles. Although these choices will have limited effect for large samples, they can differ noticeably for small samples. That can lead to different boundaries for the box part of the boxplot and different values of k to maintain $B(k, n) = 0.95$.

Consider the non-negative number $f = j + g$, where j is the integer part of f and g the fractional part. For example, if $f = 12.8$, then $j = 12$ and $g = 0.8$. Consider the ordered observations $X_{(1)} \le X_{(2)} \le X_{(3)} \le \cdots \le X_{(n)}$, then $X_{(f)} = (1 - g)X_{(j)} + gX_{(j+1)}$. The median is typically obtained as $Q_{(2)} = X_{(f)}$, where $f = (n + 1)/2$. Letting $n = 2N + 1$ for n odd, $Q_{(2)} = X_{(N+1)}$. For $n = 2N$, n even, $Q_{(2)} = (X_{(N)} + X_{(N+1)})/2$. Tukey (1977) suggested a very simple rule for obtaining $Q_{(1)}$ and $Q_{(3)}$, as $Q_{(1)} = X_{(f)}$, where $f = (j + 1)/2$ and j = the integer part of $(n + 1)/2$. Then $Q_{(3)} = X_{(n+1-f)}$. An alternative popular choice for f in $X_{(f)} = Q_{(1)}$ is $f = (n + 1)/4$.

Summary

The boxplot is a heavily used graphical tool for summarizing univariate continuous data. Although the boxplot option shown on the second from the left plot of Fig. 1 is by far the most popular version, a variety of other useful choices have been discussed. These include the notched boxplots that are useful in comparing two population medians, the Tufte version useful when comparing many samples, and plots that incorporate density information. On some occasions, professionals are content with the simpler box-and-whisker plot illustrated as the leftmost plot of Fig. 1. While the illustrations of Fig. 1 consist of vertical boxplots, these could have just as effectively been drawn horizontally.

While this write-up has been devoted to discussion of the popular univariate boxplot, there have been a number of successful introductions of bivariate boxplots. These again use robust measures, but incorporate information on the correlation between the variables. Two bivariate boxplot versions worthy of note are by Goldberg and Iglewicz (1992) and Rousseeuw et al. (1999).

Acknowledgment

The author wishes to thank Alicia Stranberg for help with generating the graph. This article was written while on a Study Leave from Temple University.

About the Author

Boris Iglewicz serves as Professor of Statistics and Director of Biostatistics Research Center, Temple University. He received his Ph.D in statistics from Virginia Tech. At Temple University Dr. Iglewicz has served as the founding director of the graduate programs in statistics and as department chair. He also received the school's Musser Leadership Award for Excellence in Research and chosen as a Senior Research Fellow. Dr. Iglewicz has published about 70 professional journal articles, books, and chapters in books. From the American Statistical Association (ASA) he was chosen as a Fellow and received the following awards and recognitions: Chapter Recognition Award; W. J. Youden Award; Don Owen Award; and SPAIG award. He also served as President of the Philadelphia Chapter of ASA. He is also a Fellow of the Royal Statistical Society, Elected member of the International Statistical Institute, and serves as Associate Editor of *Statistics in Biopharmaceutical Research*. Dr. Iglewicz is listed in American Men and Women of Science, Who's Who in America, and Who's Who in the World.

Cross References

►Data Analysis
►Exploratory Data Analysis
►Five-Number Summaries
►Outliers

References and Further Reading

Banerjee S, Iglewicz B (2007) A simple univariate outlier identification procedure designed for large samples. Commun Stat Simul Comput 36:249–263

Benjamini Y (1988) Opening the box of a boxplot. Am Stat 42: 257–262

Frigge M, Hoaglin DC, Iglewicz B (1989) Some Implementations of the boxplot. Am Stat 43:50–54
Goldberg KM, Iglewicz B (1992) Bivariate extensions of the boxplot. Technometrics 34:307–320
Hintze J, Nelson RD (1998) Violin plots: a boxplot–density trace synergism. Am Stat 52:181–184
Hoaglin DC, Iglewicz B (1987) Fine-tuning some resistant rules for outlier labeling. J Am Stat Assoc 81:1147–1149
Hoaglin DC, Iglewicz B, Tukey JW (1986) Performance of some resistant rules for outlier labeling. J Am Stat Assoc 81:991–999
Iglewicz B, Banerjee S (2001) A simple univariate outlier identification procedure. In: Proceedings of the annual meeting of the american statistical association
McGill R, Tukey JW, Larson WA (1978) Variations of the box plots. Am Stat 32:12–16
Rousseeuw PJ, Ruts I, Tukey JW (1999) The Bagplot: a bivariate boxplot. Am Stat 53:382–387
Sim CH, Gan FF, Chang TC (2005) Outlier labeling with boxplot procedures. J Am Stat Assoc 100:642–652
Tufte E (1983) The visual display of quantitative information. Graphic Press, Cheshire
Tukey JW (1977) Introductory data analysis. Addison-Wesley, Reading

Superpopulation Models in Survey Sampling

Gad Nathan
Professor Emeritus
Hebrew University of Jerusalem, Jerusalem, Israel

Classical sampling theory considers a finite population, $U = \{1, \ldots, N\}$, of known size, N, with a vector of fixed unknown values of a variable of interest, $\boldsymbol{y} = (y_1, \ldots, y_N)$. A sample of size n, $s = \{s_{i_1}, \ldots, s_{i_n}\}$, is selected by a sample design, which assigns to each possible sub-set of U a known probability – $p(s)$. The objective is to estimate some function of $\boldsymbol{y}$, which can be assumed, without loss of generality, to be the population total, $\boldsymbol{y} = \sum_{i=1}^{N} y_i$, on the basis of the sample observations, $\{y_{i_1}, \ldots, y_{i_n}\}$, and the sample probabilities – $p(s)$. Inference based only on the sample selection probabilities is known as *design based* (or ►randomization) inference and the properties of estimators are considered in this framework solely with respect to the known sample selection probabilities. Although design-based inference is widely applied in practice for the estimation of finite population parameters, it suffers from several drawbacks:

1. It can be shown that there is no unbiased estimator, say of the total, which is optimal, in the sense that its randomization variance is minimal for all sets of possible values of the population variables (Godambe 1955).
2. While the use of auxiliary data, e.g., known values of an auxiliary variable for all population units, $\boldsymbol{X} = (X_1, \ldots, X_N)$, for sample design (e.g., stratification) or for estimation (e.g., ratio estimation) is widely applied, in practice, it cannot strictly be justified under the design-based paradigm, unless some model relating the values of X and Y is assumed. For instance the efficiency of ratio estimation is based on the premise that there is a linear relationship between the values of X and Y (without an intercept) – Cochran (1977).
3. The use of sample survey data for analytical purposes, which has developed extensively over the past few decades, cannot be treated on a solid theoretical basis solely under design-based inference -see e.g., Kish and Frankel (1974). Thus, although a regression analysis can formally be carried out on sample data, the results cannot be interpreted easily when the dependent and the independent variables are considered as fixed values, rather than as realizations of random variables, i.e., unless a linear model with random errors is assumed - Brewer and Mellor (1973).

This has led sample survey theoreticians and practitioners to consider a *model based*, or *superpopulation* approach, which assumes that each population unit is associated with a random variable for which a stochastic structure is specified and the actual value associated with a population unit is considered as the realization of the random variable, rather than a fixed unknown value - Cassel et al. 1976. Thus the vector of population values, $\boldsymbol{y}$, is assumed to be the realization of a random vector variable: $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$. The form of the joint distribution of $Y_1, \ldots, Y_N$, often denoted by ξ, is usually assumed to be known, except for unknown parameters. Thus, if we assume a regression model between $\boldsymbol{Y}$ and $\boldsymbol{X}$, we might consider ξ as multivariate normal, i.e., $\boldsymbol{Y}|\boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \Sigma)$, where $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are unknown parameters.

There are several different possible interpretations of the superpopulation concept, such as the following - see also Särndal et al. 1992:

1. The finite population may be considered as actually selected from a larger universe by a real world random mechanism or process. This would be the interpretation of a statistical model in the social sciences, such as econometric models. This is the approach usually used by practitioners who wish to analyze sample survey data created by complex sample designs – see, for

instance, Nathan and Holt (1980), Skinner et al. (1989), Pfeffermann (1993) and Chambers and Skinner (2003).
2. The superpopulation joint distribution, ξ, may be considered under a Bayesian approach, as a prior distribution, which reflects the subjective belief in the unknown values of $Y_1, \ldots, Y_N$, so that we consider the problem of finding the posterior distribution of the finite population parameter, given the sample values.
3. The superpopulation distribution may be considered as reflecting nonsampling errors, such as measurement errors, which account for differences between observed values of the variables and their 'true' values.
4. The superpopulation distribution, ξ, may be considered as a purely mathematical device, not associated with any physical process or subjective belief, in order to make explicit theoretical derivations. Thus different estimators or sample designs may be considered and compared, with respect to their performance and characteristics (e.g., bias and variance), under different models. Since in most cases our certainty about the true models is very limited, this can provide a useful tool for checking the robustness of estimators and sample designs to departures from assumed models.

The rapid development of sample survey theory and practice over the past 50 years has occurred in all aspects of sample surveys. However the rapid integration of the superpopulation concept and model-based ideas in mainstream theory and practice of sample survey inference has been one of the major developments. Thirty five years ago, the fundamental divide between advocates of classical design-based inference and design, and those who preferred basing both the sample design and inference only on superpopulation models was still at its zenith and the controversies of the two previous decades, exemplified by Brewer and Mellor (1973), were still raging. The early randomization-based approach, developed by the pioneers of classical design-based sampling theory, was challenged by the study of the logical foundations of estimation theory in survey sampling, for example, Godambe (1955), and by early advocates of pure superpopulation model-based design and prediction approach to inference, for example, Royall (1970). These controversies continued to be fiercely discussed well into the 1980s, see, for example, Hansen et al. (1983), and pure superpopulation based prediction approaches are still being advocated – see Valliant (2000). However the extreme views, relating to both approaches, have mellowed considerably over the past 2 decades, and sample survey theory and practice are currently, by and large, based on a variety of combined approaches, such as model-assisted methods, which integrate superpopulation models with a randomization-based approach – see for example the variety of approaches, many of them based on superpopulation models, used in the latest *Handbook of Statistics* (volume 29), devoted to sample surveys – Rao and Pfeffermann (2009).

The superpopulation concept has served and continues to serve as an extremely important and useful tool for the development of the theory and practice of sample surveys – in their design, estimation and analysis.

About the Author

Dr. Gad Nathan is Professor Emeritus, Department of Statistics, Hebrew University, Jerusalem (since 2002). He is Past President of the Israel Statistical Association (1991–1993). Professor Nathan was Chair, Department of Statistics, Hebrew University, Jerusalem (1974–1977, and 1988–1991). He was also Director, Statistical Methods Division, Central Bureau of Statistics, Jerusalem, (1964–1969), Chief Scientist, Central Bureau of Statistics (Part-time, 1995–2001), Vice-President, International Statistical Institute, (1981–1983), and Vice-President, International Association of Survey Statisticians (1999–2001). He is Elected Member of the International Statistical Institute (1977) and Elected Fellow of the American Statistical Association (1978). He has (co-)authored about 75 publications.

Cross References

►Model Selection
►Multivariate Normal Distributions
►Nonsampling Errors in Surveys
►Random Variable
►Randomization
►Sample Survey Methods
►Sampling From Finite Populations
►Small Area Estimation

References and Further Reading

Brewer KRW, Mellor RW (1973) The effect of sample structure on analytical surveys. Aust J Stat 15:145–152
Chambers RL, Skinner CJ (eds) (2003) Analysis of survey data. Wiley, New York
Cassel CM, Särndal CE, Wretman JH (1976) Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika 63:615–620
Cochran WG (1977) Sampling techniques 3rd edn. Wiley, New York
Godambe VP (1955) A unified theory of sampling from finite populations. J R Stat Soc B 17:269–278
Hansen MH, Madow WG, Tepping BJ (1983) An evaluation of model-dependent and probability-sampling inferences in sample surveys. J Am Stat Assoc 78:776–793
Kish L, Frankel M (1974) Inference from complex samples. J R Stat Soc B 36:1–37

Nathan G, Holt D (1980) The effect of survey design on regression analysis. J R Stat Soc B 43:377–386
Pfeffermann D (1993) The role of sampling weights when modeling survey data. Int Stat Rev 61:317–337
Rao CR, Pfeffermann D (eds) (2009) Handbook of statistics, 29: sample surveys: theory, methods and inference. Elsevier, Amsterdam
Royall RM (1970) On finite population sampling theory under certain linear regression models. Biometrika 57:377–387
Särndal CE, Swensson B, Wretman JH (1992) Model assisted survey sampling. Springer, New York
Skinner CJ, Holt D, Smith TMF (eds) (1989) Analysis of complex surveys. Wiley, Chichester
Valliant R, Dorfman AH, Royall RM (2000) Finite population sampling and inference: a prediction approach. Wiley, Chichester/New York

Surveillance

Marianne Frisén
Professor
University of Gothenburg, Gothenburg, Sweden

The Need for Statistical Surveillance

The aim of statistical surveillance is the timely detection of important changes in the process that generates the data. Already at birth surveillance is used, as described by Frisén (1992). The baby might get the umbilical cord around the neck at any time during labour. This will cause a lack of oxygen, and a Caesarean section is urgent. The electrical signal of the heart of the baby during labour is the base for the surveillance system. Detection has to be made as soon as possible to ensure that the baby is delivered without brain damage.

Around 1930, Walter A. Shewhart developed the first versions of sequential surveillance by introducing control charts for industrial applications (see ►Control Charts). Although industrial applications are still important, many new applications have come into focus.

In finance, transaction strategies are of great interest and the timeliness of transactions is important. Most theory of stochastic finance is based on the assumption of an efficient market. When the stochastic model is assumed to be completely known, we can use probability theory to calculate the optimal transaction conditions. When the information about the process is incomplete, as for example when a change can occur in the process, there may be an arbitrage opportunity, as demonstrated by Shiryaev (2002). In these situations, observations should be analysed continuously to decide whether a transaction at that time is profitable as measured either by return or by risk. Statistical inference is needed for the decision. Different aspects of the subject of financial surveillance are described in the book edited by Frisén (2007). There are also other applications in the field of economics. The *detection of turning points in business cycles* is important for both government and industry.

In public health surveillance, the timely detection of various types of adverse health events is crucial. The monitoring of incidences of different diseases and symptoms is carried out by international, national and local authorities to detect outbreaks of infectious diseases. Epidemics, such as influenza, are for several reasons very costly to society, and it is therefore of great value to monitor influenza data, both for the outbreak detection and during the epidemic period in order to allocate medical resources. Methods for surveillance for common diseases also serve as models for the detection of new diseases as well as for detecting bioterrorism. Surveillance for the onset of an outbreak is described in Frisén et al. (2009). Reviews of methods for the surveillance of public health are given by Sonesson and Bock (2003) and Woodall et al. (2008).

The Statistical Surveillance Problem

Terminology

The terminology is diverse. "Optimal stopping rules" (see ►Optimal Stopping Rules) is most often used in probability theory, especially in connection with financial problems. Literature on "change-point problems" does not always treat the case of sequentially obtained observations but often refers to the retrospective analysis of a fixed number of observations. The term "early warning system" is sometimes used in economic and medical literature. "Monitoring" is most often used in medical literature and with a broad meaning. The notations "statistical process control" and "quality control" are used in the literature on industrial production.

Overviews

Surveys and bibliographies on statistical surveillance are given for example by Lai (1995), who gives a full treatment of the field but concentrates on the minimax properties of stopping rules, by Woodall and Montgomery (1999) and Ryan (2000), who concentrate on control charts, and by Frisén (2003), who characterises methods by different optimality properties. The overview by Frisén (2009) and the adjoining discussion takes up many recent issues.

Differences between hypothesis testing and surveillance

In the initial example, the decision concerning whether the baby is at risk has to be made sequentially, based on the data collected so far. Each new time demands a new decision. There is no fixed data set but an increasing number of observations. In sequential hypothesis testing, we have sequentially obtained observations and repeated decisions, but the hypotheses are fixed. In contrast, there are no fixed hypotheses in surveillance. We can never accept any null hypotheses and turn our backs on the mother, since the baby might get the umbilical cord around the neck in the next minute.

Statistical specifications

We denote the process by $X = \{X(t) : t = 1, 2, \ldots\}$, where $X(t)$ is the observation (vector) made at time t, which is usually discrete. The purpose of the monitoring is to detect a possible change, for example the change in distribution of the observations due to the baby's lack of oxygen. The time of the change is denoted by τ. Before the change, the distribution belongs to the family f^D, and after the time τ, the distribution belongs to the family f^C. At each decision time s, we want to discriminate between two events, $C(s)$ and $D(s)$. For most applications, these can be further specified as $C(s) = \{\tau \leq s\}$ (a change has occurred) and $D(s) = \{\tau > s\}$ (no change has occurred yet), respectively.

We use the observations $X_s = \{X(t); t \leq s\}$ to form an alarm criterion which, when fulfilled, is an indication that the process is in state $C(s)$, and an alarm is triggered. We use an alarm statistic, $p(X_s)$, and a control limit, $G(s)$, and the alarm time, t_A, is $t_A = \min\{s; p(X_s) > G(s)\}$. The change point τ can be regarded either as a random variable or as a deterministic but unknown value, depending on what is most relevant for the application.

Evaluation and Optimality

Quick detection and few false alarms are desired properties of methods for surveillance. Different error rates and their implications for active and passive surveillance were discussed by Frisén and de Maré (1991).

Evaluation by significance level, power, specificity, sensitivity, or other well-known metrics may seem convenient. However, these are not easily interpreted in a surveillance situation. For example, when the surveillance continues, the specificity will tend to zero for most surveillance methods. Thus, there is not one unique specificity value in a surveillance situation.

Special metrics such as the expected time to a false alarm ARL^0 and the expected delay of a warranted alarm are used (see Frisén (1992)). The expected delay is different for early changes as compared with late ones. The most commonly used delay measure is ARL^1, the expected delay for a change that appears at the start of the surveillance.

In addition, the optimality criteria are different in surveillance as compared with hypothesis testing. The minimax optimality and the expected delay over the distribution of the change point are frequently used.

Methods

In surveillance, it is important to aggregate the sequentially obtained information in order to take advantage of all information. Different ways of aggregation meet different optimality criteria. Expressing methods for surveillance through likelihood functions makes it possible to link the methods to various optimality criteria. Many methods for surveillance can be expressed by a combination of partial likelihood ratios (Frisén (2003)). The likelihood ratio for a fixed value of τ is $L(s,t) = f_{X_s}(x_s|\tau = t)/f_{X_s}(x_s|D)$. The exact formula for these likelihood components will vary between situations.

The full likelihood ratio method (LR) can be expressed as a weighted sum of the partial likelihoods $L(s,t)$. It is optimal with respect to the criterion of minimal expected delay, as demonstrated by Shiryaev (1963).

The simplest way to aggregate the likelihood components is to add them. Shiryaev (1963) and Roberts (1966) suggested what is now called the Shiryaev-Roberts method. This means that all possible change times, up to the decision time s, are given equal weight.

The method by Shewhart (1931) is simple and the most commonly used method for surveillance. An alarm is given as soon as an observation deviates too much from the target. Thus, only the last observation is considered. The alarm criterion can be expressed by the condition $L(s,s) > G$, where G is a constant.

The CUSUM method was first suggested by Page (1954). The alarm condition of the method can be expressed by the partial likelihood ratios as $t_A = \min\{s; \max(L(s,t); t = 1, 2, \ldots, s) > G\}$, where G is a constant. The CUSUM method satisfies the minimax criterion of optimality, as proved by Moustakides (1986).

The alarm statistic of the EWMA method is an exponentially weighted moving average, $Z_s = (1 - \lambda)Z_{s-1} + \lambda X(s)$, $s = 1, 2, \ldots$ where $0 < \lambda < 1$ and Z_0 is the target value. The EWMA method was described by Roberts (1959).

Complex Situations

Applications contain complexities such as autocorrelations, complex distributions, complex types of changes and

spatial as well as other multivariate settings. Thus, the basic surveillance theory has to be adapted to special cases.

Time series with special dependencies have been treated for example by Basseville and Nikiforov (1993), Schmid (1997) and Lai (1998). Surveillance for special distributions such as, for example, discrete ones were discussed for example by Woodall (1997). Complex changes such as gradual ones from an unknown baseline are of interest at the outbreak of influenza or other diseases. The maximal partial maximum likelihood will give a CUSUM variant. This was used for semiparametric surveillance by Frisén et al. (2009).

Multivariate surveillance is of interest in many areas. In industry, the monitoring of several components in an assembly process requires multivariate surveillance. An example in finance is the on-line decisions on the optimal portfolio of stocks, as described by Okhrin and Schmid (2007). The surveillance of several distribution parameters, such as the mean and the variance (see e.g., Knoth and Schmid (2002)), is another example of multivariate surveillance.

In spatial surveillance, observations are made at different locations. Most methods for spatial surveillance are aimed at detecting spatial clusters, but other relations between the variables can also be of interest. The surveillance of a set of variables for different locations is a special case of multivariate surveillance, as discussed by Sonesson and Frisén (2005) and Sonesson (2007).

About the Author

Dr. Marianne Frisén is Professor of Statistics at University of Gothenburg. She is an Elected member of the ISI. Professor Frisén has has been working in the area of surveillance for over 25 years. She has organized symposiums on financial surveillance, written numerous publications on surveillance, including the text *Financial Surveillance* (Wiley, 2007), the first book-length treatment of statistical surveillance methods used in financial analysis.

Cross References

►Detection of Turning Points in Business Cycles
►Optimal Stopping Rules
►Relationship Between Statistical and Engineering Process Control
►Sequential Probability Ratio Test
►Sequential Sampling
►Significance Testing: An Overview

References and Further Reading

Basseville M, Nikiforov I (1993) Detection of abrupt changes: theory and application. Prentice Hall, Englewood Cliffs

Frisén M (1992) Evaluations of methods for statistical surveillance. Stat Med 11:1489–1502

Frisén M (2003) Statistical surveillance. Optimality and methods. Int Stat Rev 71:403–434

Frisén M (2007) Financial surveillance, edited volume. Wiley, Chichester

Frisén M (2009) Optimal sequential surveillance for finance, public health and other areas. Editor's special invited paper. Sequential Anal 28:310–337, discussion 338–393

Frisén M, de Maré J (1991) Optimal surveillance. Biometrika 78: 271–280

Frisén M, Andersson E, Schiöler L (2009) Robust outbreak surveillance of epidemics in Sweden. Stat Med 28:476–493

Knoth S, Schmid W (2002) Monitoring the mean and the variance of a stationary process. Stat Neerl 56:77–100

Lai TL (1998) Information bounds and quick detection of parameters in stochastic systems. IEEE Trans Inform Theor 44: 2917–2929

Moustakides GV (1986) Optimal stopping times for detecting changes in distributions. Ann Stat 14:1379–1387

Okhrin Y, Schmid W (2007) Surveillance of univariate and multivariate nonlinear time series. In: Frisén M (ed) Financial surveillance. Wiley, Chichester, pp 153–177

Page ES (1954) Continuous inspection schemes. Biometrika 41: 100–114

Roberts SW (1959) Control chart tests based on geometric moving averages. Technometrics 1:239–250

Roberts SW (1966) A Comparison of some control chart procedures. Technometrics 8:411–430

Ryan TP (2000) Statistical methods for quality improvement. Wiley, New York

Schmid W (1997) Cusum control schemes for Gaussian processes. Stat Pap 38:191–217

Shewhart WA (1931) Economic control of quality of manufactured product. MacMillan, London

Shiryaev AN (1963) On optimum methods in quickest detection problems. Theor Probab Appl 8:22–46

Shiryaev AN (2002) Quickest detection problems in the technical analysis of financial data. In: Geman H, Madan D, Pliska S, Vorst T (eds) Mathematical finance – bachelier congress 2000. Springer, Berlin, pp 487–521

Sonesson C, Bock D (2003) A review and discussion of Prospective statistical surveillance in public health, JR Stat Soc A 166:5–21

Sonesson C (2007) A cusum framework for detection of space-time disease clusters using scan statistics. Stat Med 26: 4770–4789

Sonesson C, Frisén M (2005) Multivariate surveillance. In: Lawson A, Kleinman K (eds) Spatial surveillance for public health. Wiley, New York, pp 169–186

Woodall WH (1997) Control charts based on attribute data: bibliography and review. J Qual Technol 29:172–183

Woodall WH, Montgomery DC (1999) Research issues and ideas in statistical process control. J Qual Technol 31: 376–386

Woodall WH, Marshall JB, Joner JMD, Fraker SE, Abdel-Salam ASG (2008) On the use and evaluation of prospective scan methods for health-related surveillance. J R Stat Soc A 171: 223–237

Sonesson, C. and Bock, D. (2003) A review and discussion of prospective statistical surveillance in public health. *J R Stat Soc A 166*: 5–21

Survival Data

D. R. Cox
Honorary Fellow
Nuffield College, Oxford, UK

Preliminaries

The most immediate examples of survival data come from ►demography and actuarial science and concern the duration of human life. The issues of statistical analysis that arise are similar to those in many fields. Thus survival time may be the length of time before a piece of industrial equipment fails, the length of time before a firm becomes bankrupt, the duration of a period of employment or, particularly in a medical or epidemiological context, the time between diagnosis of a specific condition and death from that condition.

Depending on the perspective involved the term failure time may be used instead of survival time.

Central requirements are that for each study individual we have a clear time origin and a clear end point. For example, time may be measured from the instant an individual enters the study population and the end point may be death from a specific cause, or death (all causes) or cure. Normally the passage of time is clearly defined in the natural way. There may be other possibilities, for example the investigation of tire life in terms of km driven. In applications considerable care is needed over these definitions, ensuring that they are precise and relevant.

A common characteristic of such data is that the frequency distributions are widely dispersed with positive skewness. Another is the presence of right censoring. That is for some, or in some cases, for many individuals, all that is known is that by the end of the study the critical event in question has not occurred, implying that the survival time in question exceeds some given value. In industrial life testing censoring may be by design but more commonly it is just a feature of the data acquisition process.

Formalization

We represent survival time by a random variable T, treated for simplicity as continuously distributed; there is a closely parallel discussion for discrete random variables.

For a given population of individuals the distribution of T can be described in several mutually equivalent ways, for example by

- the survivor function

$$S(t) = P(T > t), \tag{1}$$

- the probability density function

$$f(t) = -S'(t) \tag{2}$$

- the hazard or age-specific failure rate

$$h(t) = f(t)/S(t) = -\frac{d}{dt}\log S(t). \tag{3}$$

A more interpretable specification of the hazard at time t is as a failure rate conditional on survival to time t, that is as

$$\lim P(T < t + \delta \mid t < T)/\delta$$

as δ tends to zero through positive values.

These three specifications are mathematically equivalent; all have their uses in applications.

A central role is played in some parts of the subject by the exponential distribution of rate ρ and mean $1/\rho$, namely the special case

$$S(t) = e^{-\rho t}, \quad f(t) = \rho e^{-\rho t}, \quad h(t) = \rho. \tag{4}$$

The last property shows that failure occurs at random with respect to "age." If $h(t)$ increases with t there is ageing whereas if $h(t)$ decreases with t then in a certain sense old is better than new. There are other possibilities, in particular a bath-tub effect in which high initial values are followed by a decrease followed in turn by a gradual increase.

Many other forms may be used in applications, notably the ►Weibull distribution with $h(t) = \rho(\rho t)^{\gamma}$.

Statistical Analysis

For n independent individuals from a homogenous population it is convenient to write the data in the form

$$(t_1, d_1), \ldots, (t_n, d_n). \tag{5}$$

Here for individual j, t_j is a time and if $d_j = 1$ this is the relevant value of T whereas if $d_j = 0$ the individual is right censored. This is interpreted to mean that all we know about the value of T for that individual exceeds t_j, a nontrivial assumption implying what is rather misleadingly called uninformative censoring. It excludes for example the deliberate or unwitting withdrawal of individuals from a study because of a presumption of imminent failure.

There are two broad approaches to analysis, parametric based on an assumed form for the distribution, and nonparametric.

The former is typically tackled by the method of maximum likelihood. Let θ denote the parameter specifying the distribution, for example ρ for the exponential distribution and (ρ, γ) for the Weibull distribution. Then the

likelihood is

$$\Pi\{f(t_j;\theta\}^{d_j}\{S(t_j;\theta)\}^{1-d_j}. \quad (6)$$

That is, each failed individual contributes a term depending on the density whereas each censored individual contributes a term depending on the survivor function. The method of maximum likelihood may now be applied (or a Bayesian posterior density calculated).

For the exponential distribution the likelihood takes the form

$$\rho^{\Sigma d_j}\exp(-\rho\Sigma t_j). \quad (7)$$

where Σd_j is the total number of failures. It follows that the maximum likelihood estimate of ρ, obtained by maximizing this expression with respect to ρ, is

$$\frac{\Sigma d_j}{\Sigma t_j}, \quad (8)$$

that is, the total number of failures divided by the total time at risk calculated from all individuals those who fail and those who are censored. This is sometimes called the fundamental theorem of epidemiology.

For a nonparametric analysis a limiting form of a life-table approach is used called the Kaplan-Meier method. Essentially the hazard is estimated as zero at all times at which failures do not occur and as the number of failures divided by the number at risk of failure at times at which failure does occur. The estimated survivor function is reconstructed from this by a discrete version of (3). If required, estimates of, say, the median survival time can be found by interpolation, assuming that sufficient failures have occurred to allow this part of the distribution to be estimated effectively.

Dependencies and Comparisons

Often there are more than a single group of observations and comparisons are required, say between groups of individuals treated differently. In simple cases this can be achieved either by comparing parameters in parametric models fitted separately to the different groups or by graphical comparison of the Kaplan-Meier estimates (see ►Kaplan-Meier Estimator).

In more complicated cases, for example when several explanatory variables are addressed simultaneously, models analogous to regression models are helpful. The most widely used of these is the proportional hazards model. For each individual we suppose available a vector z of explanatory variables and that the corresponding hazard function is

$$h_0(t)\exp(\beta^T z). \quad (9)$$

Here $h_0(t)$, called the baseline hazard, specifies the hazard for a reference individual with $z = 0$.

A typical example with critical event death from cardio-vascular causes might have z_1, age at entry, z_2, systolic blood pressure at entry, both typically measured from some reference level, z_3, zero for men, one for women and z_4, zero for control and one for a new drug under test. A component of β, say the first component β_1, specifies the increase in hazard per unit increase in the component z_1 of z, with all other components of z held fixed. That is for fixed gender, treatment and blood pressure the hazard increases by a factor e^{β_1} per extra year of age.

If the baseline hazard is constant or specified parametrically maximum likelihood estimation is possible, essentially generalizing (3). For example if $h_0(t)$ is an unknown constant, a baseline individual has an exponential distribution. If $h_0(t)$ is left arbitrary a modified form of likelihood-based inference is used called partial likelihood. Problems of interpretation, model choice, etc., are essentially the same as in multiple linear regression. An important possibility is that some components of z may be functions of time.

Generalizations and Literature

There are many generalizations of these ideas of which the most notable is to event-history analysis in which a sequence of events, possibly of different types, may occur on each individual.

There is a very extensive literature, some of it specific to application fields. Cox and Oakes (1984) give a broad introduction and Kalbfleisch and Prentice (2002) a more specialized and thorough account. For a discussion with attention to mathematical detail, see Andersen et al. (1993) and Aalen et al. (2008).

About the Author

Sir David Cox is among the most important statisticians of the past half-century. He has made major contributions to statistical theory, methods, and applications. He has written or co-authored 18 books and more than 300 papers, many of which are seminal works. He was editor of *Biometrika* for 25 years (1966–1991). Professor Cox was elected a Fellow of the Royal Society (F.R.S.) in 1973, knighted by Queen Elizabeth II in 1985 and became an Honorary Fellow of the British Academy in 1997. He has served as President of the Bernoulli Society (1979–1981), of the Royal Statistical Society (1980–1982), and of the International Statistical Institute (1995–1997). He is a Fellow of the Royal Danish Academy of Sciences, of the Indian Academy of Sciences and of the Royal Society of Canada and a Foreign Associate of the US National Academy of

Sciences. He has been awarded the Guy Medal in Silver, Royal Statistical Society (1961), Guy Medal in Gold, Royal Statistical Society (1973), Weldon Memorial Prize, University of Oxford (1984), Kettering Prize and Gold Medal for Cancer Research (1990), Marvin Zelen Leadership Award, Harvard University (1998). In 2010 he received the Copley Medal of the Royal Society. He has supervised or been associated with more than 60 doctoral students, many of whom have become leading researchers themselves (including a number of authors of this Encyclopedia: Anthony Atkinson, Adelchi Azzalini, Gauss Cordeiro, Vern Farewell, Roderick Little, Francisco Louzada-Neto, Peter McCullagh, and Basilio Pereira). Professor Cox holds 21 honorary doctorates, the last one from the University of Gothenburg (2007).

"His outstanding contributions to the theory and applications of statistics have a pervasive influence. For instance, the introduction of what is nowadays called 'Cox regression' in survival analysis has started a research area with numerous books and thousands of papers on statistical theory and on statistical practice. It has changed the way in which survival studies in medicine and technology are performed and evaluated." (Inauguration of Doctors Ceremony, University of Gothenburg, October 20, 2007).

Cross References

►Bayesian Semiparametric Regression
►Censoring Methodology
►Degradation Models in Reliability and Survival Analysis
►Demographic Analysis: A Stochastic Approach
►Event History Analysis
►First-Hitting-Time Based Threshold Regression
►Frailty Model
►Generalized Weibull Distributions
►Hazard Ratio Estimator
►Hazard Regression Models
►Kaplan-Meier Estimator
►Life Table
►Logistic Distribution
►Medical Research, Statistics in
►Modeling Survival Data
►Population Projections
►Statistical Inference in Ecology
►Testing Exponentiality of Distribution
►Weibull Distribution

References and Further Reading

Aalen OO, Borgan O, Gjessing HK (2008) Survival and event history analysis. Springer, New York

Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Cox DR, Oakes D (1984) Analysis of survival data. Chapman & Hall, London

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data. 2nd edn. Wiley, New York

T

Target Estimation: A New Approach to Parametric Estimation

Luisa Turrin Fernholz
Professor Emerita of Statistics
Temple University, Philadelphia, PA, USA

Introduction and Definition

Target estimation is a computer intensive procedure introduced by Cabrera and Fernholz (1999) that has proved to be effective in reducing the bias as well as the L_1 and L_2 errors of statistics in parametric settings.

For a statistical functional T, let the statistic $T(F_n)$ estimate the parameter $T(F_\theta)$, where F_n is the empirical d.f. corresponding to the sample $X_1, \ldots, X_n$ of i.i.d. random variables. Suppose that all the X_i's have common d.f. F_θ where $\theta \in \Theta$, an open subset of real numbers. If the expectation of $T(F_n)$, $g(\theta) = E_\theta(T(F_n))$, exists for all $\theta \in \Theta$ and is one-to-one and differentiable, then the functional $\widetilde{T}$ induced by T from the relation

$$g^{-1}(T) = \widetilde{T}$$

will be called the *target functional* of T. The statistic $\widetilde{T}(F_n)$ will be called the *target estimator*.

Remarks

a. Note that the target estimate of θ corresponds to choosing the value $\widetilde{\theta} = \widetilde{T}(\widehat{F}_n)$, which solves the equation

$$g(\widetilde{\theta}) = E_{\widetilde{\theta}}(T(F_n)) = T(\widehat{F}_n)$$

where $\widehat{F}_n$ is the observed value of F_n. That is, we set the expectation of a statistic equal to its observed value and we solve for θ. Also, note that g depends on the sample size n which will remain fixed.

b. It is a direct consequence of the definition that if T is a statistical functional with $g(\theta) = a\theta + b$ for $a \neq 0$, then the corresponding target estimator will be unbiased. The variance of $\widetilde{T}$ will satisfy

$$Var(\widetilde{T}) = (1/a^2)\,Var(T)$$

and the variance of the target estimator will be reduced if and only if $a^2 > 1$.

Properties of Target Estimators

For general estimators, Cabrera and Fernholz (1999) give some results regarding bias and variance reduction after targeting. These results can be summarized as follows:
If $g(\theta) > \theta$ and g is increasing, then:

1. If $1 < g'(\theta) < b$ then $|B_{\widetilde{T}}(\theta)| < |B_T(\theta)|$,
2. If $1 < |g'(\theta)|$ then $MSE(\widetilde{T}) < Var(T)$ and $E|\widetilde{T} - \theta| < E|T - \theta| + |Med(T) - E(T)|$,

where $B_T(\theta)$ and $B_{\widetilde{T}}$ denote the bias of T and $\widetilde{T}$ respectively, $Med(T)$ is the median of T, and MSE is the mean square error.

von Mises Expansions of Target Functionals

The von Mises expansions for the target functional $\widetilde{T}$ can be obtained using the Hadamard or Fréchet derivatives of the functional T. These expansions are useful to analyze the bias of $\widetilde{T}$ as well as the asymptotics and robustness properties of $\widetilde{T}$. For $T(F_n)$ the first order von Mises expansion is: $T(F_n) = \theta + \frac{1}{n}\sum_1^n \varphi(X_i) + Rem$. Then, under some regularity conditions, the remainder satisfies $\sqrt{n}Rem = o_P(1)$, and the statistic $T(F_n)$ is asymptotically normal (see Fernholz 1983). Moreover, when φ is properly normalized, the expectation of $T(F_n)$ gives: $g(\theta) = \theta + E_\theta(Rem)$ so that $T = \widetilde{T} + E_{\widetilde{T}}(Rem)$, and the bias of the target estimator is $B_{\widetilde{T}}(\theta) = E_\theta(Rem_1 - E_{\widetilde{T}}(Rem))$, which under certain conditions satisfies,

$$|B_{\widetilde{T}}(\theta)| = |E_\theta(Rem - E_{\widetilde{T}}(Rem))| < |E_\theta(Rem)| = |B_\theta(T)|.$$

Using the von Mises expansions of T, it can be shown that the ▶asymptotic normality of $\widetilde{T}$ is inherited from the asymptotic normality of T, with some gain in asymptotic efficiency when $|g'(\theta)| > 1$. The robustness aspects of target functionals are also analyzed using the von Mises approach and the influence functions of $\widetilde{T}$ and T are related by:

$$\mathrm{IF}_{\widetilde{T}}(x) = (1/g'(\theta))\mathrm{IF}_T(x).$$

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

This shows that the gross-error sensitivity of the target functional is lower when $|g'(\theta)| > 1$. See Fernholz (1997) and Cabrera and Fernholz (1999).

Target Estimation in Multidimensional Settings

Multivariate target estimation was treated in Cabrera and Fernholz (2004) where p-dimensional statistical functionals $T = (T_1, \ldots, T_p)$ estimate a p-dimensional parameter vector $\theta = (\theta_1, \ldots, \theta_p)$. In this case the expectation function $g(\theta)$, as defined in section "►Introduction and Definition", is p-dimensional, and for the simple case where g is an affine function of the parameter vector, the bias can be removed entirely and, under certain conditions, the variability of the bias corrected functional is reduced in the sense of smaller trace and smaller determinant. Examples of multivariate targeting for location-scale equivariant estimators and the location-scale exponential model are given in Cabrera and Fernholz (2004).

In practice, we seldom have linearity of the p-dimensional expectation function g. Quite often, the p-dimensional estimator T is defined implicitly and the corresponding target estimator must be found by solving multidimensional implicit equations in θ, of the form $g(\theta) = T(F_n)$ where $T(F_n)$ has been observed and $g(\theta) = E_\theta(T(F_n))$ is multidimensional. This amounts to inverting the function g which, if unknown, must first be estimated. The method of *stochastic approximation* introduced by Robbins and Munro (1951) and modified by Cabrera and Hu (2001) was successfully used to find the target estimates in many situations. For details and description of this methods see Cabrera and Fernholz (2004) and Cabrera et al. (2005).

Applications and Examples

Target estimation has been successfully used for bias and variance reduction in many cases. The following are just some of the more important cases developed:

1. *Ellipse estimation.* The case of ellipse estimation when only an arc of data points is available is of particular importance in computer vision since many real life problems encounter this difficulty. A study regarding the least squares estimators of five parameters identifying an ellipse can be found in Cabrera and Fernholz (2004) where a comparison of the target estimators with both the bootstrap (see ►Bootstrap Methods) and the jackknife estimators (see ►Jackknife) shows the advantages of the target estimation method in terms of reducing bias and lowering the variability of the estimators.
2. *Autoregressive Models.* Simulations were performed for autoregressive models AR(1) of the form $X_{t+1} = \theta X_t + \epsilon_t$, where the error term ϵ_t is Gaussian. The maximum likelihood estimator (MLE) of the parameter θ was compared to the corresponding target estimator for different sample sizes and different values of θ. These simulations showed a substantial reduction in the bias of the target estimator as compared to the bias of the MLE for every case considered, and they also showed that the MSE of the target estimator was reduced in most of the cases. See Cabrera and Fernholz (1999).
3. *Errors-in-variables Models.* General errors-in-variables models of the form $Y = a + bU + \epsilon$ when the observable variables are $X = U + \delta$, where ϵ and δ are independent Gaussian errors. In all the simulations performed for different sample sizes and different values of b the bias of the target estimator was substantially reduced when compared to the bias of the MLE, and in all cases the MSE of the target estimator was smaller than that of the MLE. See Cabrera and Fernholz (1999).
4. *Logistic Regression Models.* A treatment of logistic regression models (see ►Logistic Regression) of one and two parameters was given in Cabrera et al. (2005) where it is shown that the transformed MLE, i.e., the target estimator, has lower bias and MSE that the original MLE. It was also shown that another benefit of targeting is that it corrects the asymmetry of the statistic thus producing target statistics with more symmetric distributions.

Final Remarks

1. *Comparison to the Bootstrap.* Target estimation has been compared to other methods of reducing bias and variability such as the jackknife and the bootstrap. This comparison is treated in Cabrera and Fernholz (1999, 2004), where for different situations it was shown that targeting can provide considerable improvement over both the jackknife and the bootstrap in lowering the bias and the MSE.
2. *Median Target.* When the sampling distribution of the statistic is skewed or has heavy tails, the mean of the statistic may not be the proper measure of location to be considered, or may not even be defined. In such cases the mean target defined above may not be the proper approach. However, in these situations we can consider the median of the statistic as a function of θ by taking $g(\theta) = \text{med}_\theta T(F_n)$ and defining the *median target estimate* in an analogous way. The resulting median target estimate will always be *median unbiased* when the g function is monotone; this is a drastic difference with the mean target situation where

some additional regularity conditions for g are needed. Results in this direction can be found in Cabrera and Watson (1996) and Cabrera et al. (2005), but many open questions about median target estimates and their variability are still awaiting their answers.

About the Author

Biography of Fernholz is in ►Functional Derivatives in Statistics: Asymptotics and Robustness.

Cross References

►Bias Correction
►Bootstrap Methods
►Estimation
►Estimation: An Overview
►Functional Derivatives in Statistics: Asymptotics and Robustness
►Jackknife
►Logistic Regression

References and Further Reading

Cabrera J, Fernholz LT (1999) Target estimation for bias and mean square error reduction. Ann Stat 27 3:1080–1104
Cabrera J, Hu I (2001) Algorithms for target estimation using stochastic approximation. InterStat 02–04:1–18
Cabrera J, Watson GS (1997) Simulation methods for mean and median bias reduction in parametric estimation. J Stat Plann Inference 57(1):143–152
Cabrera J, Devas V, Fernholz LT (2005) Target estimation for the logistic regression model. J Stat Comput Simul 75: 121–140
Fernholz LT (1983) Von Mises calculus for statistical functionals. Lecture notes in statistics, vol 19. Springer, New York
Fernholz LT (1997) Target estimation and implications to robustness. In: L_1-statistical procedures and related topics. IMS Lecture notes, monograph series, vol 31, pp 363–372
Robbins H, Munro S (1951) A stochastic approximation method. Ann Math Stat 22:400–407

Telephone Sampling: Frames and Selection Techniques

James M. Lepkowski
Director, Program in Survey Methodology, Research Professor, Survey Research Center
University of Michigan, Ann Arbor, MI, USA

Telephone sampling is a set of techniques used to generate samples in telephone survey data collection. Telephone surveys have lower cost and time of data collection than face-to-face survey methods. (Telephone surveys are also conducted for other types of units, such as business establishments. This discussion is limited to telephone household surveys.). Cost and timeliness advantages outweigh potential loss in accuracy due to failure to cover households without telephones. However, since households without telephones vary in character over time and across countries and key subgroups, researchers must decide in any particular application whether non-coverage bias is a potentially serious source of error before choosing to use a telephone survey.

Telephone sampling methods use traditional sampling techniques or modifications of those techniques designed to address the nature of the materials available for sample selection. The materials, or frames, are of two basic types: lists of telephone household numbers and lists of groups of potential telephone household numbers.

Telephone household number lists come from commercial or government sources. Some cover virtually all, or a high percentage of all, telephone households in a target population, such as those obtained from a government agency providing telephone service. Alternatively, list frame numbers may be from published telephone directories that include a majority but not all telephone households. Telephone directories do not cover recent subscribers or subscribers who do not want to have a number appear in the directory, and substantial telephone household non-coverage arising from out-of-date or absent entries has led alternative frames with more complete coverage.

Telephone sampling for list frames uses traditional element sampling techniques such as systematic selection and stratified random sampling. The lists and samples contain numbers that are not telephone household numbers, which requires screening during data collection to eliminate non-household numbers. Some telephone households have more than one telephone number in the list, which in turn have higher chances of selection. Weights are used to compensate for the duplicate numbers. If persons within households are to be sub-selected, within household selection methods choose one or more sample persons within a household, yielding additional adjustment weights.

Alternative frames or sets of materials provide more complete, if not virtually complete, coverage than directory list frames. The alternative frames are used to complete through random generation of some portion of a telephone number telephone numbers where only an area code and local prefix combination are available, and are often referred to as random digit dialing (RDD; see Groves and Kahn 1979). The frame consists of all area code and local area prefixes for a country or region obtained from government or commercial sources. These combinations

are not complete telephone numbers, but randomly generated 'suffixes' added to a selected combination yield a valid and complete telephone number. The combination plus random digits cover, in principle, all telephone households provided all combinations in the region are available.

Simple RDD telephone number generation is typically very inefficient due to a large percentage of randomly generated numbers (sometimes in excess of 80 percent) that are not telephone households, increasing costs through screening to find telephone households among randomly generated numbers. Specialized techniques reduce the percentage of non-household numbers obtained, and improve efficiency. For example, Mitofsky–Waksberg RDD sampling (Waksberg 1978) is a two-stage sampling technique devised to randomly generate numbers that have a much lower percentage of non-telephone households, below 35 percent in early applications in the United States. Practical deficiencies led to variations to improve efficiency of the two-stage methods (see, for example, Potthoff 1987).

List-assisted methods seek efficiency gains as well, but start from a directory frame to extend coverage to all telephone households (Tucker et al. 2001). Many have a slight loss of coverage, though, compared to RDD methods. Numbers selected from a directory are selected and digits in the number altered to cover numbers that are not in the directory. Plus-one dialing, for example, replaces the last digit of a directory number with a number one larger – 8 instead of 7, for instance, replaces the last digit of a phone number ending in 7. While in principle this method should cover all telephone numbers, in practice the coverage is incomplete, and difficult to determine. Variations include changing the last digit or the last two digits randomly (Sudman 1973).

Commercial sources compile lists of all directory numbers in a country, metropolitan area, or region that are used to generate telephone numbers with higher levels of coverage. Phone numbers can be divided into sets of 100 consecutive numbers defined by all but the last two digits of a telephone number. For example, directory entry 7345551212 defines 100 consecutive numbers 7345551200 to 7345551299. Commercial sources use directory entries to find all 100 "banks" where at least one directory number is present. Telephone numbers are selected at random from all numbers occurring in the set of 100 'banks' that contain one or more directory numbers. These methods provide today higher efficiency than even the two-stage RDD methods (Casady and Lepkowski 1993).

Finally, dual frame sampling designs have been used to select samples separately from directory and RDD frames and combine the results in estimation (Lepkowski 1988). These methods are also currently being used to include telephone households that have only mobile or cell telephones and are not covered by list assisted sampling frames.

About the Author

Dr. James M. Lepkowski is Professor, Research Professor, and Director, Program in Survey Methodology, at the University of Michigan, USA. He directed the Michigan Summer Institute in Survey Research Techniques (1997–2004) and the Sampling Program for Survey Statisticians (since 1992). He is a Fellow of the American Statistical Association and an Elected Member of the International Statistical Institute. He has authored or co-authored more than 80 peer-reviewed papers and books and monographs, including the textbook *Survey Methodology* (2nd Edition, Wiley, 2009) and the edited volume *Advances in Telephone Survey Methodology* (Wiley, 2007).

Cross References

►Federal Statistics in the United States, Some Challenges
►Nonsampling Errors in Surveys
►Public Opinion Polls
►Representative Samples
►Sample Survey Methods
►Sampling From Finite Populations
►Statistical Fallacies

References and Further Reading

Casady RJ, Lepkowski JM (1993) Stratified telephone sampling designs. Surv Meth 19:103–113

Groves RM, Kahn R (1979) Surveys by telephone: A national comparison with personal interviews. Academic Press, New York

Lepkowski JM (1988) Telephone sampling methods in the United States. In: Groves RM et al (eds) Telephone survey methodology, Chap. 3, Wiley, New York

Potthoff RF (1987) Some generalizations of the Mitofsky–Waksberg techniques for random digit dialing. J Am Stat Assoc 82: 409–418

Sudman W (1973) The uses of telephone directories for survey sampling. J Market Res 10:204–207

Tucker C, Lepkowski JM, Piekarski L (2001) The current efficiency of list-assisted telephone sample designs. Public Opin Quart 66:321–338

Waksberg J (1978) Sampling methods for random digit dialing. J Am Stat Assoc 73:40–46

Testing Exponentiality of Distribution

John Haywood[1], Estate V. Khmaladze[2]
[1]Senior Lecturer
Victoria University of Wellington, Wellington, New Zealand
[2]Professor
Victoria University of Wellington, Wellington, New Zealand

The exponential distribution, defined on the positive half-line $\mathbb{R}^+$ with scale parameter $\lambda > 0$, has distribution function and density

$$F_\lambda(x) = 1 - e^{-\lambda x}, \qquad f_\lambda(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

It plays a very prominent role in probability theory and statistics, especially as a model for random times until some event, like emission of radioactive particles (Rutherford et al. 1910), or an earthquake (Gardner and Knopoff 1974), or failure of equipment (Pham 2003), or occurrence of abnormally high levels of a random process (Cramér and Leadbetter 1967), like unusually high prices (Shiryaev 1999), etc.

The characteristic "memoryless" property of the exponential distribution says that, if X is an exponential random variable, then

$$P\{X > y + x | X > y\} = P\{X > x\}, \quad \text{or}$$
$$1 - F_\lambda(x+y) = [1 - F_\lambda(x)][1 - F_\lambda(y)], \tag{1}$$

which means that the chances to wait for longer than some time x do not change, if you have been waiting already for some time y: X "does not remember" if waiting has occurred already or not. Connected to this is another characteristic property of the exponential distribution, which states that its failure rate is constant:

$$\frac{f_\lambda(x)}{1 - F_\lambda(x)} = \lambda. \tag{2}$$

Given a sample $X_1, \ldots, X_n$, denote by F_n and ν_n the empirical distribution function and the empirical process, respectively:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\}, \quad \nu_n(x) = \sqrt{n}[F_n(x) - F_\lambda(x)],$$

where $\mathbb{I}\{A\}$ denotes the indicator function of the event A. As is well known, after time transformation $t = F_\lambda(x)$, the process $\nu_n \circ F_\lambda^{-1}(t) = \nu_n\left(F_\lambda^{-1}(t)\right)$ converges in distribution to a standard Brownian bridge $u(t), t \in [0,1]$. Since in the majority of problems the value of the parameter λ is unknown, inference can not be based on ν_n but must use the parametric (or estimated) empirical process $\hat{\nu}_n$,

$$\hat{\nu}_n(x) = \nu_n(x, \hat{\lambda}_n) = \sqrt{n}[F_n(x) - F_{\hat{\lambda}_n}(x)],$$

where $\hat{\lambda}_n$ is an estimator of λ, based on the sample.

In any testing procedure one can use either of two types of statistics from $\hat{\nu}_n$, or a combination of the two: linear, or asymptotically linear, statistics and nonlinear omnibus statistics. Asymptotically linear statistics of the form

$$l_n(X_1, \ldots, X_n; F_{\hat{\lambda}_n}) = \int_0^\infty g(x) d\hat{\nu}_n(x) + o_P(1) \tag{3}$$

typically lead to asymptotically optimal tests against specific "local" (or contiguous) alternatives, but have very poor power against the huge majority of other alternatives. In contrast, nonlinear statistics like

$$\sup_x |\hat{\nu}_n(x)| \quad \text{or} \quad \int_0^\infty \hat{\nu}_n^2(x) dF_{\hat{\lambda}_n}(x),$$

which may not have best power against any given alternative, have reasonable power against more or less all alternatives. These are used in goodness of fit testing problems.

It is for these omnibus tests that the asymptotic behavior of the empirical process $\hat{\nu}_n$ is somewhat unpleasant: after time transformation $t = F_{\hat{\lambda}}(x)$ it does not converge to a standard Brownian bridge, but to a different Gaussian process with more complicated distribution. While it is true that the distribution of each omnibus statistic can in principle be calculated and tables prepared, this would involve a considerable amount of computational work. Below we show versions of ▶empirical processes that are distribution free and, moreover, the distribution of many statistics from these processes are already known.

Asymptotically Linear Statistics

There are several asymptotically linear statistics, which are widely used for testing exponentiality. Papers (Deshpande 1983) and (Bandyopadhyay and Basu 1989) are based on testing whether $1 - F_\lambda(bx) = [1 - F_\lambda(x)]^b$, and the test statistic is

$$D_n = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}\{X_j > bX_i\}.$$

A statistic known as the Gini index (or coefficient),

$$G_n = \frac{\sum_{i \neq j} |X_i - X_j|}{2n(n-1)\overline{X}}, \quad \text{with } \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

was originally designed as a measure of spread and is commonly used as a measure of inequality, e.g., see Deaton (1997). In Gail and Gastwirth (1978), and later Nikitin and Tchirina (1996), it was considered and recommended as a test of exponentiality.

The so-called Moran statistic was introduced in Moran (1951) as the score statistic for testing exponentiality against the alternative of a Gamma distribution and has the form

$$M_n = \frac{1}{n}\sum_{i=1}^{n} \log \frac{X_i}{\overline{X}}.$$

One more test of exponentiality, known as the Cox-Oakes statistic, was suggested in Cox and Oakes (1984) as the score test statistic against the alternative of a ▶Weibull distribution:

$$C_n = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{X_i}{\overline{X}}\right)\log \frac{X_i}{\overline{X}}.$$

One can show that all four statistics are asymptotically linear, e.g., see Haywood and Khmaladze (2008), and hence are asymptotically Gaussian. Somewhat surprisingly, although the kernels g of representation (3) in all four statistics look different, their correlation is extremely high, which means that all four statistics lead to the same test in practice; see Haywood and Khmaladze (2008).

Distribution Free Versions of Empirical Processes

As we noted above, unlike the empirical process ν_n, the time transformed parametric empirical process $\hat{\nu}_n \circ F_\lambda^{-1}$ does not converge to a standard Brownian bridge u. However, a beautiful observation, see Barlow and Campo (1975; Barlow and Proschan 1975), leads to another version of empirical process, which does. It is based on the "total time on test" (or TTT) notion of Epstein and Sobel (1953). Consider

$$\eta_n(x) = \frac{\int_0^x [1 - F_n(y)]dy}{\int_0^\infty [1 - F_n(y)]dy},$$

$$\text{where } \int_0^\infty [1 - F_n(y)]dy = \overline{X},\ x \geq 0.$$

If one interprets random variable X_i as a survival time (or time until failure) of the ith item on test, then $\eta_n(x)$ measures the time all items spent on test before the moment x, relative to the total time spent on test by all n items until they all failed. The process

$$\xi_n(x) = \sqrt{n}[F_n(x) - \eta_n(x)]$$

will converge in distribution to a Brownian bridge in time F_λ, and hence the time transformed empirical process $\xi_n \circ F_\lambda^{-1}$ converges in distribution to a standard Brownian bridge. To explain why this is true, cf. (Gill 1986; Khmaladze 1981), note that the process

$$B_n(x) = \sqrt{n}\left[F_n(x) - \int_0^x \frac{1 - F_n(y)}{1 - F(y)}dF(y)\right]$$

is a martingale (see ▶Martingales) with respect to the natural filtration $\{\mathcal{F}_x\}$ generated by F_n, for any i.i.d. observations. Using (2), in the case of the exponential distribution it reduces to

$$B_n(x, \lambda) = \sqrt{n}\left[F_n(x) - \lambda \int_0^x 1 - F_n(y)dy\right].$$

If we estimate the parameter λ through the equation $B_n(\infty, \lambda) = 0$, we get the usual estimator $\hat{\lambda}_n = 1/\int_0^\infty [1 - F_n(y)]dy = 1/\overline{X}_n$ and $B_n(x, \hat{\lambda}_n) = \xi_n(x)$. The process $B_n\left(F_\lambda^{-1}(t), \lambda\right)$ converges in distribution to standard Brownian motion on $[0,1]$ and hence $\xi_n \circ F_\lambda^{-1}$ converges to "tied up" Brownian motion, i.e., a standard Brownian bridge.

Another version of empirical process was investigated in Haywood and Khmaladze (2008). It has the form

$$w_n(x) = \sqrt{n}[F_n(x) - K(x, F_n)],$$

where

$$K(x, F_n) = \frac{\hat{\lambda}}{n}\sum_{i:X_i \leq x}\left(2X_i - \frac{\hat{\lambda}}{2}X_i^2\right) + \hat{\lambda}\left(2 + \frac{\hat{\lambda}}{2}x\right)x[1 - F_n(x)] - x\frac{\hat{\lambda}^2}{n}\sum_{i:X_i > x} X_i.$$

Asymptotically, the process w_n is also a martingale, but with respect to the "enriched" filtration $\{\hat{\mathcal{F}}_x\}$, where each σ-field is generated by the past of F_n and also the estimator $\hat{\lambda}_n$; $\hat{\mathcal{F}}_x = \sigma\{F_n(y), y \leq x, \overline{X}_n\}$. The idea behind this process follows from the general suggestion in Khmaladze (1981), but the form of compensator $K(x, F_n)$ for the exponential distribution is computationally particularly simple. Haywood and Khmaladze (2008) demonstrated quick convergence of the time transformed process $w_n \circ F_\lambda^{-1}$ to a standard Brownian motion (see ▶Brownian Motion and Diffusions).

Although not proved formally, the relationship between processes ξ_n and w_n is clear: the latter is asymptotically the innovation martingale for the former and therefore the two stay in one-to-one correspondence. The limit distribution of many statistics based on both $\xi_n \circ F_\lambda^{-1}$ and $w_n \circ F_\lambda^{-1}$ are well known.

Koul (1978) considered an empirical version of the memoryless property (1) of the exponential distribution and studied the empirical process

$$\alpha_n(x, y) = -\sqrt{n}\{1 - F_n(x + y) - [1 - F_n(x)][1 - F_n(y)]\}.$$

The asymptotic form of Koul's process is

$$\alpha_n(x, y) = \nu_n(x + y) - [1 - F_\lambda(x)]\nu_n(y) - [1 - F_\lambda(y)]\nu_n(x) + o_p(1)$$

and therefore, after the usual time transformation, it converges in distribution to β,

$$\beta(t,s) = u(ts) - tu(s) - su(t),$$

which is again a distribution free process in t and s. A particular form of this process,

$$\alpha_n(x) = -\sqrt{n}\left\{1 - F_n(bx) - [1 - F_n(x)]^b\right\}$$

with $b = 2$ was studied in Angus (1982) and Nikitin (1996). Note that the limit distributions of omnibus statistics from these α_n processes are not easy to obtain.

P-P Plots

It is easy and quick to calculate random variables $\hat{U}_i = 1 - F_{\hat{\lambda}_n}(X_i) = e^{-\hat{\lambda}_n X_i}$ and plot their ▶order statistics $\hat{U}_{(i:n)}$ against expected values $i/(n+1), i = 1,\ldots,n$, of the uniform order statistics. Under exponentiality the graph should be approximately linear, as $\hat{U}_i, i = 1,\ldots,n$ are almost independent and almost uniformly distributed on $[0,1]$: they would exactly have these properties if λ was known and used instead, but with $\hat{\lambda}_n$ they are not. Visual inspection of the graph is a useful preliminary tool. However, the normalized differences

$$\sqrt{n}\left[\hat{U}_{(i:n)} - \frac{i}{n+1}\right]$$

as a process in $t = i/(n+1)$, has the same drawback as the time transformed parametric empirical process $\hat{v}_n \circ F_{\hat{\lambda}}^{-1}$: distributions of many statistics from it are not known and would require extra computational effort.

Uniform Spacings

If $0 = V_{(0:n-1)} < V_{(1:n-1)} < \cdots < V_{(n-1:n-1)} < V_{(n:n-1)} = 1$ denote the uniform order statistics from a sample of size $n-1$, the differences $\Delta V_{(i-1:n-1)} = V_{(i:n-1)} - V_{(i-1:n-1)}$, form uniform spacings. Random variables

$$\frac{X_i}{\sum_{j=1}^n X_j} = \frac{X_i}{n\overline{X}},\ i = 1,\ldots,n,$$

have the same distribution as $\Delta V_{(i-1:n-1)}, i = 1,\ldots,n$, if and only if $X_1,\ldots,X_n$ are i.i.d. exponential random variables. This characteristic property was systematically used in testing problems pertaining to uniform spacings, (Pyke 1965).

Although the normalized spacings $n\Delta V_{(i-1:n-1)}$ form a distribution free statistic, they are dependent, and the empirical process based on them does not converge to a Brownian bridge. It can be shown that this empirical process is asymptotically equivalent to the process $\hat{v}_n \circ F_{\hat{\lambda}}^{-1}$.

Other approaches for testing exponentiality include tests based on functionals from the empirical characteristic function and Laplace transform, studied, e.g., in Baringhaus and Henze (1991), Epps and Pulley (1986) and Henze (1993), and on the empirical likelihood principle, e.g., Einmahl and McKeague (2003). Surveys on tests for exponentiality, including numerical studies of their relative power against fixed alternatives, can be found in Ascher (1990) and Henze and Meintanis (2005).

About the Authors

John Haywood is Senior Lecturer, School of Mathematics, Statistics and Operations Research, Victoria University.

Estate Khmaladze completed his Ph.D. in 1971 at V.A. Steklov Mathematical Institute, Moscow, under supervision of L. N. Bolshev, who was head of department of mathematical statistics at Steklov after N.V. Smirnov. He was awarded the title Professor in Probability Theory and Mathematical Statistics in 1992. Returning to Tbilisi permanently in 1990, he was appointed Head of Department of Probability Theory and Mathematical Statistics of A. Razmadze Mathematical Institute, 1990–1999, where he is still Honorary Member. He played key role in formation of Georgian Statistical Association and served as its Vice-President from 1991 to 1998. From 1991 to 1998 he served as Vice-President of Georgian Statistical Association. In 1996, Khmaladze moved to Sydney, Australia, and from there to New Zealand, where in 2002 he was appointed a Professor in Statistics at Victoria University of Wellington. School of Mathematics, Statistics and Computer Sciences. Professor Khmaladze was the first Soviet statistician to become an Associate Editor of *The Annals of Statistics* (1989–1991). Currently, he is Associate Editor for several international journals. *Mathematical Methods of Statistics*, *Statistics and Probability Letters*, *Annals of the Institute of Statistical Mathematics* and *Sankhya*. He is widely known for broadening applications of martingale methods in statistics in general and for *Khmaladze transformation*, in particular. His research interests include recently established connections of set-valued analysis and differential geometry to statistics and the statistical theory of diversity.

"This week sees Wellington inherit its very own 'beautiful mind' with the arrival at Victoria University of outstanding mathematician and statistician, Professor Estate Khmaladze. Victoria staff and students are extremely lucky to have someone of the calibre and experience of Professor Khmaladze among us" (Victoria University of Wellington, Media Release, March 25, 2002).

Cross References

▶Accelerated Lifetime Testing
▶Brownian Motion and Diffusions
▶Empirical Processes
▶Exponential Family Models

►Lorenz Curve
►Relationships Among Univariate Statistical Distributions
►Stochastic Processes: Applications in Finance and Insurance
►Survival Data

References and Further Reading

Angus JE (1982) Goodness-of-fit tests for exponentiality based on a loss-of-memory type functional equation. J Stat Plann Inference 6:241–251
Ascher S (1990) A survey of tests for exponentiality. Commun Stat 19:1811–1825
Bandyopadhyay D, Basu AP (1989) A note on tests for exponentiality by Deshpande. Biometrika 76:403–405
Baringhaus L, Henze N (1991) A class of consistent tests for exponentiality based on the empirical Laplace transform. Ann Inst Stat Math 43:179–192
Barlow RE, Campo R (1975) Total time on test processes and applications to failure data analysis. In: Barlow RE, Fussell J, Singpurwalla ND (eds) Reliability and fault tree analysis. SIAM, Philadelphia, pp 451–481
Barlow RE, Proschan F (1975) Statistical theory of reliability and life testing: probability models. Holt, Rinehart and Winston, New York
Cox DR, Oakes D (1984) Analysis of survival data. Chapman & Hall, London
Cramér H, Leadbetter MR (1967) Stationary and related stochastic processes: sample function properties and their applications. Wiley, New York
Deaton A (1997) The analysis of household surveys: a microeconometric approach to development policy. Johns Hopkins University Press, Baltimore
Deshpande JV (1983) A class of tests for exponentiality against increasing failure rate average alternatives. Biometrika 70:514–518
Einmahl JHJ, McKeague IW (2003) Empirical likelihood based hypothesis testing. Bernoulli 9:267–290
Epps TW, Pulley LB (1986) A test for exponentiality vs. monotone hazard alternatives derived from the empirical characteristic function. J R Stat Soc B 48:206–213
Epstein B, Sobel M (1953) Life testing. J Am Stat Assoc 48:486–502
Gail MH, Gastwirth JL (1978) A scale-free goodness-of-fit test for exponentiality based on the Gini statistic. J R Stat Soc B 40: 350–357
Gardner JK, Knopoff L (1974) Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? Bull Seismol Soc Am 64:1363–1367
Gill RD (1986) The total time on test plot and the cumulative total time on test statistic for a counting process. Ann Stat 14: 1234–1239
Haywood J, Khmaladze EV (2008) On distribution-free goodness-of-fit testing of exponentiality. J Econom 143:5–18
Henze N (1993) A new flexible class of omnibus tests for exponentiality. Commun Stat 22:115–133
Henze N, Meintanis SG (2005) Recent and classical tests for exponentiality: a partial review with comparisons. Metrika 61:29–45
Khmaladze EV (1981) Martingale approach in the theory of goodness of fit tests. Theory Probab Appl 26:240–257
Koul HL (1978) A class of tests for testing "new is better than used." Can J Stat 6:249–271
Moran PAP (1951) The random division of an interval – part II. J R Stat Soc B 13:147–150
Nikitin Y (1996) Bahadur efficiency of a test of exponentiality based on a loss of memory type functional equation. J Nonparametric Stat 6:13–26
Nikitin Y, Tchirina AV (1996) Bahadur efficiency and local optimality of a test for the exponential distribution based on the Gini statistic. J Ital Stat Soc 5:163–175
Pham H (ed) (2003) Handbook of reliability engineering. Springer, London
Pyke R (1965) Spacings (with discussion). J R Stat Soc B 27:395–449
Rutherford E, Geiger H, Bateman H (1910) The probability variations in the distribution of α particles. Philos Mag 20:698–707
Shiryaev AN (1999) Essentials of stochastic finance. Facts, models, theory. World Scientific, Singapore

Testing Variance Components in Mixed Linear Models

Mohamed Y. El-Bassiouni
Professor and Head
United Arab Emirates University, Al-Ain, UAE

Introduction

Consider the mixed model

$$Y = X\beta + Z\gamma + \varepsilon, \tag{1}$$

where Y is an $n \times 1$ observable random vector, X is an $n \times p$ known matrix, β is a $p \times 1$ vector of unknown parameters, Z is another known $n \times m$ matrix, γ is an $m \times 1$ unobservable random vector such that $\gamma \sim N_m(0, \theta_1 I)$, $\theta_1 \geq 0$, and ε is another unobservable $n \times 1$ random vector such that $\varepsilon \sim N_n(0, \theta_0 I)$, $\theta_0 > 0$. It is also assumed that γ and ε are independent and that $n > \text{rank}(X, Z) > \text{rank}(X)$. Therefore, we have

$$Y \sim N_n(X\beta, \theta_0 I + \theta_1 ZZ'). \tag{2}$$

Model (1) can be generalized to more than two variance components and has proven useful to practitioners in a variety of fields such as genetics, biology, psychology, and agriculture, where it is usually of interest to test the null hypothesis $\theta_1 = 0$ against the alternative $\theta_1 > 0$, or equivalently,

$$H_0 : \rho = 0, \qquad \text{vs} \qquad H_1 : \rho > 0, \tag{3}$$

where $\rho = \theta_1/\theta_0$.

Wald Test

Wald (1947) proposed an exact procedure to construct confidence intervals for ρ, which can be used to test the hypotheses in (3). This test is based on the usual ANOVA F-statistic and uses the readily available F tables to determine the critical region and that is why it has been widely used in applications. The Wald test was shown by Spjotvoll (1967) to be optimal against large alternatives. Further, unless the design is strongly unbalanced and the alternative is fairly small, El-Bassiouni and Seely (1988) showed that the Wald test has reasonable efficiency relative to the corresponding MP tests.

Seely and El-Bassiouni (1983) obtained the Wald test via reduction sums of squares. This circumvents the necessity of transforming to independent variables and/or modifying Wald's method as discussed by Spjotvoll (1968). They also give necessary and sufficient conditions under which the Wald test can be used in mixed models as well as a uniqueness property that allows one to immediately determine whether or not a proposed variance component test in a mixed model is the Wald test.

Likelihood Ratio (LR) Tests

Likelihood (LR) tests were developed by Hartley and Rao (1967) who showed that such tests are consistent and unbiased and recommended that the LR tests be carried out by comparing the observed values of the test statistics with the (approximate) cutoff points obtained from the standard χ^2 tables. However, such cutoff points can yield sizes quite different from the nominal sizes (Garbade 1977). Since the computation of the maximum likelihood estimates requires the numerical solution of a constrained nonlinear optimization problem, the LR tests have not been used much in practice. Nevertheless, Harville (1977) gives some results to facilitate the computation of LR tests. It should also be noted that even for balanced models, when a UMPU (uniformly most powerful unbiased) F-test is available, the LR approach does not necessarily yield the UMPU F-test (Herbach 1959). Using the likelihood induced by maximal location-invariant statistics leads to the restricted LR (RLR) tests. For balanced models, these RLR tests are for all practical purposes equivalent to the F-tests (El-Bassiouni 1981, 1982).

For a discussion of LR and RLR tests and their comparison with the Wald and LMPI (locally most powerful invariant) tests, the reader is referred to Li et al. (1996) and the references therein.

Uniformly Most Powerful Unbiased (UMPU) Tests

Optimal tests for certain functions of the parameters of the covariance matrix were developed by El-Bassiouni and Seely (1980), where the theory in Chap. 4 of Lehmann (1959) for determining UMPU tests in exponential families is applied to a zero mean multivariate normal family that admits a complete sufficient statistic. The special case when the matrices in the covariance structure commute was emphasized. It appears that while completeness buys similarity, it is the additional assumption of commutativity that buys simple test procedures. The case of a nonzero mean family was also discussed as were some results on the completeness of families of product measures.

In balanced models, Mathew and Sinha (1988) showed that the usual ANOVA F-test is UMPU and UMPIU (uniformly most powerful invariant unbiased), but in unbalanced models, no such UMP test exists (Spjotvoll 1967).

Similar and Location-Invariant Tests

In the context of unbalanced mixed models, if one has a specific alternative $\rho^* > 0$ in mind, a most powerful test among similar location-invariant tests, which is also MPI (most powerful among location- and scale-invariant tests), was developed by Spjotvoll (1967).

As $\rho^* \to \infty$, Spjotvoll (1967) showed that the MPI test reduces to the exact F-test of Wald (1947). On the other hand, to guard against small alternatives ($\rho^* \to 0$), the LMPI test was considered by Westfall (1988, 1989) who compared the Wald and LMPI tests in classification designs and concluded that the LMPI test is better in large designs whereas the Wald test may be preferable in small designs. Further, Westfall (1989) found that the Wald test is inferior to the LMPI test whenever there is a small proportion of relatively large group sizes. The ▶harmonic mean method was used by Thomas and Hultquist (1978) to construct confidence intervals for ρ in unbalanced random one-way models. The method was generalized to unbalanced mixed models by Harville and Fenech (1985). A modified harmonic mean procedure that compares favorably with the Wald and LMPI tests was proposed by El-Bassiouni and Seely (1996) for testing the hypotheses in (3).

For $\rho_\ell < \rho_0 < \rho_u$, Lin and Harville (1991) combined the two MPI tests against ρ_ℓ and ρ_u to obtain a two-sided test of $H_0 : \rho = \rho_0$ vs $H_1 : \rho \neq \rho_0$ and showed that their NP (Neyman–Pearson) test, although computationally intensive, can be better than the Wald test in some designs. Motivated by this idea, El-Bassiouni and Halawa (2003) proposed a test that combines the LMPI test ($\rho_\ell \to 0$) and the Wald test ($\rho_u \to \infty$) to obtain a test of $H_0 : \rho_0 = 0$ vs $H_1 : \rho > 0$. The combined test statistic is easily computed

and its null distribution may be approximated by a central F distribution with the degrees of freedom of the numerator adjusted in accordance with the degree of imbalance of the design. It is also shown to be a member of the complete class of tests of El-Bassiouni and Seely (1996). Numerical methods were used to show that the approximation is accurate over a wide range of conditions and that the efficiency of the combined test, relative to the power envelope, is satisfactorily high overall.

The combined test was also adapted to the case where $n = \text{rank}(X, Z)$ (El-Bassiouni and Charif 2004). Such models with zero degrees of freedom for error occur in many applications including plant and animal breeding and time-varying regression coefficients.

About the Author

Dr. Mohamed Yahia El-Bassiouni is a Professor and Head, Department of Statistics, UAE University. He obtained his Ph.D. degree in Statistics from Oregon State University in 1977 and has received the Legion of Science, First Degree Honor, from the Egyptian President in 1984, in honor of his research. He is also the Editor of the *Journal of Economic and Administrative Sciences*, UAE University, and the Regional Editor of the I*nternational Journal of Management and Sustainable Development*, United Kingdom. Professor El-Bassiouni has authored and coauthored more than 70 papers, 15 monographs, and 30 research reports, many of which are on testing and interval estimation of variance components.

Cross References

►Best Linear Unbiased Estimation in Linear Models
►Cross Classified and Multiple Membership Multilevel Models
►General Linear Models
►Linear Mixed Models
►Multilevel Analysis
►Panel Data

References and Further Reading

El-Bassiouni MY (1981) Likelihood ratio tests for covariance hypotheses generating commutative quadratic subspaces. Commun Stat A10(23):2461–2468
El-Bassiouni MY (1982) On a theorem of Graybill. Commun Stat A11(13):1519–1522
El-Bassiouni MY, Charif HA (2004) Testing a null variance ratio in mixed models with zero degrees of freedom for error. Comput Stat Data Anal 46:707–719
El-Bassiouni MY, Halawa AM (2003) A combined invariant test for a null variance ratio. Biom J 45:249–260
El-Bassiouni MY, Seely J (1980) Optimal tests for certain functions of the parameters in a covariance matrix with linear structure. Sankhya 42:64–77
El-Bassiouni MY, Seely JF (1988) On the power of Wald's variance component test in the unbalanced random one-way model. In: Dodge Y, Federov VV, Wynn HP (eds) Optimum design and analysis of experiments. Elsevier, North-Holland, pp 157–165
El-Bassiouni MY, Seely JF (1996) A modified harmonic mean test procedure for variance components. J Stat Plan Infer 49:319–326
Garbade K (1977) Two methods for examining the stability of regression coefficients. J Am Stat Assoc 72:54–63
Hartley HO, Rao JNK (1967) Maximum-Likelihood estimation for the mixed analysis of variance model. Biometrika 54:93–108
Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72:320–338
Harville DA, Fenech AP (1985) Confidence intervals for a variance ratio, or for heritability, in an unbalanced mixed linear model. Biometrics 41:137–152
Herbach LH (1959) Properties of model II-type analysis of variance tests, A: Optimum nature of the F-test for model II in the balanced case. Ann Math Stat 30:939–959
Lehmann EL (1959) Testing statistical hypotheses. Wiley, New York
Li Y, Birkes D, Thomas DR (1996) The residual likelihood ratio test for the variance ratio in a linear model with two variance components. Biom J 38:961–972
Lin TH, Harville DA (1991) Some alternatives to Wald's confidence interval and test. J Am Stat Assoc 86:179–187
Mathew T, Sinha BK (1988) Optimum tests for fixed effects and variance components in balanced models. J Am Stat Assoc 83:133–135
Seely JF, El-Bassiouni MY (1983) Applying Wald's variance component test. Ann Stat 11:197–201
Spjotvoll E (1967) Optimum invariant tests in unbalanced variance component models. Ann Math Stat 38:422–429
Spjotvoll E (1968) Confidence intervals and tests for variance ratios in unbalanced variance components models. Rev Int Stat Inst 36:37–42
Thomas JD, Hultquist RA (1978) Interval estimation for the unbalanced case of the one-way random effects model. Ann Stat 6:582–587
Wald A (1947) A note on regression analysis. Ann Math Stat 18: 586–589
Westfall PH (1988) Robustness and power of tests for a null variance ratio. Biometrika 75:207–214
Westfall PH (1989) Power comparisons for invariant variance ratio tests in mixed ANOVA models. Ann Stat 17:318–326

Tests for Discriminating Separate or Non-Nested Models

Basilio de Bragança Pereira
Professor of Biostatistics at the school of Medicine
Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Introduction

The Neyman–Pearson theory of hypothesis testing applies if the models belong to the same family of distributions. Alternatively, special procedures are needed if the models

belong to families that are separate or non-nested, in the sense that an arbitrary member of one family cannot be obtained as a limit of members of the other.

Let $y = (y_1, \ldots, y_n)$ be independent observations from some unknown distribution. Suppose that there are null and alternative hypotheses H_f and H_g specifying parametric densities $f(y, \alpha)$ and $g(y, \beta)$ for the random vector y. Hence α and β are unknown vector parameters and it is assumed that the families are separate.

The asymptotic tests developed by Cox (1961, 1962) were based on a modification of the Neyman–Pearson maximum likelihood ratio. If H_f is the null hypothesis and H_g the alternative hypothesis, the test statistics considered was

$$T_{fg} = T_{fg}(\hat{\alpha}, \hat{\beta}) - E_{\hat{\alpha}}\{T_{fg}(\alpha, \beta_\alpha)\}$$

where for a random sample of size n, $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimators of α and β respectively, $T_{fg}(\alpha, \beta) = 1_f(\alpha) - 1_g(\beta)$ is the log likelihood ratio, β_α is the probability limit, as $n \to \infty$, of $\hat{\beta}$ under H_f and the subscript α means that expectations, etc. are calculated under H_f.

Cox showed that, asymptotically, under the alternative hypothesis T_{fg} has a negative mean and that under the null hypothesis T_{fg} is normally distributed with mean zero and variance

$$V_\alpha(T_{fg}) = V_\alpha\{IT_{fg}(\alpha, \beta_\alpha)\} - C_\alpha^{-1} I^{-1} C_\alpha$$

where $C_\alpha = \partial E_\alpha\{T_{fg}(\alpha, \beta_\alpha)\}/\partial\alpha$, and I_α the information matrix of α. When H_g is the null hypothesis and H_f is the alternative hypothesis analogous results are obtained for a statistics T_{gf}. Therefore $T^*_{fg} = T_{fg}\{V_\alpha(T_{fg})\}^{-1/2}$ and $T^*_{gf} = T_{gf}\{V_\beta(T_{gf})\}^{-1/2}$ under H_f and H_g respectively can be considered as approximately standard normal variables and two-tailed tests can be performed. The outcomes of application of both tests are shown in the Table 1.

As an alternative to his test, Cox (1961) suggested combining the two models in a general model of which they would be both special cases. The density could be proportional to the exponential mixture

$$\{f(y, \alpha)\}^\lambda \{g(y, \beta)\}^{1-\lambda}$$

and inferences made about λ. It should be notice that these mixtures can be generalized for testing more than two models. In particular, the exponential mixture is the base of the tests developed in econometrics.

Cox also suggested a Bayesian approach and gives a general expression when losses are associated and a large sample approximation.

The posterior odds for H_f versus H_g is

$$\frac{\pi_f \int f(y; \alpha)\pi_f(\alpha)d\alpha}{\pi_g \int g(y; \beta)\pi_g(\beta)d\beta} = \frac{\pi_f}{\pi_g} B_{fg}(y)$$

where π_f and π_g are the prior probabilities of H_f and H_g respectively, $\pi_f(\alpha)$ and $\pi_g(\beta)$ are the prior probabilities for the parameters conditionally on H_f and H_g. $B_{fg}(y)$ is the Bayes Factor and represents the weight of evidence in the data for H_f over H_g.

One difficulty with this approach lies in the fact that the prior knowledge expressed by π_f and $\pi_f(\alpha)$ must be coherent with that of π_g and $\pi_g(\beta)$. If the parameter spaces have different dimensions and there is no simple relation between the parameters, the problem is not simple. When prior information is weak and improper prior is used there are also difficulties and paradox with the use of Bayes factors which is unspecified.

Alternative Approaches

Alternative approaches present in Cox (1961) were further developed under Cox supervision in unpublished Ph.D. thesis at Imperial College : O.A.Y. Jackson in 1968 and B. de B. Pereira in 1976 obtained further results on the modified likelihood ratio, A.C. Atkinson in 1970 developed the exponential compound model approach, J. K. Lindsey in 1972 used a direct relative likelihood approach. Later in 1980 A. C. Atkinson supervised L. R. Pericchi on the Bayesian approach. Published references from this work can be traced in Pereira (1977a, b). Further contributions of Cox in this area are Cox (1974), Cox and Brandwood (1959), Atkinson and Cox (1974), Chambers and Cox (1967).

Further alternative approaches such as: linear mixtures, relative likelihoods, tests based on information and divergence measures, ▶moment generating functions, multiple combinations, methods based on invariant statistics and method of moments and bootstrap are reviewed in Pereira (1998, 2005).

A huge amount of research on separate families of hypothesis was developed since the fundamental work of Cox (1961, 1962). In the 1980s econometricians, using the exponential compound model took a great interest in the subject. Bayesian statisticians in the 1990s developed alternative Bayes factors (see Araújo and Pereira 2007). For reviews and references see McAller et al. (1990), Gourieroux and Monfort (1994), McAller (1995), Pereira (1977b, 1981a, 1998, 2005), and Pesaran and Weeks (2001).

A test based on descriptive statistics for the mean and the variance of the log-likelihood ratio has been proposed by Vuong (1989) but this has not been compared with Cox test that has been shown to be consistent (Pereira 1977a)

Tests for Discriminating Separate or Non-Nested Models. Table 1 Possible results of Cox test

	T_{fg}		
T_{gf}	Significant negative	Not significant	Significant positive
Significant negative	Reject both	Accept H_f	Reject both
Not significant	Accept H_g	Accept both	Possible acceptance H_g
Significant positive	Reject both	Possible acceptance H_f	Reject both

and the only that can be extend to multivariate problems (Araújo et al. 2005; Timm and Al-Subaihi 2001) and that approaches the normal asymptotic result faster (Pereira 1978).

About the Author

Dr. Basilio de Bragança Pereira obtained his Ph.D. and D.I.C. from the Imperial College of Science, Technology and Medicine (1976), supervised by Sir David Cox. In 2003 he spent a year working on a Project on Neural Networks in Statistics with Professor C.R. Rao at Penn State University on a postdoctoral grant from the Brazilian Government (CAPES). He was Associate Professor at the Institute of Mathematics (1970–1989 and 1994–1997), and Professor Titular of Applied Statistics at COPPE (1989–1994, retired). Currently, he is Professor Titular of Biostatistics at the School of Medicine (since 1998) and the coordinator of the Statistical research consulting group at The University Hospital of UFRJ. He has supervised 19 PhD and 38 MSc students. He is an Elected member of the International Statistical Institute. Professor Pereira has coauthored over 70 refereed papers, and a monograph *Data Mining with Neural Networks: A Guide for Statisticians* (with C.R. Rao, available for download in TextBook Revolution).

Cross References

►Bayesian Statistics
►Econometrics
►Mixture Models
►Neyman-Pearson Lemma
►Significance Testing: An Overview

References and Further Reading

Araújo MI, Pereira BB (2007) Comparison among Bayes factors for separate models: some simulation results. Commun Stat – Simul Comput 36:297–309

Araújo MI, Fernandes M, de B Pereira B (2005) Alternative procedures to discriminate non nested multivariate linear regression models. Commun Stat – Theory Methods 34:2047–2062

Atkinson AC, Cox DR (1974) Planning experiments for discriminating between models (with discussion). J R Stat Soc B 36:321–348

Chambers EA, Cox DR (1967) Discriminating between alternative binary response models. Biometrica 54:573–578

Cox DR (1961) Tests of separate families of hypotheses. In: Proceedings of fourth Berkeley symposium, vol 1, pp 105–123

Cox DR (1962) Further results on tests of separate families of hypotheses. J R Stat Soc B 24:406–423

Cox DR (1974) Discussion of "Dempster, A.P. pg 353 – The direct use of likelihood for significance testing". In: Proceedings of conference on foundational questions in statistical inference, Aarhus. Memoirs no. 1. University of Aarhus, Aarhus

Cox DR, Brandwood L (1959) On a discriminatory problem connected with the works of Plato. J R Stat Soc B 21:195–200

Davidson R, MacKinnon JD (1983) Testing the specification of multivariate models in the presence of alternative hypotheses. J Econom 23:301–313

de B Pereira B (1977a) A note on the consistency and on finite sample comparisons of some tests of separate families of hypotheses. Biometrika 64:109–113 (correction in volume 64, page 655)

de B Pereira B (1977b) Discriminating among separate models: a bibliography. Int Stat Rev 45:163–172

de B Pereira B (1978) Empirical comparisons of some tests of separate families of hypotheses. Metrika 25:219–234

de B Pereira B (1981) Discriminating among separate models: an additionalbibliography. Int Stat Inf 62(2):3 (repr Katti SK (1982) On the preliminary test for the CEAS model versus the Thompson model for predicting soybean production. Technical Report 125, Department of Statistics, University of Missouri – Columbia)

de B Pereira B (1998) Separate families of hypotheses. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 5. Wiley, pp 4069–4074

de B Pereira B (2005) Separate families of hypotheses. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics, vol 7, 2nd edn. Wiley, pp 4881–4886

Gourieroux C, Monfort A (1994) Testing non-nested hypotheses. In: Engle R, McFadden DL (eds) Handbook of econometrics, vol IV (Chapter 44). Elsevier, London, pp 2585–2637

McAller M (1995) The significance of testing empirical non-nested models. J Econom 65:149–171

McAller M, Pesaran MH, Bera AK (1990) Alternative approaches to testing non-nested models with autocorrelated disturbances. Commun Stat – Theory Methods 19:3619–3644

Pesaran MH (1982) On the comprehensive method for testing non-nested regression models. J Econom 18:263–274

Pesaran MH, Deaton AS (1978) Testing non-nested regression models. Econometrica 46:677–694

Pesaran MH, Weeks M (2001) Non-nested hypothesis testing: an overview. In: Baltagi BH (ed) Companion to theoretical econometrics. Basil Blackwell, Oxford

Timm NH, Al-Subaihi AA (2001) Testing model specification in seemingly unrelated regression models. Commun Stat – Theory Methods 30:577–590

Uloa RD, Pesaran MH (2008) Non-nested hypotheses. In: Durlauf SN, Blume LE (eds) The new Palgrave dictionary of economics, 2nd edn. Palgrave Macmillan

Vuong QH (1989) Likelihood ratio tests for model selection and non-nested hypothesis. Econometrica 57:307–333

Tests for Homogeneity of Variance

Nataša Erjavec
Professor, Faculty of Economics
University of Zagreb, Zagreb, Croatia

Introduction

Homogeneity of variance (*homoscedasticity*) is an important assumption shared by many parametric statistical methods. This assumption requires that the variance within each population be equal for all populations (two or more, depending on the method). For example, this assumption is used in the two-sample *t*-test and ANOVA. If the variances are not homogeneous, they are said to be *heterogeneous*. If this is the case, we say that the underlying populations, or random variables, are *heteroscedastic* (sometimes spelled as heteroskedastic).

In this entry we will initially discuss the case when we compare variances of two populations, and subsequently will extend to k populations.

Comparison of Two Population Variances

The standard F-test is used to test whether two populations have the same variance. The test statistic for testing the hypothesis if ${\sigma_1}^2 = {\sigma_2}^2$ where ${\sigma_1}^2$ and ${\sigma_2}^2$ are the variances of two populations, is

$$F = \frac{{s_1}^2}{{s_2}^2}, \tag{1}$$

where ${s_1}^2$ and s_2 2 are the sample variances for two independent random samples of n_1 and n_2 observations from normally distributed populations with variances ${\sigma_1}^2$ and ${\sigma_2}^2$, respectively. If the null hypothesis is true (i.e., H_0 : ${\sigma_1}^2 = {\sigma_2}^2$), the test statistic has the F-distribution with $(n_1 - 1)$ degrees of freedom for the numerator and $(n_2 - 1)$ degrees of freedom for the denominator. The F-test is extremely sensitive to non-normality and should not be used unless there is strong evidence that the data do not depart from normality.

In practical applications, the F ratio in (1) is usually calculated so that the larger sample variance is in the numerator, that is, ${s_1}^2 > {s_2}^2$. Thus, F statistic is always greater than one and only the upper critical values of the F-distribution are used. At the significance level α, the test rejects the hypothesis that the variances are equal if $F > F_{(\alpha;n_1-1;n_2-1)}$, where $F_{(\alpha;n_1-1;n_2-1)}$ is the upper critical value of the F distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom.

Tests for Equality of Variances of *k* Populations

The **Bartlett's test** (Bartlett 1937) is used to test if k-groups (populations) have equal variances. Hypotheses are stated as follows:

$$H_0 : {\sigma_1}^2 = {\sigma_1}^2 = \ldots = {\sigma_k}^2$$

$$H_1 : {\sigma_i}^2 \neq {\sigma_j}^2 \quad \text{for at least one pair } (i,j).$$

To test for equality of variance against the alternative that variances are not equal for at least two groups, the test statistic is defined as

$$\chi^2 = \frac{(N-k)\ln\left(\dfrac{\sum_{i=1}^{k}(n_i-1){s_i}^2}{N-k}\right) - \sum_{i=1}^{k}(n_i-1)\ln\left({s_i}^2\right)}{1 + \dfrac{1}{3(k-1)}\left[\left(\sum_{i=1}^{k}\dfrac{1}{n_i-1}\right) - \dfrac{1}{N-k}\right]} \tag{2}$$

where k is the number of samples (groups), n_i is the size of the ith sample with sample variance ${s_i}^2$, and N is the sum of all samples sizes.

The test statistic follows a ▶chi-square distribution with $(k-1)$ degrees of freedom and the standard *chi-squared test* with $(k-1)$ degrees of freedom is applied.

The Bartlett's test rejects the null hypothesis that the variances are equal if $\chi^2 > \chi^2_{(\alpha,k-1)}$, where $\chi^2_{(\alpha,k-1)}$ is the upper critical value of the chi-square distribution with $(k-1)$ degrees of freedom and a significance level of α.

The test is very sensitive to departures from normality and/or to differences in group sizes and is not recommended for routine use. However, if there is strong evidence that the underlying distribution is normal (or nearly normal), the Bartlett's test has good performance.

The Levene's test (Levene 1960) is another test used to test if k groups have equal variances, as an alternative to

the Bartlett's test. It is less sensitive to departures from normality and/or to differences in group sizes and is considered to be the standard test for homogeneity of variances. The idea of this test is to transform the original values of the dependent variable Y and obtain a new variable known as the "dispersion variable." A standard ►analysis of variance based on these transformed values will test the assumption of homogeneity of variances.

The test has two options. Given a variable Y with sample of size N divided into k-subgroups, Y_{ij} will be the jth individual score belonging to the ith subgroup. The first option of the test is to define the transformed variable as the absolute deviation of the individual's score from the mean of the subgroup to which the individual belongs, that is, as $Z_{ij} = \left|Y_{ij} - \overline{Y}_{i.}\right|$ where $\overline{Y}_{i.}$ is the mean of the ith subgroup. The transformed variable is known as the dispersion variable, since it "measures" how far the individual is displaced from its subgroup mean.

The Levene's test statistic is defined as

$$F^L = \frac{(N-k)\sum_{i=1}^{k} n_i\left(\overline{Z}_{i.} - \overline{Z}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Z_{ij} - \overline{Z}_{i.}\right)^2} \tag{3}$$

where n_i is the sample size of the ith subgroup, $Z_{ij} = \left|Y_{ij} - \overline{Y}_{i.}\right|$ is the dispersion variable, $\overline{Z}_{i.}$ are the subgroup means of Z_{ij} and $\overline{Z}$ is the overall mean of Z_{ij}.

The test statistic follows the F-distribution with $(k-1)$ and $(N-k)$ degrees of freedom and the standard F-test is applied.

The Levene's test will reject the hypothesis that the variances are equal if $F^L > F^{\alpha}_{(k-1,N-k)}$ where $F^{\alpha}_{(k-1,N-k)}$ is the upper critical value of the F distribution with $(k-1)$ and $(N-k)$ degrees of freedom at the significance level α.

The second option is to define the dispersion variable as the square of the absolute deviation from the subgroup mean, that is, as $Z_{ij}^2 = \left|Y_{ij} - \overline{Y}_{i.}\right|^2$.

The Brown–Forsythe test (Brown and Forsythe 1974) is a modification of the Levene's test, based on the same logic, except that the dispersion variable Z_{ij} is defined as the absolute deviation from the subgroup median rather than the subgroup mean, that is, $Z_{ij} = |Y_{ij} - M_{i.}|$, where $M_{i.}$ is the median of the ith subgroup. Such a definition, based on medians instead of means, provides good robustness against many types of non-normal data while retaining good power, and is therefore recommended in practical applications.

The O'Brien test (O'Brien 1979) is a modification of the Levene's Z_{ij}^2 test. In the O'Brien test, the dispersion variable Z_{ij}^2 is modified in a way to include an additional scalar W (weight) to account for the suspected kurtosis of the underlying distribution. The dispersion variable in the O'Brien test is defined as

$$Z_{ij}^B = \frac{(W + n_i - 2)\, n_i Z_{ij}^2 - W\,(n_i - 1)\, s_i^2}{(n_i - 1)\,(n_i - 2)} \tag{4}$$

where Z_{ij}^2 is the square of the absolute deviation from the subgroup mean and n_i is the size of the ith subgroup with its sample variance s_i^2. W is a constant with values between 0 and 1 and is used to adjust the transformation. The most commonly used weight is $W = 0.5$, as suggested by O'Brien (1979).

The previously discussed tests are the tests that are mostly used in empirical research and easily available in most statistical software packages. However, there are also other homogeneity of variance tests, both parametric and nonparametric. Among them are Hartley's Fmax test, David's multiple test, and Cochran's G test (also known as Cochran's C test). The Bartlett–Kendall test (like Bartlett's test) uses log transformation of the variance to approximate the normal distribution. An example of a nonparametric test is the Sidney–Tukey test that uses ranks and the chi-square approximation. A good discussion on the topic can be found in Zhang (1998).

Cross References

►Analysis of Variance Model, Effects of Departures from Assumptions Underlying
►Bartlett's Test
►Heteroscedasticity
►Variance

References and Further Reading

Bartlett MS (1937) Properties of sufficiency and statistical tests. Proc R Soc Lond A 160:268–282

Brown MB, Forsythe AB (1974) Robust test for equality of variances. J Am Stat Assoc 69:364–367

Levene H (1960) Robust tests for the equality of variance. In: Olkin I (ed) Contributions to probability and statistics. Stanford University Press, Paolo Alto, pp 278–292

O'Brien RG (1979) A general ANOVA method for robust tests of additive models for variances. J Am Stat Assoc 74:877–880

Zhang S (1998) Fourteen homogeneity of variance tests: when and how to use them. Paper presented at the annual meeting of the american educational research association, San Diego, California

Tests of Fit Based on The Empirical Distribution Function

Michael A. Stephens
Professor Emeritus
Simon Fraser University, Burnaby, BC, Canada

Introduction: Tests for Continuous Distributions

Suppose a random sample $x_1, x_2, \cdots, x_n$ is given and we wish to test H_0: the parent population is the (continuous) distribution $F(x;\theta)$, where θ is a vector of parameters. The empirical distribution function (EDF) of the sample is defined by

$$F_n(x) = n(x)/n,$$

where $n(x)$ is the number of x_i which are less than or equal to x. The goodness-of-fit tests to be discussed are EDF tests, that is, based on the discrepancy

$$Y(x) = F_n(x) - F(x;\theta)$$

The most well known are the Kolmogorov-Smirnov family:

$$D_n^+ = sup\ Y(x)$$
$$D_n^- = sup\{-Y(x)\}$$
$$D_n = sup|Y(x)|$$

and the Cramér-von Mises family:

$$W_n^2 = n\int_{-\infty}^{\infty} Y^2(x)dF(x;\theta)$$

$$U_n^2 = n\int_{-\infty}^{\infty}\left\{Y(x) - \int_{-\infty}^{\infty} Y(x)dF(x;\theta)\right\}^2 dF(x;\theta)$$

and

$$A_n^2 = n\int_{-\infty}^{\infty} Y^2(x)\psi(x)dF(x;\theta)$$

where $\psi(x) = [F(x;\theta)(1-F(x;\theta))]^{-1}$

Statistic W_n^2 is the original Cramér-von Mises statistic, originally called $n\omega^2$. Statistic U_n^2 was introduced by Watson (1961) for testing distributions around a circle; it has the merit that its value does not depend on the origin used for measuring the observations. Statistic A_n^2 is the Anderson-Darling (1952) statistic: it emphasises the tails of the tested distribution.

Statistic D_n was introduced by Kolmogorov (1933). Distribution theory for the Kolmogorov–Smirnov family is known for the case when parameters are known; but when parameters are unknown and must be estimated from the sample, even asymptotic theory is not available and significance points must be obtained by Monte Carlo. Tables of significance points for testing a number of distributions are in Stephens (1986).

The statistics D_n^+ and D_n^- have good power when the sample EDF lies mostly on one side of the tested distribution, but the D_n statistic, in general, is less powerful as an omnibus test than the Cramér-von Mises statistics. More information on this statistic is given by Lopes (2010) in an article in this Encyclopedia and here it will not be considered further.

The Probability Integral Transformation

In practice, it is easier to work with the EDF of the transformed set $z_{(i)} = F(x_{(i)};\theta), i = 1,\cdots,n$; where $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ is the ordered sample. This transformation is called the probability integral transformation (PIT). If θ is known, the $z_{(i)}$ are ordered uniform variates. If θ is not known, an efficient estimate (for example, the MLE) should be used for the transformation. The EDF statistics are easier to calculate from the z-values, as follows.

Let $F_n(z)$ be the empirical distribution function of the z-values, and define

$$y_n(z) = \sqrt{n}\{F_n(z) - z\}.$$

The Cramér-von Mises statistics now become, in terms of $y_n(z)$:

$$W_n^2 = \int_0^1 \{y_n(z)\}^2 dz, \quad (1)$$

$$U_n^2 = \int_0^1 \{y_n(z) - \bar{y}\}^2 dz, \quad (2)$$

$$A_n^2 = \int_0^1 \{y_n(z)\}^2 w(z)dz, \quad (3)$$

where

$$\bar{y} = \int_0^1 y_n(z)dz \qquad \text{and} \qquad w(z) = 1/(z-z^2).$$

The computing formulas are

$$W_n^2 = \sum\left\{z_{(i)} - (2i-1)/2n\right\}^2 + 1/(12n) \quad (4)$$

$$U_n^2 = W^2 - n(\bar{z} - 0.5)^2 \quad (5)$$

and

$$A_n^2 = -n - (1/n)\sum(2i-1)\left\{\ln(z_{(i)}) + \ln(1 - z_{(n+1-i)})\right\}. \quad (6)$$

The distributions of these statistics, when estimated parameters are location or scale, will depend on the tested distribution, but not on the true values of the parameters. However, when an unknown parameter is a shape parameter, the distribution will depend on the shape.

Asymptotic theory of these statistics was first given by Anderson and Darling (1952), and Darling (1955); see also Anderson (2010), an entry in this Encyclopedia. Stephens (1976) used the theory to give significance points for tests of normality and exponentiality; points for other distributions are in Stephens (1986) and in Lockhart and Stephens (1985, 1994).

In general, W^2 and A^2 have been shown to be powerful in testing many distributions; A^2 has comparable power to the Shapiro-Wilk statistic for testing normality.

Tests for Discrete Distributions

EDF tests may be adapted for testing discrete distributions, by comparing the cumulated histogram of observed numbers in the cells with the cumulated histogram of the expected numbers. Choulakian et al. (1994) have given distribution theory for the Cramér-von Mises family when parameters are known, and Lockhart et al. (2007) have discussed the case when parameters must be estimated from the sample; see Stephens (2010), an entry in this Encyclopedia. These statistics are generally more powerful than Pearson's χ^2 statistic.

About the Author

For biography *see* the entry ►Cramér-Von Mises Statistics for Discrete Distributions.

Cross References

►Anderson-Darling Tests of Goodness-of-Fit
►Cramér-Von Mises Statistics for Discrete Distributions
►Exact Goodness-of-Fit Tests Based on Sufficiency
►Kolmogorov-Smirnov Test
►Normality Tests
►Parametric and Nonparametric Reliability Analysis

References and Further Reading

Anderson TW (2010) Anderson–Darling tests of goodness-of-fit. In: International encyclopedia of statistical science, Springer-Verlag, Berlin

Anderson TW, Darling DA (1952) Asymptotic theory of certain goodness of fit criteria based on stochastic processes. Ann Math Stat 23:193–212

Choulakian V, Lockhart RA, Stephens MA (1994) Cramer-von Mises tests for discrete distributions. Can J Stat 22:125–137

Darling DA (1955) The Cramér-Smirnov test in the parametric case. Ann Math Stat 26:1–20

Kolmogorov AN (1933) Sulla determinazione empirica di una legge di dis-tribuzione. Giorna Ist Attuari 4:83–91

Lockhart RA, Spinelli JJ, Stephens MA (2007) Cramér-von Mises statistics for discrete distributions with unknown parameters. Can J Stat 35:125–133(9)

Lockhart RA, Stephens MA (1985) Tests of fit for the Von Mises distribution. Biometrika 72:647–652

Lockhart RA, Stephens MA (1994) Estimation and tests of fit for the three-parameter Weibull distribution. J Roy Stat Soc B 56: 491–500

Lopes RHC (2010) Kolmogorov-Smirnov test. International Encyclopedia of Statistical Science, Springer-Verlag, Berlin

Stephens MA (1976) Asymptotic results for goodness-of-fit statistics with unknown parameters. Ann Stat 4:357–369

Stephens MA (1986) Tests based on EDF statistics, Chap 4. In: D'Agostino R, Stephens MA (eds) Goodness-of-fit techniques. Marcel Dekker, New York

Stephens MA (2010) EDF tests of fit. International Encyclopedia of Statistical Science, Springer-Verlag, Berlin

Watson GS (1961) Goodness-of-fit tests on a circle, 1. Biometrika 48:109–114

Tests of Independence

Bruno Rémillard
Professor
HEC Montréal, Montréal, QC, Canada

Testing for Interdependence

Testing independence between two of more components of a random vector is an important problem in statistics. For sake of simplicity, suppose that the law of each component is continuous. In the bivariate case, for testing independence between random variables X_1 and X_2, most of the tests proposed initially were based on some dependence measure ρ, taking usually value 0 under the null hypothesis of independence. Once a random sample $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$ is collected, that is, the pairs (X_{i1}, X_{i2}), $i = 1, \ldots, n$, are independent observations of (X_1, X_2), an estimator $\hat{\rho}_n$ of ρ is obtained and it is compared with the value of ρ under the null hypothesis. In general, $\hat{\rho}_n$ must be a "good" estimator of ρ in the sense that as $n \to \infty$, $n^{1/2}(\hat{\rho}_n - \rho) \rightsquigarrow N(0, \sigma_0^2)$, where "$\rightsquigarrow$" denotes convergence in law, and σ_0 is the limiting variance of $n^{1/2}\hat{\rho}_n$. The most known example is the one based on the Pearson correlation coefficient, defined by

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X, Y)\mathrm{Var}(X, Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\{E(X^2) - E^2(X)\}\{E(Y^2) - E^2(Y)\}}},$$

provided $E(X^2)$ and $E(Y^2)$ are finite. In that case,

$$\hat{\rho}_n = r_n = \frac{\sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\sqrt{\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2}\sqrt{\sum_{i=1}^n (X_{i2} - \bar{X}_2)^2}}.$$

Under the null hypothesis of independence, $\rho = 0$ and $n^{1/2} r_n \rightsquigarrow N(0,1)$, as $n \to \infty$. If in addition the joint distribution of (X_1, X_2) is Gaussian, then $\frac{r_n}{\sqrt{(1-r_n^2)/(n-2)}}$ has a Student distribution with $n-2$ degrees of freedom.

Many other popular empirical measures of dependence are based on ranks. Recall that the ranks R_{ij}, $i = 1, \ldots, n$, $j = 1, 2$, are defined as follows: R_{i1} is the rank of X_{i1} amongst $X_{11}, \ldots, X_{n1}$, while R_{i2} is the rank of X_{i2} amongst $X_{12}, \ldots, X_{n2}$, and so on, where the smallest observation has rank 1. In particular, these measures do not depend on the margins, only on the so-called copula (see ▶Copulas). That notion will be defined later. The most known rank-based measures of dependence are ▶Kendall's tau and Spearman's rho. Kendall's tau is defined by

$$\tau_n = \frac{2}{n(n-1)}(C_n - D_n),$$

where C_n is the number of concordant pairs of ranks, and D_n is the number of discordant pairs, the pairs (R_{i1}, R_{i2}) and (R_{j1}, R_{j2}) being concordant if $(R_{i1} - R_{j1})(R_{i2} - R_{j2}) > 0$ and discordant otherwise. Recall that τ_n is an estimation of $\tau = 2P\{(X_1 - Y_1)(X_2 - Y_2) > 0\} - 1$, where (Y_1, Y_2) is an independent copy of (X_1, X_2). Under the null hypothesis of independence, $\tau = 0$ and it can be shown that $n^{1/2}\tau_n \rightsquigarrow N(0, 4/9)$, as $n \to \infty$.

Spearman's rho, denoted by ρ_n^S, is simply defined as the correlation between the ranks $(R_{11}, R_{12}), \ldots, (R_{n1}, R_{n2})$. Then ρ_n^S is an estimator of ρ^S, the correlation between $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$, where F_j is the distribution function of X_j, $j = 1, 2$. Under the null hypothesis of independence, $\rho^S = 0$ and $n^{1/2}\rho_n^S \rightsquigarrow N(0,1)$, as $n \to \infty$.

All tests based on a single measure of dependence usually have the same weakness: They are not consistent for testing independence in the sense that under some alternatives, the power of the test does not tend to 1 as the sample size tends to infinity. One such example of alternative is the following: Let X_1 be uniformly distributed over $(0,1)$, denoted by $X_1 \sim \text{Unif}(0,1)$ and set $X_2 = T(X_1)$, where T is the tent map, i.e., $T(u) = 2\min(u, 1-u)$. Then $X_2 \sim \text{Unif}(0,1)$ and X_1 and X_2 are strongly dependent. However, for any of the three measures of dependence ρ stated previously, the value of ρ is 0, the same value as for independence, and it can be shown than $n^{1/2}\hat{\rho}_n \rightsquigarrow N(0, \sigma^2)$, for some $\sigma > 0$ depending on ρ. As a result, the power of the associated test of level 5% tends to $2\Phi\left(-1.96\frac{\sigma_0}{\sigma}\right)$, where σ_0^2 is the asymptotic variance under the null hypothesis of independence and Φ is the distribution function of the standard Gaussian. For example, in the case of the Pearson correlation, $\sigma_0^2 = 1$ and $\sigma^2 = 6/5$, so the power tends to 0.1024, as $n \to \infty$.

To overcome the inconsistency problem, it was suggested in Blum et al. (1961) to use statistics based on the empirical distribution function. More precisely, in the bivariate case, one can compare the joint empirical distribution function H_n, given by

$$H_n(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_{i1} \le x_1, X_{i2} \le x_2)$$

with the product of its margins, i.e., $F_{n1}(x_1) = H_n(x_1, \infty)$ and $F_{n2}(x_2) = H_n(\infty, x_2)$. It can then be shown that $\mathbb{H}_n(x_1, x_2) = n^{1/2}\{H_n(x_1, x_2) - F_{n1}(x_1)F_{n2}(x_2)\} \rightsquigarrow \mathbb{H}(x_1, x_2)$, where the convergence is in the Skorohod space $\mathcal{D}\left([-\infty, +\infty]^2\right)$ and $\mathbb{H}(x_1, x_2) = \mathbb{B}\{F_1(x_1), F_2(x_2)\}$, where F_1 and F_2 are the margins of the joint distribution function H of (X_1, X_2), and $\mathbb{B}$ is a continuous centered Gaussian process with covariance function

$$\begin{aligned}\Gamma(u_1, u_2, v_1, v_2) &= \text{Cov}\{\mathbb{B}(u_1, u_2), \mathbb{B}(v_1, v_2)\} \\ &= \{\min(u_1, v_1) - u_1 v_1\} \\ &\quad \times \{\min(u_2, v_2) - u_2 v_2\}.\end{aligned}$$

Recall that by Sklar (1959), when the marginal distributions are continuous, there exists a unique distribution function C with uniform margins, called a copula, so that

$$\begin{aligned}H(x_1, x_2) &= P(X_1 \le x_1, X_2 \le x_2) \\ &= C\{F_1(x_1), F_2(x_2)\}, \qquad x_1, x_2 \in \mathbb{R}.\end{aligned}$$

Thus X_1 and X_2 are independent if and only if the copula is the independence copula $C_\perp$ defined by

$$C_\perp(u_1, u_2) = u_1 u_2, \qquad u_1, u_2 \in [0,1].$$

That relationship lead Deheuvels (1981) to proposed tests of interdependence based on the empirical copula C_n, where

$$C_n(u_1, u_2) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\frac{R_{i1}}{n} \le u_1, \frac{R_{i2}}{n} \le u_2\right), \qquad u_1, u_2 \in [0,1].$$

The empirical copula seems to have been studied first by Rüschendorf (1976).

To tackle the d-dimensional case, $d > 2$, where the covariance of the limiting process $\mathbb{H}$ under independence is much more intricate than when $d = 2$, Blum et al. (1961) proposed a decomposition of $\mathbb{H}_n$ based on Möbius formula, leading to processes $\mathbb{H}_{n,A}$, for all $A \subset \{1, \ldots, d\}$, so that each process $\mathbb{H}_{n,A}$ is asymptotically independent of the

others and where the covariance is similar to one obtained in the bivariate case. More precisely, the covariance of the continuous centered limiting processes $\mathbb{H}_A$ is given by

$$\mathrm{Cov}\{\mathbb{H}_A(x), \mathbb{H}_A(y)\} = \prod_{j \in A} [\min\{F_j(x_j), F_j(y_j)\} - F_j(x_j)F_j(y_j)], \quad x, y \in \mathbb{R}^d.$$

That decomposition then appeared in Deheuvels (1981) for copulas, but the author came short of proposing tests of independence. That decomposition was then rediscovered by Ghoudi et al. (2001), who where also able to test independence between non-observable error terms in regression models, using the residuals. With the notable exception of the regression case, testing independence using residuals or more generally pseudo-observations can be quite difficult. See, e.g., Ghoudi and Rémillard (2004). Building on the previous work, Genest and Rémillard (2004) applied the Möbius decomposition method to empirical copulas to test interdependence and serial dependence. That lead them to define the so-called "dependogram." The work of Genest and Rémillard (2004) has been extended recently by Beran et al. (2007) and Kojadinovic and Holmes (2009) for testing independence between random vectors. In addition to test statistics constructed from empirical distribution functions, some researchers considered empirical ▶characteristic functions. See, e.g., Feuerverger (1993), Bilodeau and Lafaye de Micheaux (2005), and more recently Székely and Rizzo (2010). Because independence can be characterized in terms of characteristic functions, the associated tests are consistent in general.

Finally it is worth mentioning Genest and Rémillard (2004), Genest et al. (2006) and Genest et al. (2007) where power comparisons were made for tests of interdependence, the last two for Cramér-von Mises type test statistics.

Testing for Serial Independence

The treatment of serial dependence in (stationary) time series is almost the same as in the previous case, few modifications being necessary for taking into account their particular nature. In fact, if $Y_1, \ldots, Y_n$ represent the time series values for n consecutive periods, then in the bivariate case, one just have to define $X_{i1} = Y_i$ and $X_{i2} = Y_{i+\ell}$, for some lag $\ell \geq 1$. Then the correlation is called autocorrelation of lag ℓ, etc. The so-called correlogram of order k, introduced by Wold in his 1938 Ph.D. thesis, is the graph of the autocorrelations for lags $\ell = 1, \ldots, k$. Under the null hypothesis of serial independence, $n^{1/2} r_n(1), \ldots, n^{1/2} r_n(k)$ converge jointly to independent standard Gaussian variables. One can also adapt the rank-based measures to time series context. More precisely, if $R_1, \ldots, R_n$ are the ranks of $Y_1, \ldots, Y_n$, then one can measure dependence between the pairs $(R_i, R_{i+\ell})$, $i = 1, \ldots, n - \ell$. For more details on rank-based measures of dependence and their properties, see e.g., Hallin et al. (1985) and Ferguson et al. (2000). As before, the tests based on autocorrelations or rank-based measures are not consistent in general, so Skaug and Tjøstheim (1993) proposed to adapt the empirical distribution function methodology to time series context. More precisely, they considered the joint distribution function

$$\tilde{H}_n(x_1, x_2) = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbb{I}(Y_i \leq x_1, Y_{i+\ell} \leq x_2)$$

which was compared to $\tilde{F}_n(x_1)\tilde{F}_n(x_2)$, where $\tilde{F}_n(x) = \tilde{H}_n(x, \infty)$. It is remarkable that the limiting distribution of $n^{1/2}\{\tilde{\mathbb{H}}_n(x_1, x_2) - \tilde{F}_n(x_1)\tilde{F}_n(x_2)\}$ is the same as the limiting distribution of $\mathbb{H}_n$, defined in the previous section. That property was extended by Genest and Rémillard (2004) to the multivariate case, using the associated empirical copula and Möbius decomposition. Other work using ▶empirical processes in a serial context includes Genest et al. (2002) and Kojadinovic and Yan (2010).

Finally, one important problem in time series is checking the serial independence of the non-observable innovations, which is often considered as a test of adequacy for the underlying model. Unfortunately, in most applications, replacing the innovations by residuals changes completely the limiting distribution. See, e.g., Ghoudi and Rémillard (2004). However, using an idea of Brock et al. (1996), Genest et al. (2007) were able to propose tests of independence so that their limiting distribution was not affected by using residuals instead of innovations. However the type of models covered by their methodology is limited to additive models, so it does not include GARCH models.

About the Author

Bruno Rémillard is a Full Professor in Financial Engineering at HEC Montréal since 2001. After completing a Ph.D. in Probability at Carleton University, he was a postdoctoral fellow at Cornell University, before being a professor of Statistics at Université du Québec à Trois-Rivières. He is the author or co-author of more than fifty research articles in Probability, Statistics and Financial Engineering. In 1987, he received the Pierre-Robillard award for the best Ph.D. thesis in Probability and Statistics in Canada and in 2003, he received the prize for the best paper of the year published in the *Canadian Journal of Statistics*. He is also a consultant in the Research and Development group at Innocap since 2007, an alternative investment firm located in Montreal, owned in part by BNP-Paribas and National

Bank of Canada, where he mainly helps developing and implanting new quantitative methods for alternative and traditional portfolios.

Cross References

►Asymptotic Relative Efficiency in Testing
►Autocorrelation in Regression
►Bivariate Distributions
►Categorical Data Analysis
►Copulas
►Copulas: Distribution Functions and Simulation
►Correlation Coefficient
►Durbin–Watson Test
►Kendall's Tau
►Measures of Dependence

References and Further Reading

Beran R, Bilodeau M, Lafaye de Micheaux P (2007) Nonparametric tests of independence between random vectors. J Multivar Anal 98(9):1805–1824

Bilodeau M, Lafaye de Micheaux P (2005) A multivariate empirical characteristic function test of independence with normal marginals. J Multivar Anal 95:345–369

Blum JR, Kiefer J, Rosenblatt M (1961) Distribution free test of independence based on the sample distribution function. Ann Math Stat 32:485–498

Brock WA, Dechert WD, LeBaron B, Scheinkman JA (1996) A test for independence based on the correlation dimension. Econom Rev 15:197–235

Deheuvels P (1981) An asymptotic decomposition for multivariate distribution-free tests of independence. J Multivar Anal 11:102–113

Ferguson TS, Genest C, Hallin M (2000) Kendall's tau for serial dependence. Can J Stat 28:587–604

Feuerverger A (1993) A consistent test for bivariate dependence. Int Stat Rev 61:419–433

Genest C, Ghoudi K, Rémillard B (2007) Rank-based extensions of the Brock Dechert Scheinkman test for serial dependence. J Am Stat Assoc 102:1363–1376

Genest C, Quessy J-F, Rémillard B (2002) Tests of serial independence based on Kendall's process. Can J Stat 30:441–461

Genest C, Quessy J-F, Rémillard B (2006) Local efficiency of a Cramér-von Mises test of independence. J Multivar Anal 97:274–294

Genest C, Quessy J-F, Rémillard B (2007) Asymptotic local efficiency of Cramér-von Mises tests for multivariate independence. Ann Stat 35:166–191

Genest C, Rémillard B (2004) Tests of independence or randomness based on the empirical copula process. Test 13:335–369

Ghoudi K, Kulperger RJ, Rémillard B (2001) A nonparametric test of serial independence for time series and residuals. J Multivar Anal 79:191–218

Ghoudi K, Rémillard B (2004) Empirical processes based on pseudoobservations. II. The multivariate case. In Asymptotic methods in stochastics, Vol 44 of fields institute communications. American Mathematical Society, Providence, RI, pp 381–406

Hallin M, Ingenbleek J-F, Puri ML (1985) Linear serial rank tests for randomness against ARMA alternatives. Ann Stat 13:1156–1181

Kojadinovic I, Holmes M (2009) Tests of independence among continuous random vectors based on cramér-von mises functionals of the empirical copula process. J Multivar Anal 100(6): 1137–1154

Kojadinovic I, Yan J (2010) Tests of serial independence for continuous multivariate time series based on a Möbius decomposition of the independence empirical copula process. Ann Inst Stat Math

Rüschendorf L (1976) Asymptotic distributions of multivariate rank order statistics. Ann Stat 4(5):912–923

Skaug HJ, Tjøstheim D (1993) A nonparametric test of serial independence based on the empirical distribution function. Biometrika 80:591–602

Sklar M (1959) Fonctions de répartition á n dimensions et leurs marges. Publ Inst Stat Univ Paris 8:229–231

Székely GJ, Rizzo ML (2010) Brownian distance covariance. Ann Appl Stat 3(4):1236–1265

Time Series

Peter J. Brockwell
Professor Emeritus
Colorado State University, Fort Collins, CO, USA

A central goal of science, and indeed of a great number of human activities, is to make use of current information in order to obtain useful forecasts of what may happen in the future. If the future is completely independent of the currently available information then this information is of no help. However if there is dependence then we would like to use it to make forecasts which are as accurate as possible in some specified sense. This is one of the key goals of time series analysis (although there are others as we shall see).

A *time series* is a set of observations $\{x_t\}$, each one associated with a particular time t and usually displayed in a *time series plot*, i.e., a graph of x_t as a function of t. An example is the following graph of the natural logarithm of the daily closing value in US dollars of the Dow-Jones Industrial Average, plotted for successive trading days from August 1st, 1997 until August 5th, 2003.

In general the set of times T at which observations are recorded may be a discrete set, as is the case when observations are made at uniformly spaced times (e.g., daily rainfall, annual income etc.) or it may be a continuous interval. For reasons of space we shall restrict attention here to observations at uniformly spaced times, in which case we can label the times $1, 2, \ldots$. In order to account for randomness, we suppose that for each t the observation x_t is just one of many possible values of a random variable X_t that we *might* have observed at time t. The term *time*

series is frequently used to denote both the sequence of random variables $\{X_1, X_2, \ldots\}$ and the particular sequence of observed values $\{x_1, x_2, \ldots\}$.

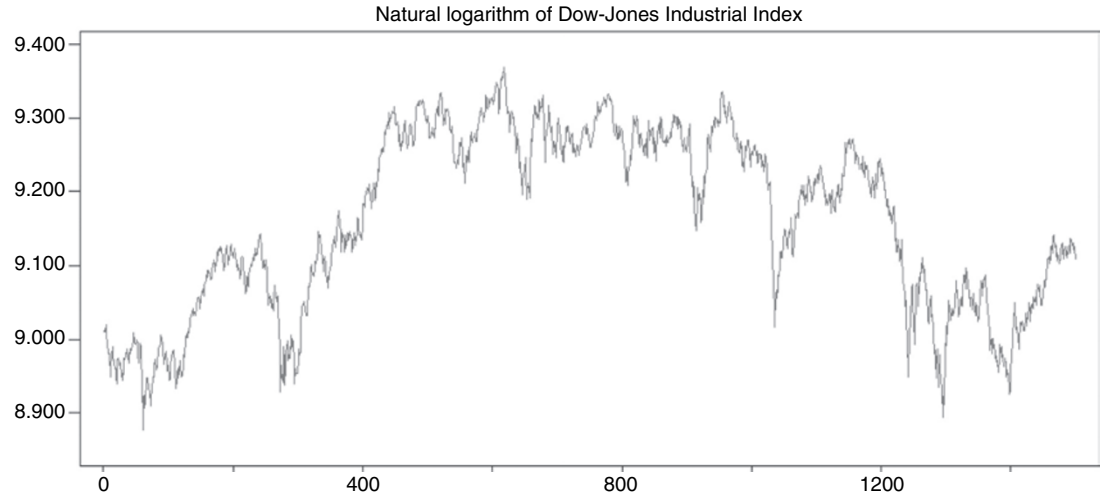
To illustrate the general problem of forecasting in concrete terms, suppose we have a sequence of jointly distributed random variables $\{X_1, X_2, \ldots\}$. Such a sequence is known as a *time series indexed by the positive integers*. Suppose also that our 'information' at time n consists of the observed values of $X_1, \ldots, X_n$. Our problem then is to predict X_{n+h}, the value of the random sequence at the future time $n + h$ using some suitably chosen function $\hat{X}_{n+h}$ of $(X_1, \ldots, X_n)$. In order to assess the performance of our forecast we need some measure of the error of $\hat{X}_{n+h}$. An especially convenient measure, if $EX_n^2 < \infty$ for all n, is the expected squared error, namely $E(X_{n+h} - \hat{X}_{n+h})^2$. Then a rather simple calculation shows that the *best* forecast, i.e., the function of $(X_1, \ldots, X_n)$ which minimizes the expected squared error is the conditional expectation, $E(X_{n+h}|X_1, \ldots, X_n)$. Unfortunately the calculation of this conditional expectation requires knowledge of the conditional distribution of X_{n+h} given $(X_1, \ldots, X_n)$ which is generally unknown and also difficult to estimate from data. (If $\{X_1, X_2, \ldots\}$ is an independent sequence then the conditional expectation is independent of $\{X_j, j \le n\}$, showing that the current information at time n is of no help in predicting X_{n+h} in this case. Time series is therefore primarily concerned with *dependent* random variables and the analysis and utilization of this dependence.) A simpler approach to forecasting X_{n+h} is to look for the *linear combination*, $\hat{X}_{n+h} = a_0 + a_1 X_n + \cdots + a_n X_1$ which minimizes the expected squared error $E(X_{n+h} - \hat{X}_{n+h})^2$. This is a much simpler problem, the solution of which depends only on the expected values EX_i and EX_iX_j, $i, j = 1, 2, \ldots$ Moreover if the joint distribution of $(X_1, X_2, \ldots, X_k)$ is multivariate normal for every positive integer k then this *best linear forecast* is the same as the best forecast.

Forecasting is just one of the many objectives of time series analysis. These depend on the particular field of application. For example, from observed values $x_1, x_2, \ldots$ of the random variables $X_1, X_2, \ldots$ we may wish to understand the mechanism generating the data or perhaps to extract a deterministic 'signal' in the data which is masked by the presence of random noise. We may simply wish to find a compact representation of the available observations or to find a mathematical model which appears to represent the observations well and to use it to simulate further realizations of the series.

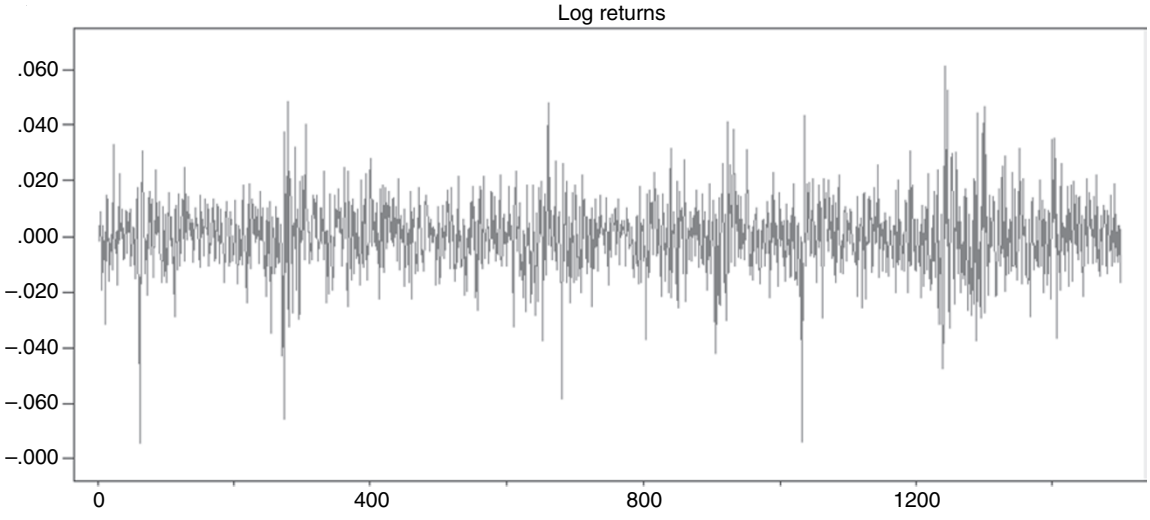
For these applications we need to find a mathematical model which gives a good representation of the data. Typically we select the best-fitting member of a specified family of models by estimating parameters from the observed data and then testing the goodness of fit of the model to the data. Once we are satisfied that the selected model is satisfactory we use it to address the questions of interest. Complete specification of a model for the time series $\{X_1, X_2, \ldots\}$ would consist of a specification of the joint distribution of $(X_1, \ldots, X_k)$ for every positive integer k. However if we are concerned with issues (such as best linear prediction) which depend only on of first and second order moments of the time series, then a model which specifies only first and second-order moments will suffice.

Much of time series analysis is concerned with *stationary* time series. It is clear that if we wish to make predictions, we must assume that *something* does not vary with time. In extrapolating deterministic functions it is common practice to assume that either the function itself or one of its derivatives is constant. The assumption of a constant first derivative leads to linear extrapolation as a means of prediction. In time series we need to predict a series that is typically not deterministic but which contains a random component. The concept of stationarity is used to extend the notion of constancy in time to incorporate randomness. Strict stationarity of the series $\{X_n\}$ means that $(X_1, \ldots, X_k)$ has the same joint distribution as $(X_{h+1}, \ldots, X_{h+k})$ for all positive integers h and k. Weak stationarity means that EX_j and $E(X_{j+h}X_j)$ exist and are both independent of j. Thus stationarity requires the probabilistic properties (or, in the case of weak stationarity, the first and second moment properties) of the series to be invariant to shifts along the time axis. Information concerning the properties of stationary processes and estimation of their parameters can be found in the many books dealing with time series analysis. Without the assumption of stationarity the formulation of appropriate models and estimation of their parameters becomes much more difficult, although in recent years progress has been made in this direction.

The practical importance of stationary processes lies in the fact that many empirically observed series, which themselves cannot be well fitted by a stationary time series model, can be simply transformed into a new series which can. If a stationary model is fitted to the transformed series, it can be used to generate forecasts of the transformed series which can then be transformed back to generate corresponding forecasts for the original series. For example if we denote by X_n the natural logarithm of the closing value of the Dow-Jones Index on day n and consider the differenced series $Y_n := X_n - X_{n-1}$ (known as the *log return* for day n) then Y_n can be rather well represented as a stationary time series. The realization of the series $\{Y_n\}$ corresponding to the realization of $\{X_n\}$ in Fig. 1 is shown in Fig. 2.



Time Series. Fig. 1 The natural logarithm of the daily closing Dow-Jones Industrial Average for successive trading days from August 1st, 1997 until August 5th, 2003



Time Series. Fig. 2 The daily log returns for the Dow-Jones Industrial Average for successive trading days from August 1st, 1997 until August 5th, 2003

The dependence between observations of a stationary time series $\{X_n\}$ is frequently measured by the *autocovariance function*,

$$\gamma(h) := E[(X_{t+h} - \mu)(X_t - \mu)],$$

where $\mu := EX_t$, or the *autocorrelation function*,

$$\rho(h) := \gamma(h)/\gamma(0),$$

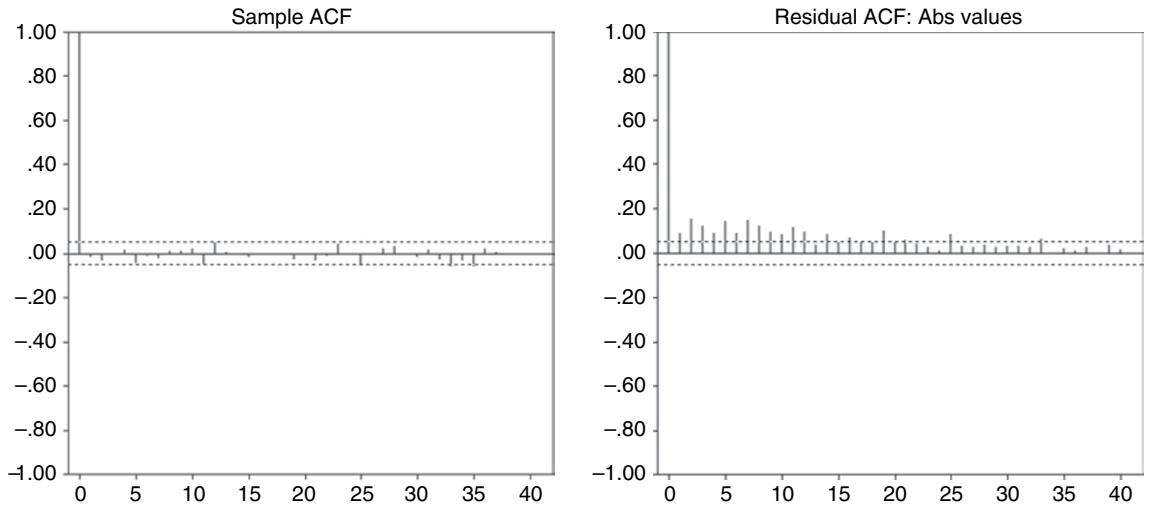
which specifies the correlation between any two observations separated by a time interval of length h. These quantities can be estimated by the *sample autocovariance function*,

$$\hat{\gamma}(h)) = n^{-1} \sum_{j=1}^{n-h} (x_{j+h} - \overline{x})(x_j - \overline{x}),$$

and *sample autocorrelation function*,

$$\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0),$$

respectively, where $\overline{x}$ denotes the sample mean, $n^{-1} \sum_{j=1}^{n} x_j$.



Time Series. Fig. 3 The sample autocorrelation function of the log returns in Fig. 2 (left) and the absolute values of the log returns (right)

The graph on the left of Fig. 3 shows the sample autocorrelation function of the differenced series in Fig. 2 with 95% significance bounds for testing the deviation of each sample autocorrelation value from zero. As there is no autocorrelation significantly different from zero from lags 1 through 40, it appears that the differenced series is uncorrelated. The best *linear* forecast of any future difference is therefore equal to the estimated mean of the differences (which is actually 0.0007). The best linear forecast of the natural logarithm of the Dow-Jones Industrial Average h trading days after August 5th, 2003 is therefore the value on August 5th (9.1090) plus $0.0007h$.

The autocorrelations in this example however do not tell the whole story. If the series of differences, instead of being merely uncorrelated with mean 0.0007, had been *independent*, then the mean value would have been the *best* rather than just the best *linear* forecast of future differences. However the graph on the right of Fig. 3, the sample autocorrelation function of the *absolute values* of the differences is clearly significantly different from zero at a number of lags. Since this implies that the absolute differences are not independent, it implies also that the differences themselves are not independent. This phenomenon of dependence with negligible correlation is a striking feature of many financial time series and has led to the development of a variety of intriguing models such as ARCH and GARCH models to account for this and related phenomena.

Probably the most widely used models for stationary time series have been the so-called ARMA (or autoregressive moving average) models. The series $\{X_n, n = 0, \pm 1, \pm 2, \ldots\}$ is said to be an ARMA(p, q) process if it is a stationary solution of the linear difference equations,

$$X_n - \phi_1 X_{n-1} - \cdots - \phi_p X_{n-p} = Z_1 + \theta_1 + \cdots \theta_q Z_{t-q},$$

where $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ are real valued coefficients, $\phi_p \neq 0$, $\theta_q \neq 0$, and $\{Z_n\}$ is a sequence of independent (or sometimes just uncorrelated) random variables, each with mean 0 and variance σ^2. Depending on the values of p and q and the coefficients ϕ_j and θ_j, an enormous range of sample autocorrelation functions can be replicated by members of the ARMA family. There is a vast literature dealing with problems of ▶model selection, estimation and forecasting for these processes. A standard technique (developed and popularized by Box and Jenkins) for dealing with observed time series which appear to be non-stationary is to apply differencing until the data appears to be representable by a stationary model and then to fit an ARMA model to the resulting series. The original data is then said to be represented by an ARIMA (or integrated ARMA) model.

In the last thirty years there has been an explosion of interest in more elaborate non-linear models to account for phenomena which cannot be accounted for in the classical linear framework provided by ARMA models. These include threshold, bilinear, ARCH, GARCH, Markov switching models and many others too numerous to be discussed here in any detail. Details can be found in some of the following references.

Acknowledgments

Work supported by NSF Grant DMS-0744058.

About the Author

Professor Brockwell is Associate Editor of the *Journal of the Japanese Statistical Society* and of the *Annals of the Institute of Statistical Mathematics*. He is a Fellow of the Institute of Mathematical Statistics, the American Statistical Association, and member of the International Statistical Institute. He was Von Neumann Guest Professor, Technical University of Munich (2001–2002). He is coauthor, with Richard Davis, of two widely used texts on Time Series Analysis: *Introduction to Time Series and Forecasting* (2nd edition, Springer, 2002), and *Time Series: Theory and Methods* (2nd edition, Springer, 1991).

Cross References

►Bayesian Approach of the Unit Root Test
►Box–Jenkins Time Series Models
►Business Forecasting Methods
►Data Mining Time Series Data
►Detecting Outliers in Time Series Using Simulation
►Detection of Turning Points in Business Cycles
►Dickey-Fuller Tests
►Exponential and Holt-Winters Smoothing
►Forecasting Principles
►Forecasting with ARIMA Processes
►Forecasting: An Overview
►Heteroscedastic Time Series
►Intervention Analysis in Time Series
►Median Filters and Extensions
►Models for Z_+-Valued Time Series Based on Thinning
►Nonlinear Time Series Analysis
►Optimality and Robustness in Statistical Forecasting
►Seasonal Integration and Cointegration in Economic Time Series
►Seasonality
►Singular Spectrum Analysis for Time Series
►Statistical Aspects of Hurricane Modeling and Forecasting
►Structural Time Series Models
►Time Series Models to Determine the Death Rate of a Given Disease
►Time Series Regression
►Trend Estimation
►Vector Autoregressive Models

References and Further Reading

Anderson TW (1971) The statistical analysis of time series. Wiley, New York
Box GEP, Jenkins GM, Reinsel GC (2008) Time series analysis: forecasting and control, 4th edn. Wiley, New York
Brockwell PJ, Davis RA (2002) Introduction to time series and forecasting, 2nd edn. Springer-Verlag, New York
Brockwell PJ, Davis RA (1991) Time series: theory and methods, 2nd edn. Springer-Verlag, New York
Fuller WA (1995) Introduction to statistical time series, 2nd edn. Wiley, New York
Hannan EJ (1970) Multiple time series. Wiley, New York
Lütkepohl H (1993) Introduction to multiple time series analysis, 2nd edn. Springer-Verlag, Berlin
Priestley MB (1981) Spectral analysis and time series. Academic Press, London
Shumway RH, Stoffer DS (2006) Time series analysis and its applications with R examples, 2nd edn. Springer-Verlag, New York
Tong H (1990) Non-linear time series: a dynamical systems approach. Oxford University Press, Oxford
Tsay RS (1990) Analysis of financial time series. Wiley, New York

Time Series Models to Determine the Death Rate of a Given Disease

Dahud K. Shangodoyin
Associate Professor
University of Botswana, Gaborone, Botswana

Statistics as a scientific subject of decision making under uncertainty is critical to the evaluation of health indicators that are of paramount importance to public health. Health issues, in most cases, are nondeterministic, which leaves their study to use the most suitable probabilistic approaches. Statistical research in health can be conducted in the following areas:

(a) ►Meta-analysis: Meta-analysis mathematically combine the results of numerous studies in order to improve the reliability of the results. Studies chosen for inclusion in a meta-analysis must be sufficiently similar in a number of characteristics in order to accurately combine their results; for instance, issues surrounding meta-analyses of individual patient data could be analyzed, and missing data can be dealt with at the patient level.

(b) Statistical Epidemiology: This aspect is broad and includes the following: (i) Clustered observational studies in which sample clusters of people are utilized for health research. This is becoming increasingly common, especially with patients in various health practices, people within health districts, and children within schools. The hierarchical nature of the data then takes on a multi-level structure that needs to be accounted for in the analysis. (ii) Ecological studies are carried out at an aggregate level, for example, the ward or district level, and can be used to investigate the relationship between socio-economic risk factors and ill-health. (iii) Longitudinal studies are useful

because following people over time is costly and time consuming and may have problems of missing data and consistency of measurement over time. Research of interest in this area could include a matched cohort study of coping and depression in parents of children newly diagnosed with terminal diseases.

(c) Survival Analysis: This is the analysis of time-to-event data, and is relevant to many clinical studies where the outcome of interest relates to the time taken for some event to occur, for instance, time to first seizure or time to death following ▶randomization.

The most important aspect of survival analysis is the measures of health indicator, especially the study of death rate for emerging and re-emerging diseases. Deliberation is continuing on how best to estimate the death rate of an emerging contagious disease, which is of paramount importance to global public health. The 2009 outbreak of influenza caused by a novel influenza A (H1N1) virus has given the World Health Organization (WHO) concern on how best to estimate the death rate arising from H1N1 throughout the world. As a matter of fact, the case of estimating the global death rate arising from the outbreak of severe acute respiratory syndrome (SARS) in 2003 also generated much public controversy (Altman (2003)). The WHO's convectional formula for computation of death rate is simply the ratio of the number of known deaths to the total number of confirmed cases (Mathers and Loncar (2006)), however, this formula is likely to underestimate the true death rate because the outcomes of many cases were still unknown or uncertain at the time these figures might have being compiled. In other words, the WHO approach has a problem of "selection" bias because the conditional probability of death among cases of known outcomes need not be equal to the unconditional probability of death. Another notable model of estimating the death rate of an emerging disease is the cohort approach. In this model cases from the same day constitute a cohort and the binomial analysis is restricted to the cases from a complete cohort, that is, cases with a known outcome at the end of the study period. The restrictions in this model lead to loss of a substantial volume of data and require some data that may not be accessible to researchers. The generalized mixed effect model of estimating death rate discussed by Chan and Tong (2006) is less biased and converges quickly to the death rate computed from the complete data, but the model specified leads to a singular precision matrix for the unknown parameters. In addition, the choice of the singular value decomposition presented may restrain this approach for practical use. Chang and Tong concluded that further research was needed on how to carry out the estimation of the conditional mean death rate with the constraint that the estimated death rate should be greater than or equal to zero. Shangodoyin (2009) proffers another method by using a novel time series model to estimate the mean death rate of an emerging or re-emerging disease with bilinear induced parameters; from the applied point of view, both the Tong and Chan (2006) and Shangodoyin (2009) models could be used by experts in monitoring and evaluating the death rate of a disease over time. For a general linear model (see ▶General Linear Model) the mean death rate could be specified as:

$$\mu_t = \sum_{a_1}^{a_2} p_j C_{t-j}$$

where $a_1, a_2 \geq 0$ are lower and upper bounds of time to death. The model is bilinear for estimating the mean deaths at time t as:

$$\mu_t = \alpha\mu_{t-1} + \beta\mu_{t-1}e_{t-1} + \sum_{1}^{u} p_j C_{t-j} + e_t.$$

By making all the necessary mathematical assumptions, the overall death rate for one-step time to death is given by

$$\dot{p}_1 = \frac{\sum_{1}^{n} \dot{\mu}_1 C_{t-1}}{\sum_{1}^{n} C_{t-1}^2}$$

where C_{t-j} is *the* number of confirmed cases at time $t - j$, $\dot{\mu}_t = \frac{\sum_{1}^{t} D_t}{t}$; $\forall t = 1, 2, \ldots$ and D_t is the *number* of deaths at time t. Readers should refer to the paper by Shangodoyin (2009) for details of the derivations.

In conclusion, statistical models play significant roles in the evaluation and monitoring of death rates from both emerging and re-emerging disease; and the use of most suitable time series models will provide the best insight to the future mortality rate for the given disease.

About the Author

Professor D. K. Shangodoyin was born in August, 1961 and started is academic career in 1986 and had taught Statistics in six African Universities. He is currently an Associate professor of Statistics at the University of Botswana, Southern Africa. He was the first alumni of the University of Ibadan in Nigeria to head the Department of Statistics at the Nigerian Premier University. He is currently the Statistics Pan African Society (SPAS) coordinator for SADC region in Africa. His broad area of research is the Theory and Application of Time Series, Bayesian inference and Econometrics modeling.

Cross References

►Meta-Analysis
►Modeling Survival Data
►Statistical Methods in Epidemiology
►Survival Data
►Time Series

References and Further Reading

Altman KL (2003, May 7) The SARS epidemic: the front-line research. New York Times

Mathers CD, Loncar D (2006) Projections of global mortality and burden of disease from 2002 to 2030. PLos Med 3(11):2011–2030

Shangodoyin DK (2009) Time series model for estimating the death rate of an emerging and re-emerging disease. In: 57th ISI Session, Durban, South Africa. www.stats.gov.za/isi2009

Tong H, Chan K (2006) Estimating the death rate of an emerging disease by Time Series Analysis. Technical Report, Department of Statistics & Actuarial Science, University of Iowa, Iowa, USA

Time Series Regression

William W. S. Wei
Professor
Temple University,
Philadelphia, PA, USA

Introduction

A regression model is used to study the relationship of a dependent variable with one or several independent variables. The standard regression model is represented by the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon,$$

where Y is the dependent variable, $X_1, \ldots, X_k$ are the independent variables, $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients, and ε is the error term. When time series data are used in the model, it becomes time series regression, and the model is often written as

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \varepsilon_t,$$

or equivalently

$$Y_t = \mathbf{X}_t'\boldsymbol{\beta} + \varepsilon_t, \tag{1}$$

where $\mathbf{X}_t' = [1, X_{1,t}, \ldots, X_{k,t}]$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]'$. The standard regression assumptions on the error variable are that the ε_t are i.i.d. $N\left(0, \sigma_\varepsilon^2\right)$. Under these standard assumptions, it is well known that the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is a minimum variance unbiased estimator, distributed as multivariate normal, $N\left(\boldsymbol{\beta}, \sigma_\varepsilon^2 \mathbf{I}\right)$. When $\mathbf{X}_t'$ is stochastic in Model (1), conditional on $\mathbf{X}_t'$, the results about the OLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ also hold as long as ε_s and $\mathbf{X}_t'$ are independent for all s and t. However, the standard assumptions associated with these models are often violated when time series data are used.

Regression with Autocorrelated Errors

When $\mathbf{X}_t'$ is a vector of a constant 1 and k lagged values of Y_t, i.e., $\mathbf{X}_t' = [1, Y_{t-1}, \ldots, Y_{t-k}]$, and ε_t is white noise, the model in (1) states that the variable Y_t is regressed on its own past k lagged values and hence is known as autoregressive model of order k, i.e., $AR(k)$ model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_k Y_{t-k} + \varepsilon_t. \tag{2}$$

The OLS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is still a minimum variance unbiased estimator. However, this result no longer holds when the ε_t are autocorrelated. In fact, when this is the case, the estimator is not even consistent and the usual tests of significance are invalid. This is an important caveat. When time series are used in a model, it is the norm rather than the exception that the error terms are autocorrelated. Even in univariate time series analysis when the underlying process is known to be an AR model as in (2), the error terms ε_t could still be autocorrelated unless the correct order of k is chosen. Thus, a residual analysis is an important step in regression analysis when time series variables are involved in the study.

There are many methods that can be used to test for autocorrelation of the error term. For example, one can use the test based on the Durbin–Watson statistic,

$$d = \frac{\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2} \approx 2(1 - \hat{\rho}_1), \tag{3}$$

where $\hat{\varepsilon}_t$ is residual series from the OLS procedure. Clearly, d lies between 0 and 4. A value close to 2 indicates no first-order autocorrelation, a value much less than 2 and close to 0 indicates a positive first-order autocorrelation and a value much greater than 2 and close to 4 indicates a negative first-order autocorrelation. To help make decision, in terms of the null hypothesis of no first-order autocorrelation against the alternative hypothesis of positive first-order autocorrelation, the critical values of Durbin-Watson, d_L and d_U can be constructed, which are functions of the number independent variables, the number of observations, and the significance level. The null hypothesis is

rejected if $0 < d < d_L$, is not rejected if $d_U < d < 2$, and inconclusive if $d_L < d < d_U$. For the null hypothesis of no first-order autocorrelation against the alternative hypothesis of negative first-order autocorrelation, the same table can be used since it is simply the mirror image of the former case when we look at the case from the endpoint of 4 instead of the endpoint of 0. Thus, the null hypothesis is rejected if $4 - d_L < d < 4$, is not rejected if $2 < d < 4 - d_U$, and inconclusive if $4 - d_U < d < 4 - d_L$.

More generally, to study the autocorrelation structure of the error term, we can perform the residual analysis with time series model identification statistics like the sample autocorrelation function (ACF) and sample partial autocorrelation function (PACF). Through these identification statistics, one can detect not only whether the residuals are autocorrelated but also identify its possible underlying model. A final analysis can then be performed on a model with autocorrelated errors as follows:

$$Y_t = \mathbf{X}_t'\boldsymbol{\beta} + \varepsilon_t \tag{4}$$

for $t = 1, 2, \ldots, n$, where

$$\varepsilon_t = \varphi_1\varepsilon_{t-1} + \cdots + \varphi_p\varepsilon_{t-p} + a_t \tag{5}$$

and the a_t are i.i.d. $N(0, \sigma^2)$.

Let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{X}_1' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix}, \text{ and } \boldsymbol{\xi} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The matrix form of the model in (4) is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \tag{6}$$

where $\boldsymbol{\xi}$ follows a multivariate normal distribution (see ►Multivariate Normal Distributions) $N(0, \boldsymbol{\Sigma})$ When $\varphi_1, \ldots, \varphi_p$, and σ^2 are known in (5), $\boldsymbol{\Sigma}$ can be easily calculated. The diagonal element of $\boldsymbol{\Sigma}$ is the variance of ε_t, the jth off-diagonal element corresponds to the jth autocovariance of ε_t, and they can be easily computed from (5). Given $\boldsymbol{\Sigma}$, the generalized least squares (GLS) estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \tag{7}$$

is known to be a minimum variance unbiased estimator.

Normally, we will not know the variance-covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{\xi}$ because even if ε_t follows an $AR(p)$ model given in (5), the σ^2 and AR parameters φ_j are usually unknown. As a remedy, the following iterative GLS is often used:

(1) Calculate OLS residuals $\hat{\varepsilon}_t$ from OLS fitting of Model (4).
(2) Estimate φ_j and σ^2 for the $AR(p)$ model in (5) based on the OLS residuals, $\hat{\varepsilon}_t$, using any time series estimation method. For example, a simple conditional OLS estimation can be used.
(3) Compute $\boldsymbol{\Sigma}$ from the model (5) using the values of φ_j and σ^2 obtained in step (2).
(4) Compute GLS estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, using the $\boldsymbol{\Sigma}$ obtained in step (3). Compute the residuals $\hat{\varepsilon}_t$ from the GLS model fitting in step (4), and repeat the above steps (1) through (4) until some convergence criterion (such as the maximum absolute value change in the estimates between iterations becoming less than some specified quantity) is reached.

More generally, the error structure can be modified to include an ARMA model. The above GLS iterative estimation can still be used except that a nonlinear least squares estimation instead of OLS is needed to estimate the parameters in the error model. Alternatively, by substituting the error model in the regression equation (4), we can also use the nonlinear estimation or maximum likelihood estimation to jointly estimate the regression and error model parameters $\boldsymbol{\beta}$ and $\varphi_j's$, which is available in standard software.

It should be pointed out that although the error term, ε_t, can be autocorrelated in the regression model, it should be stationary. A nonstationary error structure could produce a spurious regression where a significant regression can be achieved for totally unrelated series.

Regression with Heteroscedasticity

One of the main assumptions of the standard regression model in Eq. 1 or the regression model with autocorrelated errors in Eq. 4 is that the variance, σ_ε^2, is constant. In many applications, this assumption may not be realistic. For example, in financial investment, it is generally agreed that stock markets' volatility is rarely constant.

Such a model having a non-constant error variance is called a heteroscedasticity model. There are many approaches which can be used to deal with heterocedasticity. For example, the weighted regression is often used if the error variances at different times are known or if the variance of the error term varies proportionally to the value of an independent variable. In time series regression we often have the situation where the variance of the error term is related to the magnitude of the past errors. This leads to the

conditional heteroscedasticity model, introduced by Engle (1982), where in terms of Eq. 1 we assume that

$$\varepsilon_t = \sigma_t e_t, \tag{8}$$

the e_t are i.i.d. random variable with mean 0 and variance 1, and

$$\sigma_t^2 = \theta_0 + \theta_1 \varepsilon_{t-1}^2 + \theta_2 \varepsilon_{t-2}^2 + \cdots + \theta_s \varepsilon_{t-s}^2. \tag{9}$$

Given all the information up to time $(t-1)$ the conditional variance of the ε_t becomes

$$\begin{aligned} Var_{t-1}(\varepsilon_t) &= E_{t-1}\left(\varepsilon_t^2\right) = E\left(\varepsilon_t^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\right) = \sigma_t^2 \\ &= \theta_0 + \theta_1 \varepsilon_{t-1}^2 + \theta_2 \varepsilon_{t-2}^2 + \cdots + \theta_s \varepsilon_{t-s}^2. \end{aligned} \tag{10}$$

which is related to the squares of past errors, and it changes over time. A large error through ε_{t-j}^2 gives rise to the variance which tends to be followed by another large error. This is a common phenomenon of volatility clustering in many financial time series.

From the forecasting results, we see that Eq. 10 is simply the optimal forecast of ε_t^2 from the following $AR(s)$ model:

$$\varepsilon_t^2 = \theta_0 + \theta_1 \varepsilon_{t-1}^2 + \theta_2 \varepsilon_{t-2}^2 + \cdots + \theta_s \varepsilon_{t-s}^2 + a_t, \tag{11}$$

where the a_t is a $N\left(0, \sigma_a^2\right)$ white noise process. Thus, Engle (1982) called the model of the error term ε_t with the variance specification given in (8) and (9) or equivalently in (10) as autoregressive conditional heteroscedasticity model of order s ($ARCH(s)$).

Bollerslev (1986) extends the $ARCH(s)$ model to the $GARCH(r,s)$ model (generalized autoregressive conditional heteroscedasticity model of order (r,s)) so that the conditional variance of the error process is related not only to the squares of past errors but also to the past conditional variances. Thus, we have the following more general case

$$\varepsilon_t = \sigma_t e_t, \tag{12}$$

where the e_t are i.i.d. random variable with mean 0 and variance 1,

$$\sigma_t^2 = \theta_0 + \phi_1 \sigma_{t-1}^2 + \cdots + \phi_r \sigma_{t-r}^2 + \theta_1 \varepsilon_{t-1}^2 + \cdots + \theta_s \varepsilon_{t-s}^2, \tag{13}$$

and the roots of $(1 - \phi_1 B - \cdots - \phi_r B^r) = 0$ are outside the unit circle. To guarantee $\sigma_t^2 > 0$ we assume that $\theta_0 > 0$ and ϕ_i and θ_j are nonnegative.

More generally, the regression model with autocorrelated error can be combined with the conditional heteroscedasticity model, i.e.,

$$Y_t = \mathbf{X}_t' \beta + \varepsilon_t, \tag{14}$$

where

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \cdots + \phi_p \varepsilon_{t-p} + a_t, \tag{15}$$

$$a_t = \sigma_t e_t, \tag{16}$$

$$\sigma_t^2 = \theta_0 + \phi_1 \sigma_{t-1}^2 + \cdots + \phi_r \sigma_{t-r}^2 + \theta_1 a_{t-1}^2 + \cdots + \theta_s a_{t-s}^2, \tag{17}$$

and the e_t are i.i.d. $N(0,1)$. To test for the heteroscedasticity in this model, we follow:

(1) Calculate OLS residuals $\hat{\varepsilon}_t$ from the OLS fitting of (14).
(2) Fit an AR(p) model (15) to the $\hat{\varepsilon}_t$.
(3) Obtain the residuals $\hat{a}_t$ from the AR fitting in (15).
(4) Form the series $\hat{a}_t^2$, compute its sample ACF, PACF, and check whether these ACF and PACF follow any pattern. A pattern of these ACF and PACF not only indicates ARCH or GARCH errors, it also forms a good basis for their order specification.

For more detailed discussions and examples, we refer readers to Wei (2006).

About the Author

Professor Wei is a Past President, International Chinese Statistical Association (2001–2002). He was the Chair of the Department of Statistics at Temple University (1982–1987). He is a Fellow of the ASA, a Fellow of the RSS, and Elected Member of the ISI. He is currently an Associate Editor of the *Journal of Forecasting* and the *Journal of Applied Statistical Science*.

Cross References

►Autocorrelation in Regression
►Durbin–Watson Test
►Heteroscedastic Time Series
►Heteroscedasticity
►Least Squares
►Linear Regression Models
►Minimum Variance Unbiased
►Multivariate Normal Distributions
►Time Series

References and Further Reading

Bollerslev T (1986) Generalized autoregressive conditional heteroscedasticity. J Econom 31:307–327
Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50:987–1007
Wei WWS (2006) Time series analysis – Univariate and multivariate methods, 2nd edn. Pearson Addison-Wesley, Boston

Total Survey Error

Paul P. Biemer
Professor
RTI International and the University of North Carolina, Chapel Hill, NC, USA

Total survey error refers to the totality of error that can arise in the design, collection, processing and analysis of survey data. The concept dates back to the early 1940's although it has been revised and refined by a many authors over the years. Deming (1944), in one of the earliest works, describes "13 factors that affect the usefulness of surveys." These factors include sampling errors as well as nonsampling errors; i.e., the other factors that will cause an estimate to differ from the population parameter it is intended to estimate. Prior to Deming's work, not much attention was being paid to nonsampling errors and, in fact, textbooks on survey sampling made little mention of them. Indeed, classical sampling theory (Neyman 1934) assumes survey data are error free except for sampling error. The term "total survey error" originated with an edited volume of the same name (Andersen et al. (1977)).

A number of authors have provided a listing of the general sources of nonsampling error. For example, Biemer and Lyberg (2003) list five sources: specification, frame, nonresponse, measurement and data processing (including post-survey adjustment). A *specification error* arises when the concept implied by the survey question and the concept that should be measured in the survey differ. *Frame error* arises in the process for constructing, maintaining, and using the sampling frame(s) for selecting the survey sample. It includes the inclusion of nonpopulation members, exclusions of population members, and frame duplications. *Nonresponse error* encompasses both unit and item nonresponse. *Unit nonresponse* occurs when a sampled unit does not respond to any part of a ►questionnaire. *Item nonresponse* error occurs when the questionnaire is only partially completed because an interview was prematurely terminated or some items that should have been answered were skipped or left blank. *Measurement error* includes errors arising from respondents, interviewers, survey questions and factors which affect survey responses. Finally, *data processing error* includes errors in editing, data entry, coding, computation of weights, and tabulation of the survey data.

The total survey error in a survey estimator, $\hat{\theta}$, for a population parameter, θ, can be summarized by the mean squared error of the estimator defined as

$$\begin{aligned} MSE(\theta) &= E(\hat{\theta} - \theta)^2 \\ &= B^2(\hat{\theta}) + Var(\hat{\theta}) \end{aligned} \tag{1}$$

where $B^2(\hat{\theta}) = E(\hat{\theta} - \theta)$ is the bias in the estimator and $Var(\hat{\theta}) = E(\hat{\theta} - \theta)^2$ is the variance of the estimator. For estimating the population mean, biases arise from systematic errors in the survey process; i.e., errors that are either predominately positive or predominately negative. As an example, sensitive items such as drug use tend to be underreported in surveys causing a negative bias in the estimated proportion of drug users. Nonresponse can also create a bias by systematically excluding from the survey data, individuals who differ on the survey characteristics from respondents.

The variance component of the MSE arises as a result of sampling error as well as variable nonsampling errors. Variable nonsampling error can be described roughly as the error remaining after accounting for the systematic errors. Variable errors tend to fluctuate randomly from unit to unit and have little or no effect on bias. As an example, interviewer estimates of housing values or neighborhood income levels may vary randomly from their true values.

To illustrate the effects of systematic and variable error, consider a ►simple random sample of size n to estimate the mean, μ, of a large population. An elementary model for an observation, y_i, for characteristic y on sample unit i is

$$y_i = \mu_i + \varepsilon_i \tag{2}$$

where μ_i is the true value of the characteristic (i.e., the value that would have been observed without error), and ε_i is the error in the observation. Here ε_i represents the cumulative effect of all systematic and variable error sources for the ith unit. If the net error is 0, i.e., if $E(\varepsilon_i) = 0$, then there is no bias in the estimator $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$. In that case, the errors are variable; i.e., no systematic errors. When systematic errors arise in the observations, $E(\varepsilon_i) = \beta \neq 0$ where β is the bias in $\bar{y}$. Under this model, $MSE(\bar{y})$ can be written as

$$MSE(\bar{y}) = \beta^2 + \frac{\sigma_\mu^2 + \sigma_\varepsilon^2}{n} \tag{3}$$

where $\sigma_\mu^2 = Var(\mu_i)$ and $\sigma_\varepsilon^2 = Var(\varepsilon_i)$. In this expression, β is the nonsampling bias, $n^{-1}\sigma_\mu^2$ is the sampling variance and $n^{-1}\sigma_\varepsilon^2$ is the nonsampling variance.

It is often useful to decompose both the nonsampling bias and variance components further by terms representing for the various sources of error in the survey process. As an example, suppose the major sources of bias include

the sampling frame, nonresponse and measurement bias. Then bias squared component can be expanded to include bias components for these sources as follows:

$$\beta = B_{FR} + B_{NR} + B_{MEAS} \tag{4}$$

where B_{FR} denotes frame bias, B_{NR}, nonresponse bias and B_{MEAS}, measurement bias. The variance component can also be expanded to include terms for all the major contributors of variable error such as sampling error, interviewers, respondents and other variable errors. Now the MSE can be rewritten as

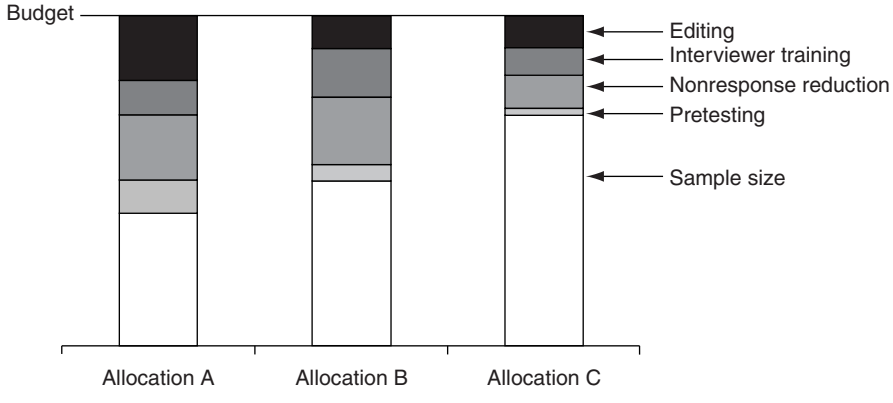
$$MSE(\bar{y}) = \left(B_{FR} + B_{NR} + B_{MEAS} + B_{DP}\right)^2 + \frac{\sigma_\mu^2 + \sigma_{int}^2 + \sigma_{res}^2 + \sigma_e^2}{n} \tag{5}$$

where σ_{int}^2 is the interviewer variance component, σ_{res}^2 is the respondent variance component, and σ_e^2 is the variance associated with all other sources. Thus, σ_ε^2 in (3) can be decomposed as $\sigma_\varepsilon^2 = \sigma_\mu^2 + \sigma_{int}^2 + \sigma_{res} + \sigma_e^2$. This form of the MSE assumes uncorrelated errors; however, the MSE can be also expanded to include correlations among the error from the same or difference error sources (see, for example, Biemer (2010)). The estimation of the components of the MSE can be quite challenging (see Mulry and Spencer 1991, for an application of the total survey error concept to the 1990 Decennial Census). Biemer (2010) provides a simplified estimator of the total MSE when multiple error sources are considered.

Finally, a critical part of the total survey error concept is error reduction and control. It is seldom possible to conduct every stage of the survey process at maximum accuracy since that would likely entail exceeding the survey budget and schedule by a considerable margin. Even under the best circumstances, some errors will necessarily remain in the data so that other, more serious errors can be avoided or reduced. For example, training interviewers adequately may require eliminating or limiting some quality control activities during data processing; but that might increase the data processing error. Efforts to reduce nonresponse bias may require substantial reductions during the survey pretesting phase to stay within budget. How should these resource allocation decisions be made? Making wise trade-offs requires an understanding of the sources of nonsampling error, their relative importance to data quality, and how they can be controlled. One answer is *optimal survey design*.

Optimal survey design aims to minimize the MSE (expressed in terms of the major error sources in the survey) subject to constraints on the survey process imposed by the budget, timeliness and other design considerations. Provide a design that is truly optimal (i.e., the best possible) may be an unattainable goal though it can be approximated. Doing so requires knowledge of the major error sources, their relative magnitudes and the most efficient and effective methods for nonsampling error reduction. Careful planning is then required to allocate survey resources to the various stages of the survey process so that the major sources of error are controlled to optimal, or near optimal levels.

To illustrate, Figure 1 depicts three possible resource allocation strategies satisfying the same budget constraint. Allocation *A* sacrifices sampling precision (i.e., sample size) for the sake of nonsampling error minimization by allocating more resources to editing, interviewer training, nonresponse reduction and questionnaire pretesting.



Total Survey Error. Fig. 1 Three potential cost allocations for the same fixed budget, each with very different implications for total survey error

Allocation *C* reduces these nonsampling error control strategies in order to boost the sample size thereby achieving greater sampling precision. Allocation *B* is a compromise between these two designs. Many other allocation schemes are possible. The challenge for the survey designer is to choose a single allocation strategy that provides the optimal balance between sampling error reduction and nonsampling error control while staying within budget. This is made even more difficult if there is insufficient information on the magnitudes of the total error components and scant knowledge regarding nonsampling error control strategies that are most effective at reducing the components of total survey error.

About the Author

Professor Biemer is RTI International Distinguished Fellow of Statistics, Associate Director of Survey Research and Development at the Odum Institute and Founding Director of the Certificate Program in Survey Methodology at the University of North Carolina, Chapel Hill. Prior to joining RTI, Professor Biemer headed the department of statistics at New Mexico State University and held a number of positions at the Bureau of the Census. Professor Biemer's book, *Introduction to Survey Quality* (with Lars Lyberg) is a widely used course text. He also co-edited following books: *Measurement Errors in Surveys*, *Survey Measurement and Process Quality*, and *Telephone Survey Methodology* which were published by John Wiley & Sons. His newest book, *Latent Class Analysis of Survey Error*, also published by John Wiley & Sons, is currently in press. He is a Fellow of the ASA and the AAAS, an Elected Member of the ISI and Associate Editor of the *Journal of Official Statistics*.

Crosss References

►Bias Analysis
►Business Surveys
►Nonresponse in Surveys
►Nonsampling Errors in Surveys
►Sample Survey Methods
►Sampling From Finite Populations

References and Further Reading

Andersen R, Kasper J, Frankel M, and Associates (1979) Total survey error. Jossey-Bass Publishers, San Francisco

Biemer PP (2010) Chapter 2 – Overview of design issues: total survey error. In: Marsden P, Wright J (eds) Handbook of survey research, 2nd edn. Bingley, United Kingdom: Emerald Group Publishing, LTD

Biemer P, Lyberg L (2003) Introduction to survey quality. Wiley, Hoboken

Deming WE (1944) On errors in surveys. Am Sociol Rev 9(4): 359–369

Mulry M, Spencer B (1991) Total error in PES estimates of population. J Am Stat Assoc 86(416):839–863

Neyman J (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J Roy Stat Soc 97:558–606

Tourism Statistics

Stephen L. J. Smith
Professor
University of Waterloo, Waterloo, ON, Canada

The development of consistent measures of tourism has challenged tourism statisticians and economists since the 1930s (Smith 2004). The challenges arise, in part, from the nature of tourism as an economic activity. Although tourism is often referred to as an industry, it is fundamentally different than conventional industries; it is these differences that complicate the measurement of tourism (the definition of "tourism" and the nature of a "tourism industry" are discussed below). Further, the development of tourism statistics consistent among nations has required extensive negotiations among national statistical agencies as well as other international organizations to reach a consensus on the definition of tourism and associated concepts.

These concepts have been operationalized through new analytical tools, particularly the Tourism Satellite Account (UNWTO 1999). International agreement on core definitions and measurement techniques has now been achieved in principle. The tasks facing tourism statisticians are to refine, apply, and extend the concepts and tools that have been developed.

Fundamental to tourism statistics is, of course, the definition of "tourism." The World Tourism Organization defines tourism as the set of activities engaged in by persons temporarily away from their usual environment for a period of not more than one year, and for a broad range of leisure, business, religious, health, and personal reasons, excluding the pursuit of remuneration from within the place visited or long-term change of residence (UNWTO 1994). Thus, tourism fundamentally is something people do in certain circumstances (particularly travel outside their usual environment), not a commodity businesses produce.

There are several related concepts that are important for tourism policy, planning, marketing, and measurement purposes. One of these is *tourism commodity* – a good or service that would be produced only in a substantially reduced volume in the absence of tourism (such as passenger air services). A *tourism industry* is an industry characterized by the production of a tourism commodity (such as an airline offering scheduled passenger service). Thus, while tourism, *per se*, is not an industry, there are tourism industries such as accommodation, passenger transportation, food service, and recreation and entertainment.

Core tourism statistics include measures of the number of visitor arrivals in a destination (annually, seasonally, and/or monthly), their spending levels (often by category of commodity purchased), numbers of businesses serving visitors (by tourism industry), numbers of tourism employees, tourism's contribution to GDP, and government revenues attributable to tourism. Many specialized statistics related to persons engaged in tourism trips are also collected such as mode(s) of travel on a trip, mode(s) of accommodation used on a trip, activities engaged in during a trip, information sources used in planning a trip, routes taken, specific destinations visited, levels of satisfaction with services consumer, and so on.

Statistics related to activities not directly associated with individual behavior on specific trips are normally not considered to be tourism statistics, even though such information may be important for other purposes. Thus, government spending on infrastructure or tourism marketing, and investment in real estate or equipment (hotels, casinos, aircraft) are not considered to be within the scope of tourism statistics because they related to forms of production, and are more properly viewed as data relating to construction, manufacturing, marketing, real estate, and other forms of economic activity.

Sources of tourism statistics are numerous and diverse. They include surveys of border-crossing counts, visitors (during a trip or afterwards), business surveys, general social surveys (especially those covering household expenditures), and administrative records such as attraction ticket sales or hotel reservation records.

About the Author

Dr. Stephen L.J. Smith is Professor in the Department of Recreation and Leisure Studies and Director of the Tourism Policy and Planning Program at the University of Waterloo, Canada. He is an Elected Fellow of the International Statistical Institute (1994) and of the International Academy for the Study of Tourism (1991). He has authored more than 200 papers and eight books. His two most recent books are *Practical Tourism Research* (published by CABI, 2010) and *The Discovery of Tourism* (published by Emerald Publishing Group, 2010). He was involved in the creation of the Canadian Tourism Satellite Account through his leadership in the Canadian National Task Force on Tourism Data. Dr. Smith is Associate Editor for *Tourism Recreation Research* and the book review editor for *Annals of Tourism Research*.

Cross References

►Economic Statistics
►Seasonality
►Statistical Fallacies

References and Further Reading

Smith SLJ (2004) The measurement of global tourism: Old debates, new consensus, and continuing challenges. In: Lew AA, Hall CM, Williams AM (eds) A companion to tourism. Blackwell, Oxford, UK, pp 25–35

UNWTO (1994) Guidelines for tourism statistics. UNWTO, Madrid, Spain

UNWTO (1999) Tourism satellite account (TS): The conceptual framework. UNWTO, Madrid, Spain

Trend Estimation

Tommaso Proietti
Professor of Economic Statistics
University of Rome "Tor Vergata", Rome, Italy

Trend estimation deals with the characterization of the underlying, or long–run, evolution of a time series. Despite being a very pervasive theme in time series analysis since its inception, it still raises a lot of controversies. The difficulties, or better, the challenges, lie in the identification of the sources of the trend dynamics, and in the definition of the time horizon which defines the long run. The prevalent view in the literature considers the trend as a genuinely latent component, i.e., as the component of the evolution of a series that is persistent and cannot be ascribed to observable factors. As a matter of fact, the univariate approaches reviewed here assume that the trend is either a deterministic or random function of time.

A variety of approaches is available, which can be classified as nonparametric (kernel methods, local polynomial regression, band-pass filters, and wavelet multiresolution analysis), semiparametric (splines and Gaussian random fields) and parametric, when the trend is modeled as a

stochastic process. They will be discussed with respect to the additive decomposition of a time series $y(t)$:

$$y(t) = \mu(t) + \epsilon(t), \quad t = 1, \ldots, n, \tag{1}$$

where $\mu(t)$ is the trend component, and $\epsilon(t)$ is the noise, or irregular, component. We assume throughout that $\epsilon(t) = 0$ is a zero mean stationary process, whereas $\mu(t)$ can be a random or deterministic function of time. The above decomposition bears different meanings in different fields. In experimental sciences $\epsilon(t)$ is usually interpreted as a pure measurement error, so that a signal is observed with superimposed random noise. However, in behavioral sciences such has economics, quite often $\epsilon(t)$ is interpreted as a stationary stochastic cycle or as the transitory component of $y(t)$. The underlying idea is that trends and cycles can be ascribed to different economic mechanisms. Moreover, according to some approaches $\mu(t)$ is an underlying deterministic function of time, whereas for other it is a random function (e.g., a random walk, or a Gaussian process), although this distinction becomes more blurred in the case of splines. For some methods, like band pass filtering, the underlying true value $\mu(t)$ is defined by the analyst via the choice of a cutoff frequency which determines the time horizon for the trend.

The simplest and historically oldest approach to trend estimation adopted a global polynomial model for μ_t: $\mu(t) = \sum_{j=0}^{p} \beta_j t^j$. The statistical treatment, based on least squares, is provided in Anderson (1971). It turns out that global polynomials are amenable to mathematical treatment, but are not very flexible: they can provide bad local approximations and behave rather weirdly at the beginning and at the end of the sample period, which is inconvenient for forecasting purposes. More up to date methodologies make the representation more flexible either assuming that certain features, like the coefficients or the derivatives, evolve over time, or that a low order polynomial representation is adequate only as a local approximation.

Local polynomial regression (LPR) is a nonparametric approach that assumes that $\mu(t)$ is a smooth but unknown deterministic function of time, which can be approximated in a neighborhood of time t by a polynomial of degree p of the time distance with time t. The polynomial is fitted by locally weighted least squares, and the weighting function is known as the kernel. LPR generates linear signal extraction filters (also known as moving average filters) whose properties depend on three key ingredients: the order of the approximating polynomial, the size of the neighborhood, also known as the bandwidth, and the choice of the kernel function. The simplest example is the arithmetic moving average $m_t = \frac{1}{2h+1} \sum_{j=-h}^{h} y_{t+j}$, which is the LPR estimator of a local linear trend ($p = 1$) in discrete time using a bandwidth of $2h + 1$ consecutive observations and the uniform kernel.

Trend filters that arise from fitting a locally weighted polynomial to a time series have a well established tradition in time series analysis and signal extraction; see Kendall et al. (1983) and Loader (1999). For instance, the Maculay's moving average filters and the Henderson (1916) filters are integral part of the X-12 seasonal adjustment procedure adopted by the US Census Bureau.
The methodology further encompasses the Nayadara-Watson kernel smoother.

An important class of nonparametric filters arises from the frequency domain notion of a band-pass filter, that is popular in engineering. An ideal low-pass filter retains only the low frequency fluctuations in the series and reduces the amplitude of fluctuations with frequencies higher than a cutoff frequency ω_c. Such a filter is available analytically, but unfeasible, since it requires a doubly infinite sequence of observations; however, it can be approximated using various strategies (see Percival and Walden 1993). Wavelet multiresolution analysis provides a systematic way of performing band-pass filtering.

An alternative way of overcoming the limitations of the global polynomial model is to add polynomial pieces at given points, called knots, so that the polynomial sections are joined together ensuring that certain continuity properties are fulfilled. Given the set of points $t_1 < \ldots < t_i < \ldots t_k$, a polynomial spline function of degree p with k knots $t_1, \ldots, t_k$ is a polynomial of degree p in each of the $k + 1$ intervals $[t_i, t_{i+1})$, with $p - 2$ continuous derivatives, whereas the $p-1$-st derivative has jumps at the knots. It can be represented as follows:

$$\mu(t) = \beta_0 + \beta_1 (t - t_1) + \cdots + \beta_p (t - t_1)^p + \sum_{i=1}^{k} \eta_i (t - t_i)_+^p, \tag{2}$$

where the set of functions

$$(t - t_i)_+^p = \begin{cases} (t - t_i)^p, & t - t_i \geq 0, \\ 0, & t - t_i < 0 \end{cases}$$

defines what is usually called the truncated power basis of degree p.

According to (2) the spline is a linear combination of polynomial pieces; at each knot a new polynomial piece, starting off at zero, is added so that the derivatives at that point are continuous up to the order $p-2$. The most popular special case arises for $p = 3$ (cubic spline); the additional *natural boundary conditions*, which constrain the

spline to be linear outside the boundary knots, is imposed. See Green and Silverman (1994) and Ruppert et al. (2003).

An important class of semiparametric and parametric time series models are encompassed by (2). The piecewise nature of the spline "reflects the occurrence of structural change" (Poirier 1973). The knot t_i is the timing of a structural break. The change is "smooth," since certain continuity conditions are ensured. The coefficients η_i, which regulate the size of the break, may be considered as fixed or random. In the latter case $\mu(t)$ is a stochastic process, η_i is interpreted as a *random shock* that drives the evolution of $\mu(t)$, whereas the truncated power function $(t-t_i)_+^p$ describes its *impulse response function*, that is the impact on the future values of the trend.

If the η_i's are considered as random, the spline model can be formulated as a ▶linear mixed model, which is a traditional regression model extended so as to incorporate random effects. Denoting $\mathbf{y} = [y(t_1), \ldots, y(t_n)]'$, $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]'$, $\boldsymbol{\epsilon} = [\epsilon(t_1), \ldots, \epsilon(t_n)]'$, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta}$,

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}, \tag{3}$$

where the t-th row of $\mathbf{X}$ is $[1, (t-1), \ldots, (t-1)^p]$, and $\mathbf{Z}$ is a known matrix whose i-th column contains the impulse response signature of the shock η_i, $(t-t_i)_+^p$.

The trend is usually fitted by penalized least squares (PLS), which chooses $\boldsymbol{\mu}$ so as to minimize

$$(\mathbf{y}-\boldsymbol{\mu})'(\mathbf{y}-\boldsymbol{\mu}) + \lambda \int \left[\frac{d^{p-1}\mu(t)}{dt^{p-1}}\right]^2 dt, \tag{4}$$

where $\lambda \geq 0$ is the smoothness parameter.

PLS is among the most popular criteria for designing filters that has a long and well established tradition in actuarial sciences and economics (see Whittaker 1923, Leser 1961, and, more recently, Hodrick and Prescott 1997). Under Gaussian independent measurement noise minimizing the PLS criterion amounts to finding the conditional mode of μ given $\mathbf{y}$, This is a solution to the smoothing problem. If $\mu(t)$ is random, the minimum mean square estimator of the signal is $\mathrm{E}(\mu(t)|\mathbf{y})$. If the model (1) is Gaussian, these inferences are linear in the observations. The computations are carried out efficiently by the Kalman filter and the associated smoother (see Wecker and Ansley 1983).

The linear mixed model representation (3) encompasses other approaches, according to which the component $\mathbf{Z}\boldsymbol{\eta}$ is a Gaussian random process (Rasmussen and Williams 2006), or a (possibly nonstationary) time series process with a Markovian representation, such as in the structural time series approach see Harvey (1989), and in the canonical decomposition of time series (see Hillmer and Tiao 1982). The Markovian nature of the opens the way to the statistical treatment by the state space methodology and signal extraction is carried out efficiently by the Kalman filter and smoother. Popular predictors, such as exponential smoothing and Holt and Winters, arise as special cases (see Harvey 1989). The representation theory for the estimator of the trend component, Wiener-Kolmogorov filter, is established in Whittle (1983).

The analysis of economic time series has contributed to trend estimation in several ways. The first contribution is the attempt to relate the trend to a particular economic mechanism. The issue at stake is whether $\mu(t)$ is better characterized as a deterministic or stochastic trends. This problem was addressed in a very influential paper by Nelson and Plosser (1982), who adopted the (augmented) Dickey Fuller test for testing the hypothesis that the series is integrated of order 1, I(1), implying that $y(t) - y(t-1)$ is a stationary process versus the alternative that it is trend-stationary, e.g., $m(t) = \beta_0 + \beta_1 t$. Using a set of annual U.S. macroeconomic time series they are unable to reject the null for most series and discuss the implications for economic interpretation. The trend in economic aggregate is the cumulative effect of supply shocks, i.e., shocks to technology that occur randomly and propagate through the economic system via a persistent transmission mechanism.

A fundamental contribution is the notion of cointegration (Engle and Granger 1987), according to which two or more series are cointegrated if they are themselves nonstationary (e.g., integrated of order 1), but a linear combination of them is stationary. Cointegration results from the presence of a long run equilibrium relationship among the series, so that the same random trends drive the nonstationary dynamics of the series; also, part of the short run dynamics are also due to the adjustment to the equilibrium.

A third contribution, related to trend estimation, is the notion of spurious cycles that may result from inappropriate detrending of a nonstationary time series. This effect is known as the Slutzky–Yule effect, and concerned with the fact that an ad hoc filter to a purely random series can introduce artificial cycles.

Finally, large dimensional dynamic factor models have become increasingly popular in empirical macroeconomics. The essential idea is that the precision by which the common components are estimated can be increased by bringing in more information from related series: suppose for simplicity that $y_i(t) = \theta_i \mu(t) + \epsilon_i(t)$, where the i-th series, $i = 1, \ldots, N$, depends on the same stationary common factor, which is responsible for the observed comovements of economic time series, plus an idiosyncratic component, which includes measurement error and local

shocks. Generally, multivariate methods provide more reliable measurements provided that a set of related series can be viewed as repeated measures of the same underlying latent variable. Stock and Watson (2002) and Forni et al. (2000) discuss the conditions on μ_t and ϵ_{it} under which dynamic or static principal components yield consistent estimates of the underlying factor μ_t as both N and the number of time series observations tend to infinity.

About the Author

Professor Proietti is Associate Editor of *Computational Statistics and Data Analysis* (Elsevier), and Co-Editor of *Statistical Methods and Applications* (Springer) He is Editor (with A.C. Harvey) of the text: *Readings in Unobserved Components Models* (Advanced Texts in Econometrics, Oxford University Press, 2005).

Cross References

►Business Forecasting Methods
►Detection of Turning Points in Business Cycles
►Dickey-Fuller Tests
►Exponential and Holt-Winters Smoothing
►Forecasting Principles
►Linear Mixed Models
►Moving Averages
►Nonparametric Estimation
►Nonparametric Regression Using Kernel and Spline Methods
►Seasonal Integration and Cointegration in Economic Time Series
►Structural Time Series Models
►Time Series

References and Further Reading

Anderson TW (1971) The statistical analysis of time series. Wiley, New York
Engle RF, Granger CWJ (1987) Co-integration and error correction: representation, estimation, and testing. Econometrica 55: 251–276
Forni M, Hallin M, Lippi F, Reichlin L (2000) The generalized dynamic factor model: identification and estimation. Rev Econ Stat 82:540–554
Green PJ, Silverman BV (1994) Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman & Hall, London
Harvey AC (1989) Forecasting, structural time series and the Kalman filter. Cambridge University Press, Cambridge
Henderson R (1916) Note on graduation by adjusted average. Trans Actuarial Soc America 17:43–48
Hillmer SC, Tiao GC (1982) An ARIMA-model-based approach to seasonal adjustment. J Am Stat Assoc 77:63–70
Hodrick R, Prescott EC (1997) Postwar U.S. business cycle: an empirical investigation. J Money Credit Bank 29(1):1–16
Kendall M, Stuart A, Ord JK (1983) The advanced theory of statistics, vol 3. Charles Griffin, London
Leser CEV (1961) A simple method of trend construction. J Roy Stat Soc B 23:91–107
Loader C (1999) Local regression and likelihood. Springer-Verlag, New York
Nelson CR, Plosser CI (1982) Trends and random walks in macroeconomic time series: some evidence and implications. J Monet Econ 10:139–162
Percival D, Walden A (1993) Spectral analysis for physical applications. Cambridge University Press, Cambridge
Poirier DJ (1973) Piecewise regression using cubic splines. J Am Stat Assoc 68:515–524
Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, Cambridge
Ruppert D, Wand MJ, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge
Stock JH, Watson MW (2002b) Forecasting using principal components from a large number of predictors. J Am Stat Assoc 97:1167–1179
Watson GS (1964) Smooth regression analysis. Shankya Series A, 26:359–372
Wecker WE, Ansley CF (1983) The signal extraction approach to nonlinear regression and spline smoothing. J Am Stat Assoc 78:81–89
Whittaker E (1923) On new method of graduation. Proc Edinburgh Math Soc 41:63–75
Whittle P (1983) Prediction and regulation by linear least squares methods, 2nd edn. Basil Blackwell, Oxford

Two-Stage Least Squares

Roberto S. Mariano
Dean and Professor of Economics and Statistics
Singapore Management University, Singapore, Singapore

In the linear regression model, $y = X_1\beta_1 + Y_1\beta_2 + u = Z\beta + u$, there are real-life situations when some of the regressors, denoted by Y_1 in the model, are correlated with the disturbance term. The vector and matrices y, X_1, and Y_1 are $N \times 1, N \times K_1$, and $N \times (G-1)$ data matrices from a sample of size N. u is the $N \times 1$ vector of disturbances, assumed to have mean zero and variance-covariance matrix $\sigma^2 I$. In this model, X_1 is assumed to be statistically independent of the disturbance term and the analysis is done conditional on X_1.

In such situations where correlation between error and regressor exists, ordinary least squares (*OLS*) estimates of regression coefficients become not only biased but also inconsistent (as sample size increases indefinitely). One of the earlier efforts to correct for this inconsistency is a

two step procedure called two-stage least squares in the econometric literature. The procedure first regresses the "disturbance-correlated" variables, Y_1, on a selected set of first-stage regressors (X) and obtains the calculated regression values $P_X(Y_1) = X(X'X)^{-1}X'Y_1$, the projection of Y_1 on the column space of X. For the second stage of the procedure, y is then regressed on X_1 and $P_X(Y_1)$ to obtain the $2SLS$ estimate b_{2SLS}. Typically $X = (X_1, X_2)$ where X_2 is $N \times K_2$, and is independent of u, and X has full column rank equal to at least $K_1 + G - 1$. Intuitively, the first-stage regression serves to "purge" Y_1 of its component that is correlated with u and this leads to consistency in the regression at the second stage where Y_1 is replaced by $P_X(Y_1)$.

$2SLS$ was developed in the econometric literature in dealing with the estimation of the linear regression model (see ►Linear Regression Models) as part of a simultaneous equations system. In this context, the joint probability distribution of y and Y_1 is specified and X is determined from the model.

$2SLS$ appeared in an earlier form as an intermediate step in the iteration towards the calculation of the limited-information-maximum-likelihood $(LIML)$ estimator in simultaneous equation models. The $2SLS$ estimator in the linear regression model also can be interpreted as an instrumental variable (IV) estimator, using the instrument matrix $W_{2SLS} = P_X Z$ for Z; that is,

$$b_{IV} = \left(W'_{2SLS}Z\right)^{-1} W'_{2SLS}y = (Z'P_XZ)^{-1}Z'P_Xy = b_{2SLS}.$$

The $2SLS$ estimator also can be interpreted as a generalized least squares (GLS) estimator in the derived linear model $X'y = X'Z\beta + X'u$.

When Z has a large dimension, modified two-stage least squares has been suggested as an alternative approach. This is also a two-step regression procedure where the first stage of $2SLS$ is modified by regressing Y_1 on H, a $N \times h$ submatrix spanning a column subspace of X. H is chosen to be of full column rank and rank $[(I - P_1)H] \geq G - 1$, where P_1 is the projection matrix on the column space of X_1. One suggested manner of constructing H is to start with X_1 and then add at least $G-1$ of the remaining columns of X or the first K_2 principal components of $(I - P_1)X_2$. In this case, the modified $2SLS$ is exactly equivalent to the IV estimator using $(X_1, P_H Y_1)$ as the instrument matrix.

Ordinary least squares also can be interpreted as an IV estimator with Z as the instrument for itself. Another variation of an IV estimator that has been suggested is Theil's k-class estimator. This uses as its instrument matrix a linear combination of the instrument matrices for OLS and $2SLS$ and k is chosen by the investigator and can be stochastic or non-stochastic. Thus, with $W_{(k)} = kW_{2SLS} + (1-k)W_{(OLS)} = kP_XZ + (1-k)Z$, the k-class estimator is

$$b_{(k)} = \left(W'_{(k)}Z\right)^{-1} W'_{(k)}y = \beta + \left(W'_{(k)}Z\right)^{-1} W'_{(k)}u,$$

Assuming that $\text{plim}(k)$ is finite, a necessary and sufficient condition for consistency of the k-class estimator is $\text{plim}(1-k) = 0$ – that is, the contribution of the OLS instrument matrix dies out in the limit.

The limited information maximum likelihood $(LIML)$ estimator is closely related to the $2SLS$ and other estimators introduced here and is a member of the k-class of estimators. Think of the linear regression equation introduced above as part of a complete simultaneous-equations model for the joint stochastic behavior of y, Y_1, and other dependent variables showing up in other equations of the model. The $LIML$ estimator of β maximizes the likelihood of (y, Y_1) subject to any identifiability restrictions, and is called limited in the sense that it ignores the dependent variables that do not show up in the regression equation. The constrained maximization process in $LIML$ reduces to minimizing the following variance ratio with respect to $\beta^* = (1, \beta')'$

$$v = (\beta^{*\prime}A\beta^*)/(\beta^{*\prime}S\beta^*) = 1 + (\beta^{*\prime}W\beta^*)/(\beta^{*\prime}S\beta^*),$$

where $Y = (y, Y_1)$; $S = Y'(I - P_X)Y$; $W = Y'(P_X - P_1)$; and $A = S + W = Y'(I - P_1)Y$.

This minimization problem yields the solution b_{LIML} as a characteristic vector of A with respect to S corresponding to the smallest root h of $\det(A - vS) = 0$, and

$$h = \left(b^*_{LIML}{}'Ab^*_{LIML}\right) / \left(b^*_{LIML}{}'Sb^*_{LIML}\right).$$

Note that $(\beta^{*\prime}W\beta^*)$ is the marginal regression sum of squares due to X_2 given X_1 in the regression of $Y\beta^*$ on X, while $\beta^{*\prime}S\beta^*/(N-K)$ provides an unbiased estimator of the error variance in the regression equation. Thus $LIML$ minimizes the marginal contribution of X_2 given X_1 relative to an estimate of the error variance. $2SLS$ simply minimizes this marginal contribution in absolute terms.

The $LIML$ estimator b^*_{LIML} needs to be normalized to have a unit value in its first element, to be comparable with the other estimators we have discussed so far. With such a normalization, the $LIML$ estimator of β_1 and β_2 turns out to be a k-class estimator as well, where the value of k is h, the smallest root of $\det(A - vS) = 0$. Note that $h = 1 + f$, where f is the smallest root of $\det(W - vS) = 0$. Thus, $LIML$ is an IV estimator also, whose instrument matrix is a linear combination of the OLS and $2SLS$ instrument matrices, with k stochastic and k at least equal to unity.

Two-stage least squares and the other estimators discussed above have been analyzed for statistical properties in small samples, under the standard large-sample asymptotics, and in alternative nonstandard asymptotic settings such as error variances going to zero, number of instruments going to infinity at the same rate as sample size, and so-called weak instrument asymptotics.

About the Author

Roberto S. Mariano received his PhD in Statistics in 1970, Stanford University. Currently, he is Professor of Economics and Statistics and Dean, School of Economics, Singapore Management University. He is also Director, Sim Kee Boon Institute for Financial Economics, and Co-Director, Center for Financial Econometrics. Before joining the Singapore Management University he was Professor of Economics and Statistics, Department of Economics, University of Pennsylvania (1980–2004). Dr Mariano is a Fellow, Econometric Society (2009–present), and a Fellow, Wharton Financial Institutions Center, Wharton School (2004–present). He has authored numerous research papers and books on econometric methodology and applications and has served on the editorial board of several international professional journals in economics and statistics. He was the principal investigator in research projects funded by the United Nations, US National Science Foundation, the Rockefeller Foundation, the US Department of Commerce and the US Department of Agriculture. In Singapore, he worked on a research project for the Ministry of Manpower entitled "Macroeconometric Sectoral Modeling for Manpower Planning in Singapore" from March 2000–March 2002 where he was the principal investigator with Nobel Laureate Lawrence R. Klein.

Cross References

►Econometrics
►Instrumental Variables
►Least Squares
►Linear Regression Models
►Method Comparison Studies
►Properties of Estimators

References and Further Reading

Anderson TW, Rubin H (1949) Estimation of the parameters of a single equation in a complete system of stochastic equations. Ann Math Stat 20:46–63

Basman RL (1957) A generalized classical method of linear estimation of coefficients in a structural equation. Econometrica 25:77–83

Bound J, Jaeger DA, Baker RM (1995) Problems with instrumental variable estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J Am Stat Assoc 90:443–450

Mariano RS (1977) Finite-sample properties of instrumental variable estimators of structural coefficients. Econometrica 45:487–496

Mariano RS (2001) Chapter 6: Simultaneous equation model estimators: statistical properties and practical implications. In: Baltagi B (ed) Companion to theoretical econometrics. Blackwell Publishers, Oxford, pp 122–143

Phillips PCB (1983) Exact small sample theory in the simultaneous equations model. In: The handbook of econometrics, Volume II, Elsevier Science, North Holland, Amsterdam, pp 881–935

Theil H (1953) Repeated least squares applied to complete equation systems. Central Planning Bureau mimeograph, The Hague

U

Unbiased Estimators and Their Applications

Mikhail Nikulin[1], Vassiliy Voinov[2]
[1]Professor
University Victor Segalen, Bordeaux, France
[2]Professor
KIMEP, Almaty, Kazakhstan

Probability models and statistical inferential methods are widely used in the study of various physical, chemical, engineering, biological, medical, social, and other phenomena. As a rule, these models depend on unknown parameters, the values of which are to be estimated. Methods of mathematical statistics and, in particular, methods of statistical estimation of parametric functions are mainly used in processing the results of experiments. The theory of unbiased estimation plays a very important role in the theory of point estimation, since in many real situations it is of importance to obtain the unbiased estimator that will have no systematical errors (see, e.g., Fisher (1925), Stigler (1977)). The problem of unbiased estimation attracted the attention of famous statisticians in the late 1940's: Neyman (1944), Cramér (1946), Kolmogorov (1950) , Halmos (1946), Lehmann (1983), Rao (1949), etc. A great amount of work has been carried out in this field up to the present time: an elegant theory of unbiased estimation based on the theory of sufficient statistics has been constructed, techniques for constructing the best unbiased estimators have been well developed and a great number of theoretical and applied problems have been solved (see Rao (1965), Zacks (1971), Voinov and Nikulin (1993, 1996)). Unbiased in the mean or simply unbiased estimator is a statistic, the mathematical expectation of which equals the quantity to be estimated.

Suppose that, using the realization of a random variable X that takes values in a sample space $(\mathcal{X}, \mathcal{B}, P_\theta, \theta\in\Theta)$, a parametric function $f : \Theta \to \mathcal{Y}$ that maps the parametric space Θ into a certain set $\mathcal{Y}$ has to be estimated. Suppose that such an estimator $T = T(X)$ of $f(\theta)$, $T : \mathcal{X} \to \mathcal{Y}$, has been constructed. If the statistic T is such that the unbiasedness equation

$$\mathbf{E}_\theta T = \int_{\mathcal{X}} T(x)dP_\theta(x) = f(\theta), \quad \theta \in \Theta,$$

holds, then T is called an unbiased in the mean or simply unbiased estimator for $f(\theta)$. Median and mode unbiased estimators can also be considered (see Voinov and Nikulin (1993)) but they have much less applications compared to unbiased in the mean ones.

Example 1 Let $X = (X_1, \ldots, X_n)$ be a sample of size n, i.e., $X_1, \ldots, X_n$ are independent identically distributed random variables. If $\mathbf{E}X_1^2$ exists, then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

will be the unbiased estimators of the mean $\mathbf{E}X_1$ and the variance $\mathbf{Var}X_1$ of X_1.

Example 2 Let $X = (X_1, \ldots, X_n)$ be a sample. Suppose that the distribution function F of the random variable X_1 is unknown. We have a non-parametric problem, since in continuous case the parameter F is infinite-dimensional, $F \in \mathcal{F} = \Theta$, where $\mathcal{F}$ is space of a distribution function. Consider the statistic

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le x\}}, \quad x \in R^1.$$

$F_n(\cdot)$ is the empirical distribution function based on the sample $X = (X_1, \ldots, X_n)$. The statistic $F_n(x)$ is an unbiased estimator for the unknown distribution function $F(x)$, since for any $x \in R^1$

$$\mathbf{E}_F F_n(x) = F(x), \quad F \in \mathcal{F}.$$

Example 3 Let X be a random variable following the geometric distribution with parameter of succes θ, i.e., for any $k = 1, 2, 3, \ldots$.

$$\mathbf{P}_\theta\{X = k\} = \theta(1-\theta)^{k-1}, \quad 0 \le \theta \le 1.$$

The unique solution of the unbiasedness equation

$$\sum_{k=1}^{\infty} T(k)\theta(1-\theta)^{k-1} = \theta, \quad \theta \in \Theta = [0,1],$$

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

is

$$T(X) = \begin{cases} 1, & \text{if } X = 1, \\ 0, & \text{if } X \geq 2. \end{cases}$$

One sees that the statistic $T(x)$ is good only when θ is close to 0 or 1, otherwise $T(x)$ contains no useful information.

Example 4 Suppose that a random variable X possesses the discrete Pólya distribution with a parameter $\theta = (p, \lambda)^T$

$$\mathbf{P}_\theta\{X = x\} = \binom{n}{x} \frac{p^{[x;\lambda]}(1-p)^{[n-x;\lambda]}}{1^{[n;\lambda]}};$$
$$x = 0, 1, \ldots, n; \quad 0 < p < 1,$$

where λ is supposed to be known such that $p+\lambda(n-1) > 0$; $1 - p + \lambda(n-1) > 0$; and $a^{[r;\lambda]}$ is the generalized power of a defined by $a^{[r;\lambda]} = \prod_{h=0}^{r-1}(a + \lambda h)$; $a^{[0;\lambda]} = 1$. This distribution and its generalizations are often used in the quality control, (see Lumelskii et al. (2007)). If $\lambda = 0$, then the discrete Pólya distribution reduces to the ▶binomial distribution. For negative values of λ the Pólya distribution reduces to a hypergeometric probability distribution. The unbiased estimator $\hat{p}$ of the parameter p and the unbiased estimator $\hat{\mathbf{Var}}_\theta(\hat{p})$ of its variance $\mathbf{Var}_\theta(\hat{p}) = \frac{p(1-p)(1+n\lambda)}{n(1+\lambda)}$ are (Lumelskii et al. 2010):

$$\hat{p} = \frac{X}{n}, \quad \hat{\mathbf{Var}}_\theta(\hat{p}) = \frac{X(n-X)(1+n\lambda)}{n^2(n-1)}.$$

In assessing the properties of point statistical estimators a statistician concentrates his attention on three main features of their quality: consistency, unbiasedness and risk. Consistency is known to be the asymptotical property appearing only when the dimension of the vector of observation X tends to infinity. It is the risk of estimator, in particular, the quadratic risk that is the main characteristic of its quality for small n. For an unbiased estimator it coincides with its variance. So it is natural that from two unbiased estimators of the same parameter the best estimator will be that whose variance is smaller. From this point of view the Rao-Cramér information inequality plays a very important role, indicating the existence of the lower bound of a point estimator risk function with respect to the square loss function. The Rao-Cramér inequality has a very simple form for unbiased estimators. Namely, if $T = T(X)$ is an unbiased estimator for a function $f(\theta)$, i.e., $\mathbf{E}_\theta T = f(\theta)$, then under some regularity conditions on the family $\{P_\theta\}$ and function f, the Rao-Cramér inequality implies that

$$\mathbf{Var}_\theta T = \mathbf{E}_\theta(T - f(\theta))^2 \geq \frac{[f'(\theta)]^2}{I(\theta)}, \tag{1}$$

where $I(\theta)$ is the Fisher information on θ, contained in the observation of X. Thus in the right side of the inequality (1) one can see a lower bound $[f'(\theta)]^2/I(\theta)$ for the variance of an unbiased estimator T of $f(\theta)$. In particular , when $f(\theta) \equiv \theta$, from (1) it follows that

$$\mathbf{Var}_\theta T = \mathbf{E}_\theta(T - \theta)^2 \geq \frac{1}{I(\theta)}. \tag{2}$$

A statistical estimator, for which the equality is attained in (1) or (2), is called efficient. The most important problem of the theory of unbiased estimation is the construction of the efficient estimators, if it is possible. Note that the lower bound is not an exact lower bound, so a statistician has to search for the best unbiased estimators whose variances reaches the exact lower bound. These estimators are known also as the minimum variance unbiased estimators (MVUEs). In this context an important role is played by Rao-Blackwell-Kolmogorov theorem (see ▶Rao–Blackwell Theorem), which allows to construct an unbiased estimator of minimal variance. This theorem asserts that if the family $\{P_\theta\}$ has a sufficient statistic $U = U(X)$ and $T = T(X)$ is an arbitrary unbiased estimator of a function $f(\theta)$, then the statistic $T^* = \mathbf{E}\{T|U\}$, obtained by averaging T over the fixed sufficient statistic U, has a risk not exceeding that of T relative to any convex loss function for all $\theta \in \Theta$. If the family $\{P_\theta\}$ is complete, then the statistic T^* will be unique. From this theorem it follows that the best unbiased estimators must be constructed in terms of sufficient statistics. There exist tables of MVUEs of unknown parameters for more then 40 univariate and multivariate probability distributions (Voinov and Nikulin 1993, 1996).

Example 5 Let $X = (X_1, \ldots, X_n)$ be a normal $N(\mu, \sigma^2)$ sample. Denote $\theta = (\mu, \sigma^2)$. In this case maximum likelihood estimator $\hat{\theta} = \left(\bar{X}_n, s_n^2\right)$ for θ is the complete minimal sufficient statistic for the family of normal distributions, where $s_n^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2/n$ is the variance of the empirical distribution. Following to Kolmogorov (1950) it is easy to construct the best unbiased estimator $\Phi^*(x)$ for the distribution function $\Phi\left(\frac{x-\mu}{\sigma}\right)$ of the normal law $N(\mu, \sigma^2)$:

$$\Phi^*(x) = \mathbf{E}\left(F_n(x)|\bar{X}_n, s_n^2\right)$$
$$= S_{n-2}\left\{\frac{\sqrt{n-2}\left(\frac{x-\bar{X}_n}{s_n}\right)}{\sqrt{n-1-\left(\frac{x-\bar{X}_n}{s_n}\right)^2}}\right\}, \quad x \in R^1,$$

where $S_f(\cdot)$ is the Student distribution function with f degrees of freedom. It is the best unbiased estimator and it differs from the maximum likelihood estimator $\Phi\left(\frac{x-\bar{X}_n}{s_n}\right)$,

which is biased, and from another unbiased estimator $F_n(x)$.

To this end we would like to note that the theory of unbiased estimation, the theory of sufficiency, and the theory of constructing the best unbiased estimators are used today, e.g., for constructing the modified chi-squared type tests (Bol'shev and Mirvaliev (1978), Voinov and Nikulin (1993), Greenwood and Nikulin (1996), Chichagov (2006)), for constructing confidence intervals (Lumelskii et al. (2010)), etc.

About the Author

For biography *see* the entry ▶Chi-Squared Goodness-of-Fit Tests: Drawbacks and Improvements.

Cross References

▶Best Linear Unbiased Estimation in Linear Models
▶Cramér–Rao Inequality
▶Estimation
▶Estimation: An Overview
▶Minimum Variance Unbiased
▶Properties of Estimators
▶Rao–Blackwell Theorem
▶Sufficient Statistics

References and Further Reading

Bol'shev LN, Mirvaliev M (1978) Chi-square goodness-of-fit test for the Poisson, binomial, and negative binomial distributions. Theor Probab Appl 23:481–494

Blackwell D (1947) Conditional expectation and unbiased sequential estimation. Ann Math Stat 18:105–110

Chichagov VV (2006) Unbiased estimators and chi-squared statistics for one-parameter exponential family. In: Statistical methods of estimation and hypotheses testing, vol 19. Perm State University, Perm, Russia, pp 78–89

Cramér H (1946) Mathematical methods of statistics. Prinston University Press, Prinston

Fisher R (1925) Statistical methods for research workers, Olivier and Boyd, Edinburgh and London

Greenwood PE, Nikulin MS (1996) A guide to chi-squared testing. Wiley, NY

Halmos PR (1946) The theory of unbiased estimation. Ann Math Stat 17:34–43

Kolmogorov AN (1950) Unbiased estimators. Izvestia Acad Sci USSR, Ser Math 14(4):303–326

Lehmann EL (1983) Theory of point estimation. Wiley, NY

Lumelskii YA, Voinov VG, Nikulin MS, Feigin P (2007) On a generalization of the classical random sampling scheme and unbiased estimation. Comm Stat Theor Meth 36:693–705

Lumelskii YA, Voinov VG, Voinov EV, Nikulin MS (2010) Approximate confidence limits for a proportion of the Pólya distribution (Communication in statistics – theory and methods to appear)

Neyman J (1944) Statistical estimation as a problem of a classical theory of probability. Uspekhi Matematicheskikh Nauk 10: 207–229

Rao CR (1949) Sufficient statistics and minimum variance estimates. Proc Cambridge Phil Soc 45:213–218

Rao CR (1965) Linear statistical inferences and their applications. Wiley, New York

Stigler SM (1977) Do robust estimators work with real data? Ann Stat 5:1055–1098

Voinov VG, Nikulin MS (1993) Unbiased estimatores and their applications, vol 1: univariate case. Kluwer Academic Publishers, Dordrecht

Voinov VG, Nikulin MS (1996) Unbiased estimatores and their applications, vol 2: multivariate case. Kluwer Academic Publishers, Dordrecht

Zacks S (1971) The theory of statistical inference. Wiley, New York

Uniform Distribution in Statistics

Vesna Jevremović
Associate Professor, Faculty of Mathematics
University of Belgrade, Belgrade, Serbia

Uniform distribution, the simplest probability distribution, plays an important role in Statistics since it is indispensable in modeling random variables, and therefore in traditional and Quasi-Monte Carlo simulation. It is often used to represent the distribution of roundoff errors in values tabulated to the nearest k decimal places (Johnson et al. 1995). We can distinguish between the continuous and discrete uniform distribution.

Properties of the Uniform Distribution

The *continuous* random variable X is said to be uniformly distributed, or having rectangular distribution on the interval $[a,b]$, and we write $X : U(a,b)$, if its probability density function (p.d.f) equals $f(x) = \frac{1}{b-a}$, $x \in [a,b]$, and 0 elsewhere. It follows that the distribution function is $F(x) = \frac{x-a}{b-a}$, $x \in [a,b]$. The moments are $m_r = \frac{1}{r+1}\frac{b^{r+1}-a^{r+1}}{b-a}$, $r \in N$, while the central moments are $\mu_{2k-1} = 0$, $\mu_{2k} = \frac{1}{2k+1}\left(\frac{b-a}{2}\right)^{2k}$, $k \in N$ (Djorić et al. 2007). The distribution mode is not unique; the median is obviously $(a+b)/2$ because of the symmetry of the uniform distribution. From the symmetry it follows also that the Pearson coefficient of skewness is 0, while the coefficient of kurtosis is −1.2.

The uniform distribution is a special case of the ▶beta distribution. Namely, the uniform distribution $U(0,1)$ is $B_2(1,1)$ distribution. Moreover, if $(X_1,\ldots,X_n)$ is a random sample from the $U(0,1)$ distribution, then the kth order statistic $X_{(k)}$ has the beta distribution $B_2(k,n-k+1)$.

Unlike most continuous distributions, the uniform distribution has a discrete counterpart. A random variable X that assumes a finite number of distinct values $x_1, \ldots, x_n$ each with the same probability

$$P(X = x_j) = \frac{1}{n} \quad j = 1, \ldots, n,$$

where n is a positive integer, is called a *discrete uniform* distribution.

The relationship between the continuous uniform distribution $U(0,1)$ and the discrete uniform distribution $Y : \begin{pmatrix} 0 & 1 & 2 & \cdots & 9 \\ 0.1 & 0.1 & 0.1 & \cdots & 0.1 \end{pmatrix}$ is given by the following theorem (Соболь 1973).

Let $\gamma = 0.u_1u_2 \ldots u_n \ldots$ be a realization of a random variable $X : U(0,1)$. Then $u_1, u_2, \ldots, u_n, \ldots$ are independent realizations of a discrete uniform distribution Y, and vice versa.

If $X : U(a,b)$, where a and/or b are unknown, we may estimate them using a sample $(X_1, \ldots, X_n)$ from this distribution. Since the uniform distribution is not regular in the Rao-Cramér sense, the estimators based on the maximum likelihood method have some distinctive, interesting properties. We will provide some examples, "where standard frequentist inference procedures are not applicable" (Rohde 2007).

Let us first analyze the case of $U(0,b)$. Maximum likelihood method (ML) yields the estimator $\hat{b} = \max\limits_{1\le j\le n} X_j = X_{(n)}$, and the method of moments (MM) gives $\overline{b}_n = 2\overline{X}_n$. We have:

$$E(\hat{b}) = nb/(n+1), \quad Var(\hat{b}) = nb^2/(n+2)(n+1)^2$$
$$E(\overline{b}_n) = b, \quad Var(\overline{b}_n) = b^2/3n.$$

We can see that the ML estimator is biased but has smaller variance than the unbiased MM estimator. The estimator $\frac{n+1}{n}\hat{b}$ is unbiased and its variance is still smaller than the variance of the MM estimator (see Larsen and Marx 2006, p. 391). In addition, the variance of this estimator is less than the lower bound of variance from the Rao-Cramér inequality, which is b^2/n. Furthermore, the statistic $\hat{b} = \max\limits_{1\le j\le n} X_j = X_{(n)}$ is a complete sufficient statistic for the parameter b; it is also consistent and even squared-error consistent since $\lim\limits_{n\to\infty} E(\hat{b} - b)^2 = 0$. Using ▶order statistics we can find other unbiased estimators such as $\hat{b}_1 = (n+1)\min\limits_{1\le j\le n} X_j = (n+1)X_{(1)}$, or, in the case of odd n, $n = 2k+1$, the estimator based on the sample median $\hat{b}_2 = 2X_{(k+1)}$.

In the case of $U(a,0)$ distribution, the ML estimator for a is $\hat{a} = \min\limits_{1\le j\le n} X_j = X_{(1)}$.

As a second example we take the distribution $X : U(a-1/2, a+1/2)$, for which the ML estimator is not unique, and every statistic V satisfying the inequality $\max\limits_{1\le j\le n} X_j - \frac{1}{2} \le V \le \min\limits_{1\le j\le n} X_j + \frac{1}{2}$ could be taken as an ML estimator for a. For more discussion on this distribution, see Dexter and Hogg (2000) and Romano and Siegel (1986, p. 182).

Finally in the case of the family of distributions $X : U(-a,a)$, $a > 0$, we have to say that this family is not complete, and the joint sufficient statistic for a is $(X_{(1)}, X_{(n)})$, while the ML estimator $\hat{a} = \max(-X_{(1)}, X_{(n)})$ is a minimal sufficient statistic for this parameter. More examples could be found in Hogg et al. (2005) and Larsen and Marx (2006).

Uniform Distribution and Modeling of Random Variables

Modeling random variables is the first step in Monte Carlo methods (see ▶Monte Carlo Methods in Statistics), and for this step the uniform distribution is necessary. Methods for modeling random variables (simulation) vary depending on the nature of random variables, but they all use modeled values of the uniform distribution $U(0,1)$. A single value of the uniform distribution $U(0,1)$ is referred to as "random number," and will be denoted γ. To be more precise, Monte Carlo methods use pseudo-random numbers, i.e., series of numbers from the interval $(0,1)$ having statistical properties of the random sample from the uniform distribution. These numbers are generated by computers using some appropriate devices or formulae, as it is explained in [6] and in ▶Uniform Random Number Generators. The accuracy of the Monte Carlo methods generally improves with an increase of pseudo-random numbers used.

When modeling random variables, the next two properties of the uniform distribution are often used: (1) If $X : U(0,1)$, then $1 - X : U(0,1)$ and (2) If the random variable X has the distribution function $F(x)$, then random variable $F(X)$ is uniformly distributed on the interval $[0,1]$, i.e., $F(X) : U(0,1)$.

The basic ideas and methods for modeling random variables are as follows.

(a) *Discrete r.v.* Let X be a discrete random variable (r.v.) with a *finite* set of values and with distribution $P(X = x_j) = p_j$, $j = 1, \ldots, n$ where $\sum\limits_{k=1}^{n} p_k = 1$. Let γ be one random number. If $\gamma \le p_1$, then we assume that the value x_1 of r.v. X is realized. If $\sum\limits_{j=1}^{k-1} p_j < \gamma \le \sum\limits_{j=1}^{k} p_j$, then we assume

that the value x_k of r.v. X is realized. In this way, for every realization of r.v. X one random number is used.

The same idea could be applied in modeling realizations of some random event using the corresponding indicator r.v.

This is the general approach for modeling discrete r.v., while there are many special solutions depending on the nature of r.v. to be modeled [6].

If X is a discrete r.v. with an *infinite* set of values, then in purpose of modeling r.v. X we use truncated r.v. X_Z with finite set of values such that $P(X \neq X_Z) = 1 - \delta$, where δ is an arbitrarily chosen small positive number. Realizations of this r.v. X_Z are modeled using the procedure previously described and are taken as realizations of the r.v. X.

(b) *Continuous r.v.* If X is a continuous r.v., then its realizations could be modeled using *the inversion method* (one random number per realization) or *the rejection method* (the number of random numbers used has the geometric distribution, and so could be infinite) or some special method as is the case with the normal distribution. The inverse function method cannot be applied with the normal distribution, while the rejection method can, using the corresponding truncated r.v., but is not usually used. One of the many procedures of modeling values for a normally distributed r.v. (Соболь 1973) is based on a central limit theorem (see ►Central Limit Theorems), and uses random numbers. Let $Y_1, Y_2, \ldots$ be independent r.v. with the uniform distribution $U(0,1)$. The sum $S_n = \sum_{j=1}^{n} Y_j$ has expectation and variance $E(S_n) = \frac{n}{2}$, $Var(S_n) = \frac{n}{12}$, and following the central limit theorem the r.v. $S_n^* = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} = \sqrt{\frac{3}{n}} \sum_{j=1}^{n} (2Y_j - 1)$ converges in distribution to the normally distributed r.v. $N(0,1)$.

The convergence is fast and with $n = 12$ the difference between S_n^* and $N(0,1)$ is small enough, so we can take $s_{12}^* = \sum_{j=1}^{12} y_j - 6$ as a realized value for r.v. with $N(0,1)$ distribution. In this way, one needs 12 random numbers y_j, $j = 1,12$ for one realization of r.v. with $N(0,1)$ distribution. It should be noticed that even $n = 6$ gives a good result (Соболь 1973).

More on simulation can be found in the entry ►Nonuniform Random Variate Generations.

Cross References

►Bivariate Distributions
►Cramér–Rao Inequality
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview
►Uniform Random Number Generators
►Univariate Discrete Distributions: An Overview

References and Further Reading

Соболь И (1973) *Численные методы Монте Карло*, Наука, Москва (in Russian)
Dexter CW, Hogg RV (2001) A little uniform density with big instructional potential. J Stat Educ 9(2)
Djorić D, Jevremović V et al (2007) Atlas raspodela. Gradjevinski fakultet, Beograd (in Serbian)
Hogg RV, McKean JW, Craig AT (2005) Introduction to mathematical statistics. Pearson Education International, Upper Saddle River
Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distributions, vol 2, 2nd edn. Wiley-Interscience, New York
Larsen RJ, Marx ML (2006) An introduction to mathematical statistics and its applications. Pearson International Edition International, Upper Saddle River
Rohde S (2007) Using the uniform distribution in teaching the foundations of statistics. ISI 56th session. Lisbon, Portugal
Romano JP, Siegel AF (1986) Counterexamples in probability and statistics. Chapman & Hall/CRC Press, New York

Uniform Experimental Design

Kai-Tai Fang
Professor, Director
Beijing Normal University-Hong Kong Baptist University
Zhuhai, China

In the past decades computer experiments or computer-based simulation have become a hot topic in statistics and engineering. The uniform experimental design (UD for short) proposed by Fang and Wang (Fang 1980; Wang and Fang 1981) is one of the space-filling designs for computer experiments and is also one of robust designs for experiments with model uncertainty. Suppose that there are s factors, $X_1, \ldots, X_s$, in an experiment. The experimental region of the factors is denoted by $\mathcal{T}$, very often, $\mathcal{T}$ is a rectangle $[a_1, b_1] \times \cdots \times [a_s, b_s]$. From now on we always assume $\mathcal{T}$ to be a rectangle. Without any generality we can assume that $\mathcal{T}$ is a unit cube, $[0,1]^s$, in the space R^s. A uniform design is a set of points that are uniformly scattered over the region $\mathcal{T}$ in a certain sense.

Consider the problem of estimating the response (y) as a function of several controlled factors $(X_1, \ldots, X_s)$ in a computer experiment. The true model is

$$y = f(\boldsymbol{x}) = f(x_1, \ldots, x_s),\ \boldsymbol{x} = (x_1, \ldots, x_s) \in \mathcal{T}, \quad (1)$$

where the function f is known, but it is too complicated to manage and to analyze. Researchers want to find an approximate model or a meta model, $y = g(x_1, \ldots, x_s)$, to replace the true model in practice. In physical experiments there exists random error and the true model can been expressed as

$$y = f(\boldsymbol{x}) = f(x_1, \ldots, x_s) + \varepsilon,\ \boldsymbol{x} = (x_1, \ldots, x_s) \in \mathcal{T}, \quad (2)$$

where ε stands for the random error. In traditional experiment designs, such as factorial designs and optimal regression designs, the function f in (2) is known up to some unknown parameters, but in many experiments including high tech experiments the function f is completely unknown. The experimenter wants to estimate the true model f by an experiment.

Theory of the uniform design consists of two parts: design and modeling. The following gives a brief introduction to these two parts.

Design Let $\mathcal{P} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a set of points on $\mathcal{T}$. A measure of uniformity of $\mathcal{P}$ on $\mathcal{T}$ is a function of $\mathcal{P}$. There are so many existing measures in the literature, among them the so-called various discrepancies, such as the star L_p-discrepancy and the centered L_2-discrepancy, have been popularly used. The computational formula for the centered L_2-discrepancy (CD for short) is given by

$$\begin{aligned}(CD(\mathcal{P}))^2 = &\left(\frac{13}{12}\right)^s - \frac{2}{n}\sum_{k=1}^{n}\prod_{j=1}^{s}\left[1 + \frac{1}{2}\left|x_{kj} - 0.5\right|\right.\\ &\left. - \frac{1}{2}\left|x_{kj} - 0.5\right|^2\right]\\ &+ \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\prod_{i=1}^{s}\left[1 + \frac{1}{2}\left|x_{ki} - 0.5\right|\right.\\ &\left. + \frac{1}{2}\left|x_{ji} - 0.5\right| - \frac{1}{2}\left|x_{ki} - x_{ji}\right|\right], \end{aligned} \quad (3)$$

where $\boldsymbol{x}_k = (x_{k1}, \ldots, x_{ks}) \in [0,1]^s, k = 1, \ldots, n$ are experimental points. The smaller value of CD, the better uniformity the set $\mathcal{P}$ has. A n-run UD is a set of n points on $[0,1]^s$ with the minimum pre-decided discrepancy. Due to the computation complexity for searching a UD when n and s increase, nearly UDs are recommended. There are many ways to find UDs such as the good lattice point method, the cutting method, the resolvable balanced incomplete block design method, etc. Readers can find many used UDs at the web site "http://math.hkbu.edu.hk/UniformDesign."

Modeling Based on the experimental data one wishes to find a metamodel to fit the data. This metamodel should approximate the true model well over the region $\mathcal{T}$, where the true model is known in computer experiments and may be unknown in some physical experiments. There are many modeling techniques in the literature, such as linear, quadratic, polynomial or nonlinear regression (See ▶Nonlinear Regression), local polynomial regression, spline, Kriging, the Bayesian approach, and neural network (see ▶Neural Networks), etc. The reader may refer to Fang et al. (2005) and references therein for details.

It has been shown that the UD is robust against the model change. Many users appreciate advantages of the uniform design: (a) high representativeness in the studied experimental domain; (b) do not impose strong assumptions on the underlying model; and (c) flexibility in the number of runs and the number of factors. The uniform design has been widely used not only in various fields such as industry, space engineering, chemical engineering, and high tech, but also in numerical optimization algorithms, artificial neural network (ANN), ▶data mining, and so on. The uniformity measure (CD, for example) has played an important role in the construction of the supersaturated design, comparison of orthogonal arrays, detection of non-isomorphic orthogonal arrays.

About the Author

Dr Fang was the Deputy Director of Institute of Applied Mathematics. Academia Sinica, March 1984–1992 President, The Uniform Design Association of China (1994–2003), Honorary President of Anhui Society of Applied Statistics (2001–2004) and Honorary President of The Uniform Design Association of China (2003–2007), (2007–2011). Currently, he is Professor and Director, Institute of Statistics and Computational Intelligence BNU-HKBU United International College, Zhuhai Campus of Beijing Normal University, China. He is Elected Fellow of the Institute of Mathematical Statistics (1992) and the American Statistical Association (2001). He has delivered lectures at more than 70 universities worldwide.

Cross References

▶Design of Experiments: A Pattern of Progress
▶Factorial Experiments
▶Optimum Experimental Design
▶Statistical Design of Experiments (DOE)

References and Further Reading

Fang KT (1980) The uniform design: application of number-theoretic methods in experimental design, Acta Math Appl Sin 3:363–372

Fang KT, Li R, Sudjianto A (2005) Design and modeling for computer experiments. Chapman & Hall/CRC, London

Fang KT, Lin DKJ (2003) Uniform designs and their application in industry. In: Khattree R, Rao CR (eds) Handbook on statistics 22: statistics in industry. Elsevier/North-Holland, Amsterdam, pp 131–170
Fang KT Lin DKJ, Winker P, Zhang Y (2000) Uniform design: theory and applications. Technometrics 42:237–248
Fang KT, Wang Y (1994) Number-theoretic methods in statistics. Chapman and Hall, London
Hickernell FJ (1998) A generalized discrepancy and quadrature error bound. Math Comp 67:299–322
McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21:239–245
Wang Y, Fang KT (1981) A note on uniform distribution and experimental design. Chin Sci Bull 26:485–489

Uniform Random Number Generators

Pierre L'Ecuyer
Professor, Canada Research Chair in Stochastic Simulation and Optimization
Université de Montréal, Montréal, QC, Canada

Introduction

A growing number of modern statistical tools are based on *Monte Carlo* ideas; they sample independent random variables by computer to estimate distributions, averages, quantiles, roots or optima of functions, etc. These methods are developed and studied in the abstract framework of probability theory, in which the notion of an infinite sequence of independent random variables uniformly distributed over the interval $(0,1)$ (i.i.d. $\mathcal{U}(0,1)$), for example, is well-defined, and the theory is built under the assumption that such random variables can be sampled at will. But in reality, the notion of i.i.d. random variables cannot be implemented exactly on current computers. It can be approximated to some extent by physical devices, but these approximations are cumbersome, inconvenient, and not always reliable, so they are rarely used for computational statistics. Random number generators used for Monte Carlo applications are in reality deterministic algorithms whose behavior *imitates* i.i.d. $\mathcal{U}(0,1)$ random variables. They are pure masquerade. It may be surprising that they work so well, but fortunately they do, or at least some of them do. Here we briefly explain how they are built and tested, and what is the theory behind. Many widely available generators should be avoided and we give examples. We point out reliable ones that can be recommended. More detailed discussions can be found in L'Ecuyer (2004), L'Ecuyer and Panneton (2009), and L'Ecuyer and Simard (2007).

Physical Devices

Hardware devices such as amplifiers of heat noise in electric resistances, photon counting and photon trajectory detectors, and several others, can be used to produce sequences of random bits, which can in turn be used to construct a sequence of floating-point numbers between 0 and 1 that provides a good approximation of i.i.d. $\mathcal{U}(0,1)$ random variables. Most of these devices sample a signal at a given (low) frequency and return 1 if the signal is above a given threshold, 0 otherwise. To improve uniformity and reduce the dependence between successive bits of this sequence, the bits can be cleverly combined via simple operations such as exclusive-or and addition modulo 2, to produce a higher-quality sequence, but at a lower frequency (Chor and Goldreich 1988). These types of "truly random" sequences are needed for applications such as cryptography and gambling machines, for example, where security and unpredictability are essential. But for Monte Carlo algorithms and computational statistics in general, sufficiently good statistical behavior can be achieved by much more practical and less cumbersome algorithmic generators, which require no special hardware.

Algorithmic Generators

These generators are in fact deterministic algorithms that implement a finite-state automaton. They are often called *pseudorandom*. For the remainder of this article, a *random number generator* (RNG) means a system with a finite set of states $\mathcal{S}$, a transition function that determines the next state from the current one, and an output function that assigns to each possible state a real number in $(0,1)$. The system starts from an initial state s_0 (the *seed*) and at each step i, its output u_i and the next state $s_i \in \mathcal{S}$ are determined uniquely by the output function and the transition function, respectively. The output values $\{u_i,\ i \geq 1\}$ are the so-called *random numbers* (an abuse of language) returned by the RNG. In practice, a few truly random bits could be used to select the seed s_0 (although this is rarely done), then everything else is deterministic. In some applications such as for gambling machines in casinos, for example, the state is reseeded frequently with true random bits coming from a physical source, to break the periodicity and determinism. But for Monte Carlo methods, there is no good reason for doing this, so we assume henceforth that no such reseeding is done.

Because the number of states is finite, the RNG will eventually revisit a state that it has already seen, and from then on the same sequence of states (and output values)

will repeat over and over again. That is, for some $l \geq 0$ and $j > 0$, we have $s_{i+j} = s_i$ and $u_{i+j} = u_i$ for all $i \geq l$. The smallest $j > 0$ that satisfies this condition is the *period length* ρ of the RNG. It can never exceed the total number of states. This means that if the state fits in b bits of memory, the period cannot exceed 2^b. This is not really restrictive, because no current computer can generate more than (say) 2^{96} numbers in a lifetime, and a state of three 32-bit integers would suffice to achieve this.

Compared with generators that exploit physical noise, algorithmic RNGs have the advantage that their sequence can be repeated exactly, as many times as we want, without storing it. This is convenient for program verification and debugging, and turns out to be even more important (crucial, in fact) for key variance reduction methods such as common random numbers, which are used all the time for comparing similar systems, for ▶sensitivity analysis, for sample-path optimization, for external control variates, and for antithetic variates, for example (Asmussen and Glynn 2007; Glasserman 2004)

Multiple Streams and Substreams

In modern simulation software, RNGs are often implemented to offer multiple streams and substreams of random numbers. In object-oriented implementations, a *stream* can be seen as a object that produces a long sequence of $\mathcal{U}(0,1)$ random numbers, and such objects can be created in a practically unlimited number, just like other types of objects. These streams are usually partitioned into substreams and methods (or procedures) are readily available to jump ahead to the next substream, or rewind to the beginning of the current substream, or to the beginning of the stream (L'Ecuyer 2008; L'Ecuyer et al. 2002). Of course, a good implementation must make sure that the streams and substreams are long enough so there is no chance of overlap.

To give an example where these facilities are useful, suppose we want to simulate two similar systems with well-synchronized common random numbers (Asmussen and Glynn 2007; Glasserman 2004; Law and Kelton 2000). Think for instance of a large supply chain model or a queueing network, for which we need to estimate the sensitivity of some performance measure with respect to a small change in the operating policy or in some parameter of the system. We want to simulate the system n times (say), with and without the change, with the same random numbers used for the same purpose (as much as possible) in the two systems. The latter is not always easy to implement, because often, the random numbers are generated in a different order for the two systems, and their required quantity is random and differs across the two systems. For that reason, one would usually create and assign a different random stream for each type of random numbers used in these systems (e.g., each type of arrival, each type of service time, routing decisions at each node, machine breakdowns, etc.) (Law et al. 2000; L'Ecuyer 2008). To make sure that the same random numbers from each stream are reused for the two systems for each simulation run, one would simply advance all streams to a new substream at the beginning of a simulation run, simulate the first system, bring these streams back to the beginning of the current substream, simulate the second system, then advance them again to their next substream for the next simulation run. Good simulation and statistical software tools now incorporate these types of facilities.

Basic requirements

From what we have seen so far, obvious requirements for a good RNG are a very long period, the ability to implement the generator easily in a platform-independent way, the possibility of repeating the same sequence over an over again, the facility of splitting the sequence into several disjoint streams and substreams and jumping across them quickly, and of course good speed for the generator itself. Nowadays, fast generators can produce over 100 million $\mathcal{U}(0,1)$ random numbers per second on laptop computers. But these requirements are not sufficient. To see this, consider a RNG that returns $u_i = (i/10^{100}) \mod 1$ at step i. It has all the above properties, but no reasonable statistician would trust it, because of the obvious correlation between the successive outputs. So what else do we need?

Multivariate Uniformity

Both uniformity and independence are covered by the following (joint) statement: For every number of dimensions $s > 0$, the vector of s successive output values $(u_0, \ldots, u_{s-1})$ of the RNG is a random vector with the uniform distribution over the unit hypercube $(0,1)^s$. Of course, this cannot be true, because there are just a finite number of possibilities for that vector. These possibilities are the vectors in the set $\Psi_s = \{(u_0, \ldots, u_{s-1}) : s_0 \in \mathcal{S}\}$, whose cardinality cannot exceed $|\mathcal{S}|$. For a random initial seed, we basically pick a point at random in Ψ_s as an approximation of picking it at random in $(0,1)^s$. For the approximation to be good, Ψ_s must provide a very even (uniform) coverage of the unit hypercube, for s as large as possible. Good RNGs are constructed based on a mathematical analysis of this uniformity. A large Ψ_s (i.e., a large $\mathcal{S}$) is needed to provide a good coverage in high dimensions, and this is the main motivation for having a large state space.

There is no universal measure of this uniformity; in practice, the measure is defined differently for different

classes of RNGs, depending on their mathematical structure, in a way that it is computable without generating the points explicitly (which would be impossible). This is the main reason why the most popular RNGs are based on linear recurrences: their period length and uniformity can be analyzed much more easily than for nonlinear RNGs. To design a RNG, we first select a construction type that can be implemented in an efficient way, and a size of the state space, then we search for parameters that provide a maximal period given that size and the best possible uniformity of Ψ_s for all s up to a certain preselected threshold. After that, the RNG is implemented and submitted to empirical statistical tests.

Empirical Statistical Testing

An unlimited number of statistical tests can be applied to RNGs. These tests take a stream of successive output values and look for evidence against the null hypothesis that they are the realizations of independent $\mathcal{U}(0,1)$ random variables. The hypothesis is rejected when the p-value of the test is extremely close to 0 (which typically indicates strong departure from uniformity) or 1 (which indicates excessive uniformity). As a very simple illustration, one might partition the unit hypercube $(0,1)^s$ in k boxes of volume $1/k$, sample n "independent" points at random in $(0,1)^s$ by generating $s\,\mathcal{U}(0,1)$ random variates for each point, and count the number of times C that a point falls in a box already occupied. If both k and n are very large and $\lambda = n^2/(2k)$ is not too large, then C is supposed to behave approximately as a Poisson random variable with mean λ, which is approximately the same as a normal random variable with mean and variance λ when λ is not too small. If c denotes the realization of C, then the p-value can be approximated by the probability that such a Poisson random variable is at least c. If the p-value is much too small (C is much too large), this means that the points tend to fall in the same boxes more often than they should, whereas if it is too large (C is too small), this means that the points fall too rarely in the same boxes. The latter represents a form of excessive uniformity which is also a departure from randomness.

If the outcome is suspicious but unclear (for example, a small p-value but not excessively small), one can reapply the test (independently), perhaps with a larger sample size. Typically, when the suspicious p-value really indicates a problem, increasing the sample size will clarify things rapidly. When problems are detected, it is frequent to find p-values smaller than 10^{-15}, for example. And this happens for many RNGs used in popular software products (L'Ecuyer and Simard 2007).

It is known that constructing a RNG that passes all possible tests is impossible (L'Ecuyer 2004). The common practice is to forget about the very complicated tests that are too difficult to find and implement, and care only about relatively simple tests. In fact, one could argue that the difference between the good and bad RNGs is that the bad ones fail very simple tests whereas the good ones fail only very complicated tests.

Collections of statistical tests for RNGs have been proposed and implemented in (Knuth 1998; L'Ecuyer and Simard 2007; Marsaglia 1996), for example. Statistical tests can never prove that a RNG is defect-free. They can catch some problems and miss others. For this reason, theoretical tests that measure the uniformity by examining the mathematical structure are more important. Empirical tests can improve our confidence in some RNGs and help us discard bad ones, but they should not be taken as the primary selection criterion.

Linear recurrences modulo *m*

Most algorithmic RNGs in simulation software have a transition function defined by a linear recurrence of the type

$$x_i = (a_1 x_{i-1} + \cdots + a_k x_{i-k}) \mod m, \qquad (1)$$

for some positive integers k and m, and coefficients $a_1, \ldots, a_k$ in $\{0, 1, \ldots, m-1\}$, with $a_k \neq 0$. This recurrence can also be written in matrix form as $\mathbf{x}_i = \mathbf{A}\mathbf{x}_{i-1} \mod m$ where $\mathbf{x}_i = (x_{i-k+1}, \ldots, x_i)^t$. One can obtain a period length of $m^k - 1$ by taking m as a prime number and choosing the coefficients a_j appropriately (Knuth 1998; L'Ecuyer 1996). The output can be defined as $u_i = x_i/m$, or $u_i = (x_i + 1)/(m+1)$, or $u_i = (x_i + 1/2)/m$, for example. This type of RNG is known as a *multiple recursive generator* (MRG). For $k = 1$, we obtain the classical (but obsolete) *linear congruential generator* (LCG). It is easy to advance the state of the MRG by an arbitrary number of steps in a single large jump: $\mathbf{x}_{i+v} = (\mathbf{A}^v \mod m)\mathbf{x}_i \mod m$, after $\mathbf{A}^v \mod m$ has been precomputed (L'Ecuyer 2006).

The uniformity of Ψ_s for the MRG can be measured by exploiting the fact that this point set has a lattice structure. Figures of merit (some related to the so-called spectral test) have been defined to measure the quality of this lattice (Knuth 1998; L'Ecuyer 1999a).

Typically, m is chosen as a prime number that fits the 32-bit or 64-bit word of the computer and the multipliers a_j are chosen so that the recurrence can be computed very quickly. But the quest for speed often goes too far. For example, popular types of generators known as lagged-Fibonacci, add-with-carry, and subtract-with-borrow (which are slight modifications of the MRG) employ only two nonzero coefficients, say a_r and a_k, both equal to ± 1. It turns out that all triples of the form

(u_i, u_{i-r}, u_{i-k}) produced by these generators lie in only two parallel planes in the three-dimensional unit cube (L'Ecuyer 1997). These generators have already given totally wrong results in real-life Monte Carlo applications and should not be used. LCGs with modulus $m \leq 2^{64}$ should also be discarded, because their state space is too small.

An effective construction technique for good MRGs is to combine (say) two or three of them, for example by adding their outputs modulo 1. The idea is to select the components so that a fast implementation is available, while the combined MRG has a long period and its point set Ψ_s has good uniformity. Good parameters can be found by extensive computer searches. For specific constructions of this type, see (L'Ecuyer 1999a; L'Ecuyer 2006; L'Ecuyer and Simard 2007). One of them is the `MRG32k3a` generator (L'Ecuyer 1999a; L'Ecuyer et al. 2002), now available with streams and substreams in many statistical and simulation software products.

Linear Recurrences Modulo 2

Given that current computers work in binary arithmetic, it is no surprise that many of the fastest good RNGs are based on linear recurrences modulo 2. That is, recurrence (1) with $m = 2$. This can be framed in matrix notation (L'Ecuyer 2006; L'Ecuyer and Panneton 2009) as:

$$\mathbf{x}_i = \mathbf{A}\mathbf{x}_{i-1} \bmod 2,$$
$$\mathbf{y}_i = \mathbf{B}\mathbf{x}_i \bmod 2,$$
$$u_i = \sum_{\ell=1}^{w} y_{i,\ell-1} 2^{-\ell}$$

where $\mathbf{x}_i = (x_{i,0}, \ldots, x_{i,k-1})^{\mathrm{t}}$ is the k-bit *state vector* at step i, $\mathbf{y}_i = (y_{i,0}, \ldots, y_{i,w-1})^{\mathrm{t}}$ is a w-bit *output vector*, k and w are positive integers, $\mathbf{A}$ is a $k \times k$ binary matrix, $\mathbf{B}$ is a $w \times k$ binary matrix, and $u_i \in [0,1)$ is the *output* at step i. In practice, the output can be modified slightly to make sure that the generator never returns exactly 0.

Many popular generators belong to this class, including the Tausworthe or linear feedback shift register (LFSR) generator, polynomial LCG, generalized feedback shift register (GFSR), twisted GFSR, Mersenne twister, WELL, xorshift, linear cellular automaton, and combinations of these (L'Ecuyer 1999b, 2004; L'Ecuyer and Panneton 2009; Matsumoto and Nishimura 1998). The largest possible period is $2^k - 1$, reached when the characteristic polynomial of $\mathbf{A}$ is a primitive polynomial modulo 2. The matrices $\mathbf{A}$ and $\mathbf{B}$ are always selected to allow a fast implementation by using just a few simple binary operations such as or, exclusive-or, shift, and rotation, on blocks of bits, while still providing good uniformity for the point set Ψ_s. This uniformity is assessed by measures of equidistribution of the points in the diadic rectangular boxes obtained by partitioning $(0,1)$ for each axis j into intervals of lengths 2^{-q_j} for some integers $q_j \geq 0$ (L'Ecuyer and Panneton 2009). Combined generators of this type, obtained by a bitwise exclusive-or of the output vectors $\mathbf{y}_i$ of two or more generators from that class, are equivalent to yet another generator from the same class (L'Ecuyer 1999b; L'Ecuyer and Panneton 2009). Their motivation is the same as for combined MRGs.

Nonlinear Generators

Linear RNGs have a regular structure that can eventually be detected by statistical tests cleverly designed for that detection. Cryptologists know well about that and use (slower) nonlinear RNGs for that reason. For Monte Carlo, the linearity itself is practically never a problem, because the random numbers are almost always transformed in some nonlinear way by the simulation algorithm. But there are situations where linearity matters. For example, if we generate large random binary matrices and the rank of the matrix must have the right distribution, then we should not use a linear generator modulo 2, because there are too many linear dependencies between the bits (L'Ecuyer and Simard 2007).

A nonlinear RNG can be obtained, for example, by simply adding a nonlinear output transformation to a linear RNG, or by shuffling its output values using another generator, or by using a nonlinear recurrence in the construction, or by combining two generators of different types, such as an MRG and a generator based on a linear recurrence modulo 2. For nonlinear RNGs, the uniformity of Ψ_s is generally too difficult to analyze. But for the last type of combination just mentioned, useful bounds can be obtained on uniformity measures (L'Ecuyer and Granger-Piché 2003). It is also important to understand that combining generators does not necessarily leads to an improvement. Nonlinear RNGs are also slower in general than their linear cousins. On the other hand, they tend to perform better in empirical statistical tests (L'Ecuyer and Simard 2007).

Recommendations

When asked for recommendations on uniform RNGs, my natural response is to say which ones I use for my own experiments. The RNG I use most of the time in my lab is `MRG32k3a`, from (L'Ecuyer 1999a). It is very robust and reliable, based on a solid theoretical analysis, and it also provides multiple streams and substreams (L'Ecuyer et al. 2002). It is not the fastest one, though. If the uniform RNG's speed really matters, good alternatives are `MRG31k3p` from (L'Ecuyer and Touzin 2000),

LFSR113 and LFSR258 from (L'Ecuyer 1999b), and some small WELL generators from (Panneton et al. 2006), for which multiple streams and substreams are also available in (L'Ecuyer 2008). For very fast generators with huge periods, the Mersenne twister MT19937 (Matsumoto and Nishimura 1998) and the WELL generators of (Panneton et al. 2006) are good choices. On the other hand, these generators have a very large state, so using them for multiple streams is not very efficient. They are more appropriate for situations where a single stream suffices.

The list of widely-used generators that should be discarded is much longer, as can be seen from the empirical results in (L'Ecuyer and Simard 2007). Do not trust blindly the software vendors. Check the default RNG of your favorite software and be ready to replace it if needed. This last recommendation has been made over and over again over the past 40 years. Perhaps amazingly, it remains as relevant today as it was 40 years ago.

Acknowledgments

This work has been supported by the Natural Sciences and Engineering Research Council of Canada Grant No. ODGP0110050 and a Canada Research Chair to the author.

About the Author

Pierre L'Ecuyer is a Professor in the "Département d'informatique et de recherche opérationnelle" at the University of Montreal. He is Elected Fellow of the Institute for Operations Research and the Management Sciences (2006). Professor L'Ecuyer was awarded the Toshiba Chair at Waseda University, Tokyo (1992), the Jacob Wolfowitz Prize for Theoretical Advances in the Mathematical and Management Sciences in 2000, and the Urgel-Archambault Prize from ACFAS (2002). He has published more than 200 refereed scientific articles and book chapters in various areas, including random number generation, quasi-Monte Carlo methods, efficiency improvement in simulation, sensitivity analysis and optimization for discrete-event simulation models, simulation software, stochastic dynamic programming, and applications in finance, manufacturing, telecommunications, and service center management. He also developed software libraries and systems for the theoretical and empirical analysis of random number generators and quasi-Monte Carlo point sets, and for general discrete-event simulation. He is currently Editor-in-Chief of the ACM Transactions on Modeling and Computer Simulation. He is also Associate Editor for the *ACM Transactions on Mathematical Software*, *Statistics and Computing* (Springer-Verlag), *Management Science*, *International Transactions in Operational Research*, *Cryptography and Communications—Discrete Structures, Boolean Functions and Sequences* (Springer-Verlag). He has been a reviewer for 108 international scientific journals since 1985.

Cross References

►Computational Statistics
►Copulas: Distribution Functions and Simulation
►Detecting Outliers in Time Series Using Simulation
►Divisible Statistics
►Inverse Gaussian Distribution
►Monte Carlo Methods in Statistics
►Non-Uniform Random Variate Generations
►Pareto Sampling
►Simple Random Sample
►Uniform Distribution in Statistics

References and Further Reading

Asmussen S, Glynn PW (2007) Stochastic simulation. Springer-Verlag, New York

Chor B, Goldreich O (1988) Unbiased bits from sources of weak randomness and probabilistic communication complexity. SIAM J Comput 17(2):230–261

Glasserman P (2004) Monte Carlo methods in financial engineering. Springer-Verlag, New York

Knuth DE (1998) The art of computer programming, vol 2: seminumerical algorithms, 3rd edn. Addison-Wesley, Reading, MA

Law AM, Kelton WD (2000) Simulation modeling and analysis, 3rd edn. McGraw-Hill, New York

L'Ecuyer P (1996) Combined multiple recursive random number generators. Oper Res 44(5):816–822

L'Ecuyer P (1997) Bad lattice structures for vectors of non-successive values produced by some linear recurrences. INFORMS J Comput 9(1):57–60

L'Ecuyer P (1999a) Good parameters and implementations for combined multiple recursive random number generators. Oper Res 47(1):159–164

L'Ecuyer P (1999b) Tables of maximally equidistributed combined LFSR generators. Math Comput 68(225):261–269

L'Ecuyer P (2004) Chapter II.2: Random number generation. In: Gentle JE, Haerdle W, Mori Y (eds) Handbook of computational statistics, Springer-Verlag, Berlin, pp 35–70

L'Ecuyer P (2006) Chapter 3: Uniform random number generation. In: Henderson SG, Nelson BL (eds) Simulation, Handbooks in operations research and management science, Elsevier, Amsterdam, pp 55–81

L'Ecuyer P (2008) SSJ: A Java library for stochastic simulation, software user's guide. Available at http://www.iro.umontreal.ca/~lecuyer.

L'Ecuyer P, Granger-Piché J (2003) Combined generators with components from different families. Math Comput Simul 62: 395–404

L'Ecuyer P, Panneton F (2009) F2-linear random number generators. In: Alexopoulos C, Goldsman D, Wilson JR (eds) Advancing the frontiers of simulation: a festschrift in honor of George Samuel Fishman. Springer-Verlag, New York, pp 169–193

L'Ecuyer P, Simard R (2007) TestU01: a C library for empirical testing of random number generators. ACM Trans Math Softw 33(4):22

L'Ecuyer P, Simard R, Chen EJ, Kelton WD (2002) An object-oriented random-number package with many long streams and substreams. Oper Res 50(6):1073–1075

L'Ecuyer P, Touzin R (2000) Fast combined multiple recursive generators with multipliers of the form $a = \pm 2^q \pm 2^r$. In: Joines JA, Barton RR, Kang K, Fishwick PA (eds) Proceedings of the 2000 winter simulation conference. IEEE Press, Piscataway, NJ, pp 683–689

Marsaglia G (1996) DIEHARD: a battery of tests of randomness. http://www.stat.fsu.edu/pub/diehard.Accessed3Aug2010

Matsumoto M, Nishimura T (1998) Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans Model Comput Simul 8(1):3–30

Panneton F, L'Ecuyer P, Matsumoto M (2006) Improved long-period generators based on linear recurrences modulo 2. ACM Trans Math Softw 32(1):1–16

Univariate Discrete Distributions: An Overview

Adrienne W. Kemp
University of St. Andrews, St. Andrews, UK

Introduction

A random variable (rv) is said to be discrete if it can take a finite or a countably infinite number of values, i.e., has a discrete state space. These values need not be equally spaced but almost all discrete random variables of use in statistics take equally spaced values and so are said to have lattice distributions. Examples are numbers of aircraft accidents, numbers of bank failures, cosmic ray counts ($x = 0, 1, 2, \ldots$), counts of occupants per car ($x = 1, \ldots$), and numbers of albino children in families of 6 children ($x = 0, 1, \ldots, 6$). Lattice variables are not restricted to count events; they can also be obtained by discretization of continuous measurements, e.g., flood heights ($x = 0, 0.5, 1, 1.5, \ldots$ meters).

The set of all possible outcomes from an experiment or sampling scheme is called the sample space, Ω. In univariate situations a single real value is associated with every outcome. The function X that determines these numerical values is the random variable and the individual values that it takes are denoted by x. The set of values that X can take is called its support.

Distributions of rv's are concerned with the probabilities with which the observed values occur. If the method of experimentation or the method of sampling is stochastic (probabilistic), not deterministic, then every value x occurs with a probability $\Pr[X = x] = p(x) = p_x$ called its probability mass function (pmf). Necessary constraints on the probabilities are $p_x \geq 0$ and $\sum_x p_x = 1$. The distribution of a rv depends therefore on Ω, X and $\Pr[X = x]$. Discrete distributions are called logconvex when $p_x p_{x+2}/p_{x+1}^2 > 1$ and logconcave when $p_x p_{x+2}/p_{x+1}^2 < 1$.

Important mathematical functions associated with a discrete distribution are the cumulative distribution function (cdf) (a step function)

$$F(x) = \Pr[X \leq x] = \sum_{y \leq x} p_y,$$

the probability generating function (pgf)

$$G_X(z) = G(z) = E[z^X] = \sum_{x=0}^{\infty} p_x z^x,$$

and the characteristic function (cf)

$$\varphi_X(t) = \varphi(t) = E[e^{itX}] = \sum_x \Pr[X = x] e^{itx} = G_X(e^{it}).$$

A cf for a discrete distribution is infinitely divisible if $\varphi(t) = \{\varphi_n(t)\}^n$ for all positive integers n, where $\varphi_n(t)$ is a cf. It is decomposible if there exist two nondegenerate cf's, $\varphi_1(t)$ and $\varphi_2(t)$ such that $\varphi(t) = \varphi_1(t)\varphi_2(t)$. Usually the cf for a limiting distribution is the limiting cf .

Important survival concepts are the survival (survivor) function

$$S_0 = 1, \quad S_t = 1 - \Pr(T < t) = \sum_{j \geq t} p_j, \quad t = 1, 2, \ldots,$$

and the hazard function (often called the failure rate, FR)

$$h_t = p_t / \sum_{j \geq t} p_j = (S_t - S_{t+1})/S_t.$$

A discrete distribution is said to have a monotonically non-decreasing failure rate with time (IFR) or a monotonically non-increasing failure rate with time (DFR) according as $p_{t+1}/p_t \gtrless p_{t+2}/p_{t+1}$.

Historical Perspective

Until the mid-twentieth century interest in discrete distributions centered mainly on (i) the solution of particular problems such as the number of tails thrown before the appearance of the first head, and (ii) the empirical fitting of discrete data such as haemocytometer counts. The distributions in general use were the binomial, hypergeometric, uniform (discrete rectangular), Poisson and negative binomial. Others were mainly used in limited application areas, for example in linguistics the Zipf-Estoup distribution (Estoup 1916) with pmf

$$p_x = x^{-\eta} / \sum_{x=1}^{\infty} x^{-\eta}, \ \eta > 1, \ x = 1, 2, \ldots \quad \text{and pgf}$$

$$G(z) = \sum_{x=1}^{\infty} z^x x^{-\eta} / \sum_{x=1}^{\infty} x^{-\eta},$$

in taxonomy Yule's (1925) distribution with pmf

$$p_x = \rho(\rho!)(x-1)!/(x+\rho)!,\ \rho > 0,\ x = 1,2,\ldots \text{ and pgf}$$
$$G(z) = \rho z\ {}_2F_1[1,1;\rho+2;z]/(\rho+1),$$

and in ecology Fisher's (1943) logarithmic distribution with pmf

$$p_x = -[\ln(1-\theta)]^{-1}\theta^x/x,\ 0 < \theta < 1,\ x = 1,2,\ldots \text{ and pgf}$$
$$G(z) = \ln(1-\theta z)/\ln(1-\theta).$$

Post mid-twentieth century there has been a shift away from the graduation of data toward the creation of distributions with more complicated underlying mathematical models. With this came an increased understanding of families of distributions and the realization that the same distribution, e.g., the negative binomial, can arise from several different models.

The very broad class of power series distributions has pmf's of the form

$$\Pr[X = x] = \frac{a_x\theta^x}{\eta(\theta)}, \quad \theta > 0, \quad a_x \geq 0,$$
$$\eta(\theta) = \sum_{x=0}^{\infty} a_x\theta^x, \quad x = 0,1,\ldots;$$

θ is the power parameter and $\eta(\cdot)$ is the series function. These are discrete linear exponential distributions and so have important inference properties. This class was explored in depth by Patil; see e.g., Patil (1986). It includes most of the distributions mentioned above. Generalized power series distributions and modified power series distributions are extensions of this family.

The Katz family is a discrete analogue of the Pearson system of continuous distributions; see Katz (1965). Their pmf's satisfy

$$p_{x+1}/p_x = (a+bx)/(1+x),\ a > 0,\ b < 1,\ x = 1,2,\ldots.$$

For Ord's (1972) difference equation family we have

$$p_x - p_{x-1} = (a-x)p_x/\{(a+b_0) + (b_1-1)x + b_2x(x-1)\},$$

where x takes a range of integer values. Sundt and Jewell (1981) and Klugman et al. (1998) have found distributions of these kinds, with the possible modification of p_0, very useful in actuarial studies.

Kemp's (1974) family of generalized hypergeometric probability distributions (GHPD) have pgf's have the form

$$G(z) = {}_pF_q[a_1,\ldots,a_p;b_1,\ldots,b_q;\lambda z]/{}_pF_q$$
$$[a_1,\ldots,a_p;b_1,\ldots,b_q;\lambda]$$

where ${}_pF_q[a_1,\ldots,a_p;b_1,\ldots,b_q;y]$ is a generalized hypergeometric function; Kem and Kemp (1974) family of generalized hypergeometric factorial moment distributions (GHFD) have pgf's of the form

$$G(z) = {}_pF_q[a_1,\ldots,a_p;b_1,\ldots,b_q;\lambda(1-z)].$$

These families include very many distributions used in applied statistics, including important matching and occupancy distributions; models and properties are discussed in Johnson et al. (2005).

Other important families are the Lagrangian family (see Consul and Famoye 2005) and the order-k family (see Balakrishnan and Koutras 2002). There has been renewed interest in q-hypergeometric series distributions and in the Lerch family in the last twenty years; details and references are in Johnson et al. (2005).

Discrete stochastic processes are random processes that have a discrete state space and evolve in time. They include random walks (see ▶Random Walk), the Poisson process (see ▶Poisson Processes), ▶Markov chains, birth-and-death processes and branching processes. Doob (1953) and Feller (1968) brought the work of earlier probabilists to the attention of applied statisticians who came to realize that these processes provide further models for many existing discrete distributions; see e.g., Jones and Smith (2001).

Distributions can be made more flexible by weighting. Let X be a rv with pmf p_x, and suppose that when the event $X = x$ occurs the probability of recording it is $w(x)$. Then the pmf for the ascertained distribution is a weighted distribution with pmf

$$p_x^* = w(x)p(x)/\sum_x w(x)p(x).$$

Hurdle models assume different underlying statistical processes, $f_a(x)$ below the hurdle and $f_b(x)$ above the hurdle. When the hurdle is at zero (the most often used hurdle), the outcome has the pmf

$$p_0 = f_a(0),\ \ p_x = \{1-f_a(0)\}f_b(x)/\{1-f_b(0)\},\ x = 1,2,\ldots;$$

see e.g., Winkelmann (2000). The hurdle-at-zero model allows for over- and under-inflation of the probability at zero, also over- and under-dispersion.

Models of physical situations often involve the combination of distributions. Three important ways are convolution, mixing, and compounding.

The distribution of the sum $X = X_1 + X_2$ of two independent rv's X_1 and X_2 with pgf's $G_1(z)$ and $G_2(z)$ has the pmf

$$\Pr[X = i] = \sum_j \Pr[X_1 = i-j]\Pr[X_2 = j] \quad \text{and pgf}$$
$$G(z) = G_1(z)G_2(z).$$

If A, B and C are the names of the distributions of X_1, X_2 and X, then C is called the convolution of X_1 and X_2; we write C ~ A*B. Kemp's (1987) weapon defense model involves the convolution of several independent binomial distributions.

Secondly, consider a rv with pmf $\Pr[X = x|\theta_1, \ldots, \theta_k]$ dependent on the parameters θ_i, $i = 1, \ldots, k$, where some or all of the parameters vary. The outcome is a mixture distribution (sometimes called a compound distribution in the early literature) with the pmf $E[\Pr[X = x|\theta_1, \ldots, \theta_\ell]]$, where expectation is with respect to the joint distribution of the varying parameters. If only one parameter varies, then the mixture is denoted symbolically by $\mathcal{F}_1 \underset{\Theta}{\wedge} \mathcal{F}_2$, where $\mathcal{F}_1$ is the original distribution and $\mathcal{F}_2$ is the distribution of Θ (the mixing distribution). If Θ has a discrete distribution with probabilities p_j^*, $j = 0, 1, \ldots$, then the outcome has pmf

$$\Pr[X = x] = \sum_{j \geq 0} p_j^* \Pr[X_j = x|\theta].$$

Gelfand and Soloman's (1975) analysis of jury decisions used a mixture of two binomials with different values of p. Such mixtures have a Bayesian interpretation; if p_j^* is the pmf for a discrete prior distribution then $p_j^* \times \Pr[X_j = x|\theta] / \Pr[X = x]$ is the pmf for a posterior distribution. When the mixing distribution is continuous, the outcome has pmf

$$\Pr[X = x] = \int \Pr[X = x|\theta] f(\theta) d\theta,$$

where the probability density function of Θ is $f(\theta)$ and integration is over all values of Θ. For the beta-binomial distribution the binomial parameter p has a beta distribution; see e.g., the study concerning dead fetuses by Brooks et al. (1997).

A third method for combining distributions was called "generalizing" or "compounding" in the early literature. A more recent, less confusing, name is "random [stopped] sum." Suppose that the size N of the initial generation in a branching process has the pgf $G_1(z)$, and that each individual i in this initial generation gives rise independently to Y_i first generation individuals; suppose also that the pgf for Y_i is $G_2(z)$. The total number of first generation individuals is then $S_N = Y_1 + Y_2 + \ldots + Y_N$, with pgf

$$E[z^{S_N}] = E_N[E[z^{S_N}|N]] = G_1(G_2(z)).$$

Here S_N has a "generalized" distribution and the distribution of Y_i is the "generalizing" distribution. More recently S_N is said to have a randomly $\mathcal{F}_1$-stopped summed-$\mathcal{F}_2$ distribution, where $G_1(z)$ is the pgf for $\mathcal{F}_1$ and $G_2(z)$ is the pgf for $\mathcal{F}_2$. It is symbolized by $S_N \sim \mathcal{F}_1 \bigvee \mathcal{F}_2$. An early example is Neyman's (1939) model for the distribution of insect larvae; it assumes Poissonian numbers of clusters of eggs per unit area and Poissonian numbers of eggs per cluster. Mixture and random stopped-sum distributions are discussed at length in Johnson et al. (2005).

For very large data sets, the ability to regress on elements of particular interest is important. This has led to the use of discrete distributions that are particularly suitable for regression models such as those that are linear exponential.

In recent years there has been much interest in resampling methods of analysis where discrete distributions are formed by resampling from discrete data. Further exploration of the properties of such distributions would be useful; see e.g., Good (1999) for methodology and Gentle (2002) for theory.

Properties

The moment properties of a discrete distribution are easily obtainable from the pgf.

The (uncorrected) ▶moment generating function (mgf) is $M(t) = G(e^t)$ whence

$$\mu'_r = \sum_{x=0}^{\infty} x^r p_x = [d^r G(e^t)/dt^r]_{t=0}, \quad r = 1, 2, \ldots.$$

The central (corrected) moment generating function (cmgf) is

$$e^{-\mu t} M(t) = e^{-\mu t} G(e^t) = \sum_1^{\infty} \mu_r t^r / r!.$$

The cumulant generating function (cgf) is

$$K(t) = \ln G(e^t) = \sum_1^{\infty} \kappa_r t^r / r!.$$

The factorial moment generating function (fmgf) is

$$E[(1+t)^X] = G(1+t) = 1 + \sum_{r \geq 1} \mu'_{[r]} / t^r r!$$

and the factorial cumulant generating function (fcgf) is

$$\ln G(1+t) = \sum_1^{\infty} \kappa_{[r]} t^r / r!.$$

The fmgf is useful because the (descending) factorial moments are obtainable by successive differentiation of the pgf:

$$\mu'_{[r]} = \sum_{j=r}^{\infty} \frac{j!}{(j-r)!} p_j = \left[\frac{d^r G(z)}{dz^r}\right]_{z=1} = \left[\frac{d^r G(1+t)}{dt^r}\right]_{t=0};$$

the moments can then be obtained from the factorial moments as

$$\mu = \mu'_{[1]}, \quad \mu'_2 = \mu'_{[2]} + \mu, \quad \mu'_3 = \mu'_{[3]} + 3\mu'_{[2]} + \mu,$$
$$\mu'_4 = \mu'_{[4]} + 6\mu'_{[3]} + 7\mu'_{[2]} + \mu, \quad \text{etc.;}$$

in general

$$\mu'_r = \sum_{j=0}^{r} S(r,j)\mu'_{[j]} \quad \text{and} \quad \mu'_{[r]} = \sum_{j=0}^{r} s(r,j)\mu'_j$$

where $S(r,j)$ and $s(r,j)$ are the Stirling numbers of the second kind and first kind, respectively.

The first uncorrected moment $\mu = \mu'_1 = \mu'_{[1]}$ is the mean; the second central moment $\mu_2 = \mu'_2 - \mu^2$ is the variance ($\mathrm{Var}(X) \equiv \sigma_X^2$), and its positive square root is the standard deviation σ. The coefficient of variation is σ/μ. Moment ratios are used as indices of shape; $\alpha_3(X) = \sqrt{\beta_1(X)} = \mu_3(\mu_2)^{-3/2}$ is used as an index of skewness and $\alpha_4(X) = \beta_2(X) = \mu_4(\mu_2)^{-2}$ is used as an index of kurtosis.

The distribution of the sum $X = X_1 + X_2$ of two independent rv's X_1 and X_2 with pgf's $G_1(z)$ and $G_2(z)$ has the pgf $G(z) = G_1(z)G_2(z)$. Their difference $X_1 - X_2$ has the pgf $G_{X_1-X_2}(z) = G_1(z)G_2(1/z)$; $X_1 - X_2$ may take negative values.

The median for a discrete distribution with $2N + 1$ points of support is the value of the $(N + 1)$th point of support; for a discrete distribution with $2N$ points of support the median is usually taken to be the average of the Nth and $(N + 1)$th points of support. A mode of a discrete distribution is at $X = x$ if

$$\Pr[X = x - a] \le \cdots \le \Pr[X = x] \quad \text{and}$$
$$\Pr[X = x] \ge \cdots \ge \Pr[X = x + b]$$

where $0 \le a < b$. A discrete distribution is unimodal if it has only one mode; otherwise it is multimodal. A discrete distribution is said to have a half-mode at $X = 0$ and to be sesquimodal if

$$p_0 > p_1 \ge p_2 \ge \ldots.$$

Estimation methods for the parameters of discrete distributions have been studied extensively but it is difficult to construct confidence intervals

$$\Pr(\theta \in \{c \le \theta \le d\}) = 1 - \alpha$$

for an exact value of the confidence level α. Estimation by the method of moments is usually simpler than maximum likelihood. It equates the first k sample moments about zero,

$$m'_r = n^{-1}\sum_{j=1}^{n} x_j^r, \; r = 1, \ldots, k,$$

to their corresponding theoretical expressions, μ'_r (k is the number of unknown parameters). The method of moments and zero frequency uses the equations for $k - 1$ moments and one equating p_0 to its expected value.

Maximum likelihood estimation with its very desirable properties is achieved by solving the equations

$$\frac{\partial L(x_1, x_2, \ldots, x_n|\theta_1, \theta_2, \ldots, \theta_k)}{\partial \theta_i} = 0, \qquad i = 1, 2, \ldots, k.$$

Their solution often requires iteration. Computer packages or optimization methods are used; good initial estimates are helpful. Several variants of ML estimation have been developed such as conditional, profile and marginal likelihood procedures; see Johnson et al. (2005) for references. Bayesian methods of estimation for discrete distributions have been studied in depth; see e.g., Congdon (2003).

About the Author

Adrienne W. (Freda) Kemp is Honorary Senior Lecturer at the Mathematical Institute, University of St Andrews, Scotland. She is known as a co-author of the text *Univariate Discrete Distributions* (with Norman L. Johnson and Samuel Kotz, Wiley, 3rd edition, 2005).

Cross References

►Binomial Distribution
►Distributions of Order K
►Geometric and Negative Binomial Distributions
►Hypergeometric Distribution and Its Application in Statistics
►Markov Chains
►Mixture Models
►Poisson Distribution and Its Application in Statistics
►Poisson Processes
►Random Walk
►Relationships Among Univariate Statistical Distributions
►Statistical Distributions: An Overview

References and Further Reading

Balakrishnan N, Koutras MV (2002) Runs and scans with applications. Wiley, New York
Brooks RJ, James WH, Gray E (1991) Modelling sub-binomial variation in the frequency of sex combinations in litters of pigs. Biometrics 47:403–417
Congdon P (2003) Applied Bayesian modelling. Wiley, Chichester
Consul PC, Famoye F (2005) Lagrangian probability distributions. Birkhäuser, Boston
Doob JL (1953) Stochastic processes. Wiley, New York
Estoup JB (1916) Les Gammes Sténographiques. Institut Sténographique, Paris
Feller W (1968) An introduction to probability theory and its applications, vol 1, 3rd edn. Wiley, New York

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol 12:42–58
Gelfand AE, Soloman H (1975) Analysing the decision making process of the American jury. J Am Stat Assoc 70:305–310
Gentle JE (2002) Elements of computational statistics. Springer, New York
Good PI (1999) Resampling methods. Birkhäuser, Boston
Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, Hoboken
Jones PW, Smith P (2001) Stochastic processes, an introduction. Arnold, London
Katz L (1965) Unified treatment of a broad class of discrete probability distributions. In: Patil GP (ed) Classical and contagious discrete distributions. Statistical Publishing Society/Pergamon, Calcutta/Oxford, pp 175–182
Kemp AW (1968) A wide class of discrete distributions and the associated differential equations. Sankhya Ser A 30:401–410
Kemp AW and Kemp CD (1974) A family of distributions defined via their factorial moments. Comm Stat 3:1187–1196
Kemp AW (1987) A Poissonian binomial model with constrained parameters. Nav Res Logist 34:853–858
Klugman SA, Panjer HH, Willmot GE (1998) Loss models: from data to decisions. Wiley, New York
Neyman J (1939) On a new class of "contagious" distributions applicable in entomology and bacteriology. Ann Math Stat 10:35–57
Ord JK (1972) Families of frequency distributions. Griffin, London
Patil GP (1986) Power series distributions. In: Kotz S, Johnson NL, Read CB (eds) Encyclopedia of statistical sciences, vol 7. Wiley, New York, pp 130–134
Sundt B, Jewell WS (1981) Further results on recursive evaluation of compound distributions. ASTIN Bull 18:27–39
Winkelmann R (2000) Econometric analysis of count data. Springer, New York
Yule GU (1925) A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F.R.S. Philos Trans R Soc Lond Ser B 213:21–87

U-Statistics

Neville C. Weber
Professor and Head of School of Mathematics and Statistics
University of Sydney, Sydney, NSW, Australia

The class of U-statistics was introduced in the seminal paper of Hoeffding (1948), motivated by the work of Halmos (1946) on unbiased estimators. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed observations from a distribution $F(x, \theta)$, where θ is some parameter of interest. If g is a function of m variables such that $Eg(X_1, \ldots, X_m) = \theta$ then a natural unbiased estimator for θ can be constructed by evaluating g at all subsets of size m that can be formed from the observations and then averaging these values. Write

$$U_n = n_{(m)}^{-1} \sum g(X_{i_1}, \ldots, X_{i_m}),$$

where $n_{(m)} = n(n-1)\ldots(n-m+1)$ and the sum is over all m-tuples $(i_1, \ldots, i_m)$ of distinct elements of $\{1, 2, \ldots, n\}$. Note we can consider the symmetric function formed by first averaging over all permutations of a given set of indices, so $h(x_1, \ldots, x_m) = (m!)^{-1} \sum g(x_{i_1}, \ldots, x_{i_m})$, where the sum is over all $m!$ permutations $(i_1, \ldots, i_m)$ of $(1, \ldots, m)$. Thus, in general, U is defined in terms of a function h, called the kernel, that is symmetric in its arguments and

$$U_n = \binom{n}{m}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_m \le n} h(X_{i_1}, \ldots, X_{i_m}).$$

The value m is known as the degree of the kernel.

U-statistics are the focus of much research as many commonly used statistics, including nonparametric and spatial statistics, can be expressed in this format by appropriate choice of kernel, or they can be closely approximated by U-statistics. For example, the kth sample moment is a U-statistic of degree 1 with kernel $h(x) = x^k$; if $h(x, y) = \frac{1}{2}(x - y)^2$ then we obtain the sample variance, and if $h(x, y) = I(x + y \le 0)$ we obtain the Wilcoxon one-sample statistic. In the case where X_i is a bivariate observation and $h(\mathbf{x}, \mathbf{y}) = I((x_1 - x_2)(y_1 - y_2) > 0)$ then $2U_n - 1$ is Kendall's coefficient of concordance, τ.

There are natural extensions to multiple sample U-statistics. If we have c independent samples, then the kernel is a function of m_j terms from the jth sample, $j = 1, \ldots, c$, and the statistic is formed by averaging over all possible combinations of m_1 terms from sample 1, m_2 terms from sample 2 and so on. We will restrict attention to the single sample case. For a comprehensive introduction to the broad range of U-statistics, see Lee (1990).

For a U-statistic with kernel h let $h_c(x_1, \ldots, x_c) = E(h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m))$ and let $\sigma_c^2 = \mathrm{Var}(h_c(X_1, \ldots, X_c))$. If $\sigma_c^2 = 0$, $c = 1, \ldots, d-1$ and $\sigma_d^2 > 0$ then U_n is said to be degenerate of order $d-1$. The variance of a U-statistic can be written as

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \sigma_c^2.$$

General moment results can be derived. When $E|h(X_1, \ldots, X_m)|^r < \infty$, for $r \ge 2$, $E|U_n - \theta|^r = O(n^{-r/2})$.

There has been a rich theory developed describing the behavior of U-statistics. Key to these results are several useful representations. First, provided $E|h(X_1, \ldots, X_m)| < \infty$, the U-statistic can be represented as a reverse

martingale and this leads naturally to various strong convergence results.

The most widely used representation is the H-decomposition developed by Hoeffding where $U_n = \theta + \sum_{j=1}^{m} \binom{m}{j} U_{nj}$, where U_{nj} is a U-statistic of degree j with kernel g_j. The U_{nj} are uncorrelated and $Eg_j(x_1, \ldots, x_{j-1}, X_j) = 0$.

From Hoeffding's original paper if $\sigma_1^2 > 0$, that is the U-statistic is non-denegerate, then $\sqrt{n}(U_n - \theta)$ converges in distribution to an asymptotic normal distribution with mean 0 and variance $m^2\sigma_1^2$. If the U-statistic is degenerate then the asymptotic behavior is more complex. For example, consider the Rayleigh statistic which is a U-statistic with kernel $h(x, y) = 2\cos(x - y)$. If X_i has a uniform distribution on $(0, 2\pi)$ then U_n is degenerate of order 1 and $2n^{-1}U_n$ has an asymptotic, mean adjusted, chi-squared distribution with 2 degrees of freedom. If $\sigma_1 = \ldots = \sigma_{d-1} = 0$, $\sigma_d > 0$, $d > 1$ then $n^{d/2}(U_n - \theta)$ has a nondegenerate limit which can be expressed in terms of multiple Wiener integrals (see Dynkin and Mandelbaum 1983). For a thorough coverage of the asymptotic theory of U-statistics including asymptotic expansions, large deviation results, rates of convergence and the law of the iterated logarithm, see the monograph by Koroljuk and Borovskikh (1994).

The theory of U-statistics has been extended to statistics of this format based on weakly dependent samples, such as observations based on samples from finite populations, and to statistics where the kernel depends on the sample size, n. Further there is a well developed theory for ►empirical processes related to the empirical distribution function H_n of $h(X_{i_1}, \ldots, X_{i_m})$. U-processes are collections of U-statistics indexed by families of kernels, $\{U_n(h), h \in \mathcal{H}\}$. The study of these processes was originally motivated by a problem on cross-validation in density estimation. See de la Peña and Giné (1999) for an explanation of decoupling theory and its application to the theory of these processes.

About the Author

Neville Weber is Head of the School of Mathematics and Statistics and Professor of Mathematical Statistics at The University of Sydney. He has been an Associate Editor for the "*International Statistical Review*," "*The Australian Journal of Statistics*," and "*The Bulletin of the Australian Mathematical Society*." In 2001 he was made an Honorary Life Member of the Statistical Society of Australia, Inc.

Cross References

►Almost Sure Convergence of Random Variables
►Empirical Processes
►Kendall's Tau
►Weighted U-Statistics
►Wilcoxon-Signed-Rank Test

References and Further Reading

de la Peña VH, Giné E (1999) Decoupling. From dependence to independence. Springer, New York

Dynkin EB, Mandelbaum A (1983) Symmetric statistics, Poisson point processes and multiple Weiner integrals. Ann Stat 11:739–745

Halmos PR (1946) The theory of unbiased estimators. Ann Math Stat 17:34–43

Hoeffding W (1948) A class of statistics with asymptotically normal distribution. Ann Math Stat 19:293–325

Koroljuk VS, Borovskikh YV (1994) Theory of U-statistics. Kluwer, Dordrecht

Lee AJ (1990) U-statistics. Theory and practice. Marcel Dekker, New York

V

Validity of Scales

Elisabeth D. Svensson
Professor Emerita
Örebro University, Örebro, Sweden

The concepts of quality of measurements made by rating scales and multi-scale questionnaires are *validity* and *reliability*. Corresponding concepts for quantitative data (interval and ratio data) are accuracy and precision. A rating scale is *valid* if it measures what it is intended to measure in the specific study. The *validity* of self-estimated subjective phenomena is relative and cannot be assessed absolutely. The validity of a scale is study specific, and must be considered each time the scale or the ►questionnaire is chosen for a new study. Therefore there are various concepts of validity, each addressing a specific type of quality assessment. The main concepts are *criterion, construct*, and *content* validity, but a large number of sub concepts are used. The meaning of these concepts is not univocal and depends on applications and research paradigms. *Criterion validity* refers to the conformity of a scale to a true state or a gold standard, and depending on the purpose of the study sub concepts like *clinical, predictive* and *concurrent validity* will be used.

Construct validity refers to the consistency between scales having the same theoretical definition in the absence of a true state or a gold standard. Sub concepts like *convergent, descriptive, discriminant, divergent, factorial, translation validity and parallel reliability* have been used in studies. *Biologic validity* refers to the closeness of scale assessments to the hypothesized expectation when comparing with other measures in a specific population. Discriminative rating scales are used to distinguish between individuals or groups, when no external criterion is available, then *discriminant validity* is to be assessed. *Parallel reliability* refers to the interchangeability of scales.

The concept *content validity* refers to the completeness of the scale or multi-scale questionnaire in the coverage of important areas. Sub concepts like *face, ecological, decision, consensual, sampling validity, comprehensiveness* and *feasibility* have been used.

Assessments on *rating scales* generate ordinal data having rank-invariant properties only, which means that the responses indicate a rank order and not a mathematical value. The results of statistical treatments of data must not being changed when relabeling the ordered responses. Appropriate statistical methods for evaluation of criterion and construct validity often refer to the order consistency or to the relationship between the scales of comparison.

The scatter plot of 48 paired assessments of perceived back pain on a visual analogue scale (VAS) and on a verbal descriptive scale (VDS-5) having five ordered categorical responses is shown the Fig. 1.

As evident from the plot there is a large overlapping between the assessments. The probability of discordance in paired observations (X,Y),

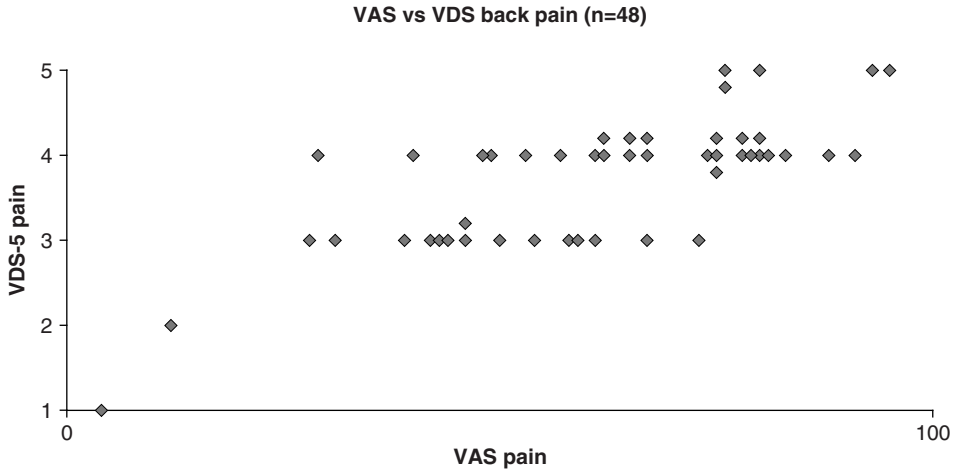
$$P\left[(X_\ell < X_k) \cap (Y_\ell > Y_k)\right] + P\left[(X_\ell > X_k) \cap (Y_\ell < Y_k)\right],$$

is estimated by the empirical measure of disorder D. In this case D equals 0.07, which means that 7% of all possible combinations of different pairs are disordered. The expected pattern of complete order consistency, the rank-transformable pattern of agreement (RTPA), is constructed by pairing off the two sets of distributions of data against each other. The measure of disorder expresses the observed dispersion of pairs from this order consistent distribution of inter-changeability between the scales. The cut-off response values for inter-scale calibration are also provided, and it is obvious that there is no linear correspondence between VAS and discrete scale assessments (see Fig. 2).

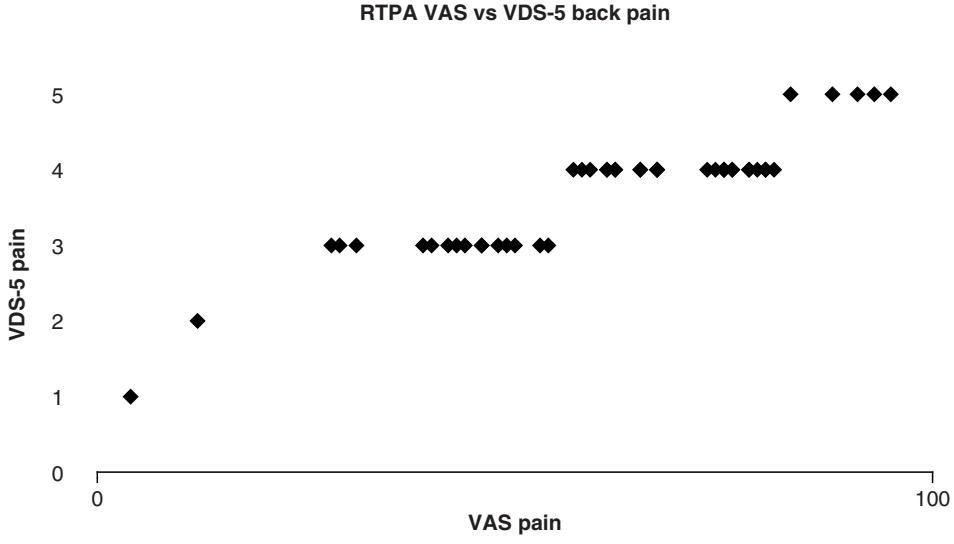
There are other measures that could be applied to evaluation of various kinds of validity of scales. Dependent on the purpose the Spearman rank-order correlations coefficient, The Goodman-Kruskal's gamma, the Kendall's tau-b (see ►Kendall's Tau), the Somers delta or the Stuart's tau-c could be suitable. Spearman rank order correlation coefficient is a commonly used non-parametric measure of association. However, a strong association does not necessarily mean a high level of order consistency, and does not indicate that two scales are interchangeable.

The Pearson correlation coefficient, the t-test and the ►Analysis of Variance are also common in validity studies. A serious drawback is that these methods assume normally

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,



Validity of Scales. Fig. 1 The distribution of paired assessments of back pain on a visual analogue pain scale and a five point verbal descriptive pain scale



Validity of Scales. Fig. 2 The rank-transformable pattern of agreement, RTPA, uniquely defined by the two sets of frequency distributions of data in Fig. 1

distributed quantitative data, and such requirements are not met by data from rating scales. When applying statistical methods on data that do not have the assumed properties then the results run the risk of being invalid and unreliable.

About the Author

For biography *see* the entry ►Ranks.

Cross References

►Analysis of Variance
►Correlation Coefficient
►Kendall's Tau
►Nonparametric Statistical Inference
►Parametric Versus Nonparametric Tests
►Rating Scales
►Scales of Measurement and Choice of Statistical Methods
►Student's *t*-Tests

References and Further Reading

Svensson E (2000a) Comparison of the quality of assessments using continuous and discrete ordinal rating scales. Biometrical J 42:417–434

Svensson E (2000b) Concordance between ratings using different scales for the same variable. Stat Med 19(24):3483–3496

Svensson E, Schillberg B, Kling AM, Nyström B (2009) The balanced inventory for spinal disorders. The validity of a disease specific questionnaire for evaluation of outcomes in patients with various spinal disorders. Spine 34(18):1976–1983

Variables

Rabija Somun- Kapetanović
Professor
University of Sarajevo, Sarajevo, Bosnia and Herzegovina

A variable is a characteristic that can take several values of a set of possible data upon which a measure or a quality can be applied. Thus, a variable varies in value among subjects in a sample or population. Each subject of the observed set has a particular value for a variable.

Examples of variables are gender (with values being female and male), nationality (American, French, German, . . .), level of education (Ph.D., Master, Bachelor, Baccalaureate, . . .), number of children in a family (0, 1, 2, . . .), and annual income in Euros.

Variables can be classified in many ways and terminology varies between different fields. For example, we may classify variables as (a) *qualitative* or *quantitative*, (b) *independent* or *dependent*, (c) *univariate* (one dimensional) or *multivariate* (multidimensional), (d) *latent* (hidden) or *observed*, (e) *endogenous* or *exogenous*, (f) *explanatory, intermediate*, or *response*, and (g) *monitoring* or *moderating*. Classification is further complicated because mixtures of different types occur quite commonly. Depending on the nature of measurement there are also different scales for measuring the variable: *nominal, ordinal, interval*, and *ratio* measurements. The scale of measurement determines the amount of information contained in a set of data and shows the most appropriate statistical methods for analyzing that data. We will focus only on the distinction between qualitative and quantitative variables.

Qualitative Variables

Qualitative variables contain values that express a quality in a descriptive way, such as sex, nationality, or level of education. Qualitative variables, also called categorical variables, are divided into nominal and ordinal ones.

- *Nominal variables* imply the fact that the labels are unordered. Indeed, there is no criterion that allows determining a label (a value) to be greater than or smaller than other labels. Thus the gender and nationality are nominal variables. Accordingly, the marital status, name, and country of residence are also nominal variables, which are measured on a nominal scale.
- *Ordinal variables* represent labels that can be ordered according to some logical criterion. Hence, the level of education is an ordinal variable as are opinions concerning a subject (excellent, good, poor. . .). The set of labels that satisfies a hierarchical criterion and is measured on an ordinal scale is an ordinal variable.

Mathematical operations are not allowed in qualitative variables, but for ordinal variables, counting and comparison are permitted. Qualitative nominal and ordinal variables can be numerically encoded. Indeed, for instance, it can be supposed that the variable "gender" takes the value 1 for female and 2 for male. Also, if the variable considered is an opinion, the value 1 can be used to represent excellent, 2 for good, and 3 for poor. However, these numbers have no meaning as such and cannot be the object of any mathematical operations.

Quantitative Variables

Quantitative variables are expressed through measurable values, that is, in terms of numbers. They can be measured on an interval or ratio scale and can be classified as either discrete or continuous.

- *Discrete variables* take only a countable and usually finite number of real values that are the result of a counting process. These variables typically take integer values. For instance, discrete variables are the number of children in a family, the number of students attending a class, and the number of employees in a company.
- *Continuous variables* take an infinite number of real values arising from a measuring process. In practice the number of values that continuous variables can take depends on the precision of the measuring instruments. For instance, the height or the weight is expressed in decimal points when they are measured.

In practice, it is sometimes difficult to distinguish discrete and continuous variables because of the way they are actually measured. Quantitative variables can be used to perform more admissible mathematical operations. The use of quantitative variables is widespread because it contributes to obtaining important results as more statistical methods for analyzing can be applied.

Cross References

►Data Analysis
►Dummy Variables
►Exchangeability
►Instrumental Variables

►Random Variable
►Rating Scales
►Scales of Measurement
►Scales of Measurement and Choice of Statistical Methods

References and Further Reading

Anderson DR, Sweeney DJ, Williams TA (2008) Essentials of statistics for business and economics, 5th edn. Cincinnati, South Western Educational Publishing

Berenson ML, Levine DM, Krehbiel TC (2007) Basic business statistics: Concepts and applications, 11th edn. New Yersey, Prentice Hall

Dehon C, Droesbeke JJ, Vermandele C (2008) Eléments de statistique, 5th edn. Editions de l'Universié de Bruxelles, Ellipses, Bruxelles, Paris

Wonnacott TH, Wonnacott RJ (1990) Introductory statistics for business and economics, 4th edn. Wiley, New York

Variance

Abdulbari Bener[1], Miodrag Lovric[2]
[1]Professor
Weill Cornell Medical College, Doha, Qatar
[2]Professor
University of Kragujevac, Kragujevac, Serbia

The term "variance" was coined by Ronald Fisher in 1918 in his famous paper on population genetics, The Correlation Between Relatives on the Supposition of Mendelian Inheritance, published by Royal Society of Edinburgh: "It is … desirable in analyzing the causes of variability to deal with the square of the standard deviation as the measure of variability. We shall term this quantity the Variance …" (p. 399). Interestingly, according to O. Kempthorne, this paper was initially rejected by the Royal Society of London, "probably the reason was that it constituted such a great advance on the thought in the area that the reviewers were unable to make a reasonable assessment."

The variance of a random variable (or a data set) is a measure of variable (data) dispersion or spread around the mean (expected value).

Definition Let X be a random variable with second moment $E(X^2)$ and let $\mu = E(X)$ be its mean. The variance of X is defined by (see, e.g., Feller 1968, p. 228)

$$Var(X) = E\left[(X-\mu)^2\right] = E(X^2) - \mu^2. \tag{1}$$

The variance of a random variable is also frequently denoted by $V(X)$, σ_X^2 or simply σ^2, when the context is clear. The positive square root of variance is called the standard deviation.

From (1), the variance of X can be interpreted as the "mean of the *squares* of deviations from the mean" (Kendall 1945, p. 39). Since the deviations are squared, it is clear that variance cannot be negative. Variance is a measure of dispersion "since if the values of a random variable X tend to be far from their mean, the variance of X will be larger than the variance of a comparable random variable Y whose values tend to be near their mean" (Mood et al. 1974, p. 67). It is obvious that a constant has variance 0, since there is no spread. Because the deviations are squared, the variance is expressed in the original units squared (inches^2, euro^2) which are difficult to interpret.

To compute the variance of a random variable, it is required to know the probability distribution of X. If X is a discrete random variable, then

$$Var(X) = \sum_i (x_i - \mu)^2 P(X = x_i) = \sum_i x_i^2 P(X = x_i) - \mu^2. \tag{2}$$

When X is a continuous random variable with probability density function $f(x)$, then

$$Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx - \mu^2. \tag{3}$$

Example 1 If X has a Uniform distribution on $[a, b]$, with pdf $1/(b-a)$, then

$$E(X) = \frac{1}{b-a}\int_a^b x dx = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2},$$

and

$$E(X^2) = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}.$$

Hence the variance is equal to

$$Var(X) = E(X^2) - \mu^2 = \frac{(b-a)^2}{12}.$$

The following table provides expressions for variance for some standard univariate discrete and continuous probability distributions.

The Cauchy distribution possesses neither mean nor variance.

Next, we list some important properties of variance.

1. The variance of a constant is 0; in other words, if all observations in the data set are identical, the variance takes its minimum possible value, which is zero.
2. If b is a constant then

Distribution	Notation	Variance
Bernoulli	$Be(p)$	pq
Binomial	$Bin(n,p)$	npq
Geometric	$Ge(p)$	q/p^2
Poisson	$Po(\lambda)$	λ
Uniform	$U(a,b)$	$(b-a)^2/12$
Exponential	$Exp(\lambda)$	$1/\lambda^2$
Normal	$N(\mu,\sigma)$	σ^2
Standard Normal	$N(0,1)$	1
Student	$t(\nu)$	$\nu(\nu-2)$ for $\nu>2$
F	$F(\nu_1,\nu_2)$	$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ for $\nu_2>4$
Chi-square	$Chi(\nu)$	2ν

$$Var(X+b) = Var\,X,$$

which means that adding a constant to a random variable does not change the variance.

3. If a and b are constants, then

$$Var(aX+b) = a^2\,Var\,X$$

4. If two variables X and Y are independent, then

$$Var(X+Y) = Var\,X + Var\,Y$$
$$Var(X-Y) = Var\,X + Var\,Y$$

5. The previous property can be generalized, i.e., the variance of the sum of independent random variables is equal to the sum of variances of these random variables

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i).$$

This result is called Bienaymé equality (see Loève 1977, p. 12, or Roussas p. 171).

6. If two random variables X and Y are independent and a and b are constants, then

$$Var(aX+bY) = a^2\,Var\,X + b^2\,Var\,Y.$$

In practice, the variance of a population, σ^2, is usually not known, and therefore it can only be estimated using the information contained in a sample of observations drawn from that population. If $x_1, x_2, \ldots, x_n$ is a random sample of size n selected from a population with mean μ, then the sample variance is usually denoted by s^2 and is defined by

$$S^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}, \tag{4}$$

where $\bar{x}$ is the sample mean. The sample variance depicts the dispersion of sample observations around the sample mean. The squared deviations in (4) are divided by $n-1$, not by n, in order to obtain the unbiased estimator of the population variance, $E(s^2) = \sigma^2$. The factor $1/(n-1)$ increases sample variance enough to make it unbiased. This factor is known as Bessel's correction (after Friedrich Bessel). Although the sample variance defined as in (4) is an unbiased estimator of population variance, the same does not relate to its square root, standard deviation; the sample standard deviation is a *biased* estimate of the population standard deviation.

Example 2 The first column of the following table contains first five measurements of the speed of light in suitable units (000 km/s) from the classical experiments performed by Michelson in 1879 (data obtained from the Ernest N. Dorsey's 1944 paper "The Velocity of Light").

x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	x_i^2
299.85	−0.048	0.002304	89,910.0225
299.74	−0.158	0.024964	89,844.0676
299.90	0.002	0.000004	89,940.0100
300.07	0.172	0.029584	90,042.0049
299.93	0.032	0.001024	89,958.0049
Σ 1499.49	0.000	0.057880	449,694.1099

Since the sample mean is equal to $\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{1499.49}{5} = 299.898$ using the formula given in (4) results in the variance value

$$S^2 = \frac{\sum(x_i-\bar{x})^2}{n-1} = \frac{0.057880}{4} = 0.01447.$$

In the past, instead of the "definitional" formula (4), the following (so-called shorthand) formula was commonly used, but it has become obsolete with the wide access of

statistical software, spreadsheets, and Internet java applets:

$$S^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{449,694.1099 - \frac{1,499.49^2}{5}}{4}$$

$$= 0.01447.$$

About the Author

Abdulbari Bener, Ph.D., has joined the Department of Public Health at the Weill Cornell Medical College as Research Professor of Public Health. Professor Bener is Director of the Medical Statistics and Epidemiology Department at Hamad Medical Corporation/Qatar. He is also an advisor to the World Health Organization and Adjunct Professor and Coordinator for the postgraduate and master public health programs (MPH) of the School of Epidemiology and Health Sciences, University of Manchester. He is Fellow of Royal Statistical Society (FRSS) and Fellow of Faculty of Public Health (FFPH). Dr Bener holds a Ph.D. degree in Medical Statistics (Biometry) and Genetics from the University College of London, and a B.Sc. degree from Ankara University, Faculty of Education, Department of Management, Planning and Investigation. He completed research fellowships in the Departments of Genetics and Biometry and Statistics and Computer Sciences at the University College of London. He has held academic positions in public health, epidemiology, and statistics at universities in Turkey, Saudi Arabia, Kuwait, the United Arab Emirates, Qatar, and England. Professor Bener has been author or coauthor of more than 430 published journal articles; Editor, Associate Editor, Advisor Editor, and Asst. Editor for several Journals; and Referee for over 23 journals. He has contributed to more than 15 book chapters and supervised thesis of 40 postgraduate students (M.Sc., MPH, M.Phil. and Ph.D.).

Cross References

►Expected Value
►Mean Median and Mode
►Mean, Median, Mode: An Introduction
►Semi-Variance in Finance
►Standard Deviation
►Statistical Distributions: An Overview
►Tests for Homogeneity of Variance

References and Further Reading

Fisher R (1918) The correlation between relatives on the supposition of mendelian inheritance. Philos Trans Roy Soc Edinb 52: 399–433

Dorsey EN (1944) The velocity of light. T Am Philos Soc 34(Part 1): 1–110, Table 22

Feller W (1968) An introduction to the probability theory and its applications, 3rd edn. Wiley, New York

Kempthorne O (1968) Book reviews. Am J Hum Genet 20(4): 402–403

Kendall M (1945) The advanced theory of statistics. Charles Griffin, London

Loève M (1977) Probability theory I, 4th edn. Springer, New York

Mood AM, Graybill FA, Boes DC (1974) Introduction to the theory of statistics, 3rd edn. McGraw-Hill, London

Roussas G (1997) A course in mathematical statistics, 2nd edn. Academic, Hardcover

Variation for Categorical Variables

Tarald O. Kvålseth
Professor Emeritus
University of Minnesota, Minneapolis, MN, USA

By definition, a categorical variable has a measurement scale that consists of a set of categories, either nominal (i.e., categories without any natural ordering) or ordinal (i.e., categories that are ordered). For a categorical variable with n categories and the probability distribution $P_n = (p_1, \ldots, p_n)$ where $p_i \geq 0$ for $i = 1, \ldots, n$ and $\sum_{i=1}^{n} p_i = 1$, some measurement of variation (dispersion) is sometimes of interest. Any such measure will necessarily depend on whether the variable (or set of categories or data) is nominal or ordinal.

Nominal Case

In the nominal case, variation is generally considered to increase strictly as the probabilities (or proportions) $p_i (i = 1, \ldots, n)$ become increasingly equal, with the variation being maximum for the uniform distribution $P_n^1 = (1/n, \ldots, 1/n)$ and minimum for the degenerate distribution $P_n^0 = (0, \ldots, 0, 1, 0, \ldots, 0)$ and for any given n. In terms of *majorization theory* (Marshall and Olkin 1979, Ch. 1), this requires that a nominal variation measure be strictly Schur-concave. Another typically imposed requirement is that the measure should be normed to the [0,1] - interval for ease of interpretation.

The best known measures meeting those two requirements are the *index of qualitative variation* (IQV), the normed entropy (H^*), and the normed form of the *variation ratio* (VR^*) defined as follows (e.g., Weisberg 1992):

$$IQV = \left(\frac{n}{n-1}\right)\left(1 - \sum_{i=1}^{n} p_i^2\right), \qquad (1)$$

$$H^* = \frac{-\sum_{i=1}^{n} p_i \log p_i}{\log n}, \tag{2}$$

$$VR^* = \left(\frac{n}{n-1}\right)(1 - \max\{p_1, \ldots, p_n\}). \tag{3}$$

Note that the logarithmic terms in (2) can be to any base since such terms appear both in the numerator and denominator. Those three measures range in value from 0 (when $P_n = P_n^0$) to 1 (when $P_n = P_n^1$) where

$$P_n^0 = (0, \ldots, 0, 1, 0, \ldots, 0), \quad P_n^1 = (1/n, \ldots, 1/n), \tag{4}$$

and for any given n. The measures in (1) and (2) can be seen to be strictly Schur-concave, while VR^* in (3) is Schur-concave but not strictly so (see Marshall and Olkin 1979, Ch. 3). Also, while IQV and H^* are continuous functions of all the probability components $p_1, \ldots, p_n$, VR^* is a function of only the modal probability.

Although IQV and H^* in (1)–(2) have a number of nice properties, they both lack an important one: they both overstate the true extent of variation. To illustrate this fact, consider $P_2 = (0.75, 0.25)$ for which each element is the arithmetic mean of the corresponding elements of $P_2^0 = (1, 0)$ and $P_2^1 = (0.5, 0.5)$ so that one would reasonably expect that the variation for this P_2 should be 0.5, i.e., the mean of the variations for P_2^0 and P_2^1 (i.e., 0 and 1, respectively). However, one finds the $IQV(0.75, 0.25) = 0.75$ and $H^*(0.75, 0.25) = 0.81$. In order for a variation measure to take on reasonable numerical values, and thereby avoid invalid and misleading results and conclusions, Kvälseth (1995) proposed the following *coefficient of nominal variation* (CNV) as a simple transformation of IQV:

$$CNV = 1 - \sqrt{1 - IQV}. \tag{5}$$

Besides having the same types of properties as IQV, this CNV takes on values that appear to be entirely reasonable throughout the $[0, 1]$ – interval. For instance, $CNV(0.75, 0.25) = 0.50$ as is only reasonable.

Note also that the CNV in (5) can be expressed in terms of metric distances as follows. In terms of the Euclidean distance $d_2(X, Y) = \left[\sum_{i=1}^{n}(x_i - y_i)^2\right]^{1/2}$ between the two points $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$, CNV can be expressed as

$$CNV = 1 - \frac{d_2\left(P_n, P_n^1\right)}{d_2\left(P_n^0, P_n^1\right)}, \tag{6}$$

for any distribution P_n, with P_n^0 and P_n^1 defined in (4). That is, CNV is the relative extent to which the Euclidean distance $d_2\left(P_n, P_n^1\right)$ is less than its maximum possible value. Or, CNV is the relative (metric) proximity of P_n to P_n^1. Thus, the expression in (6) provides CNV with a reasonable interpretation and a solid basis.

In terms of the standard deviation s of $p_1, \ldots, p_n$ (using the usual divisor $n - 1$), it is readily seen that CNV is given by

$$CNV = 1 - s\sqrt{n}. \tag{7}$$

Similarly, in terms of the pair-wise differences between the p_i's,

$$CNV = 1 - \left(\frac{1}{n-1}\sum_{1 \le i < j \le n}\sum |p_i - p_j|^2\right)^{1/2}. \tag{8}$$

A parameterized family of such difference-based variation measures may also be formulated (Kvälseth 1998), but no other family member appears to be superior to CNV.

Ordinal Case

In the ordinal case, and when the order information is accounted for, it is considered that variation is zero for the degenerate distribution P_n^0 and maximal for the polarized distribution $P_n^{(1)}$ defined as

$$P_n^0 = (0, \ldots, 0, 1, 0, \ldots, 0), \quad P_n^{(1)} = (0.5, 0, \ldots, 0, 0.5), \tag{9}$$

(see, e.g., Leik 1966; Weisberg 1992). When the n categories are ordered, it makes sense to use cumulative probabilities $F_i = \sum_{j=1}^{i} p_j$ for $i = 1, \ldots, n$ with $F_n = 1$. Thus, for any given $P_n = (p_1, \ldots, p_n)$, and for the particular distributions in (9), the following cumulative distributions can be defined:

$$F_{(n)} = (F_1, \ldots, F_{n-1}, 1), \quad F_{(n)}^0 = (0, \ldots, 0, 1, 1, \ldots, 1),$$
$$F_{(n)}^{(1)} = (.5, \ldots, .5, 1). \tag{10}$$

A measure of variation for ordinal categorical data may then be based on cumulative probabilities.

The first such proposed measure appears to be Leik's (1966) ordinal variation measure (LOV), which can be expressed as

$$LOV = 1 - \frac{\sum_{i=1}^{n-1} |2F_i - 1|}{n - 1}, \tag{11}$$

which ranges in value from 0 to 1, equals 0 for $F_{(n)}^{(0)}$ and 1 for $F_{(n)}^{(1)}$ in (10). An alternative measure is the *coefficient of ordinal variation* (COV) by Kvälseth (1995a,b) defined,

and somewhat analogous to *CNV* in (5), as

$$COV = 1 - \sqrt{1 - \Delta^*}, \quad \Delta^* = \frac{2}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n}|i-j|p_i p_j$$

$$= \frac{4}{n-1}\sum_{i=1}^{n-1} F_i(1-F_i) \tag{12}$$

where $COV \in [0,1]$, $COV\left(F_{(n)}^0\right) = 0$, and $COV\left(F_{(n)}^{(1)}\right) = 1$. The *COV* can also be expressed as

$$COV = 1 - \left(\frac{\sum_{i=1}^{n-1}|2F_i - 1|^2}{n-1}\right)^{1/2}. \tag{13}$$

It would appear from (11) and (13) that *LOV* and *COV* are both members of the same family. In fact, expressed in terms of an α – order arithmetic mean, both measures belong to the family of ordinal variation measures

$$OV_\alpha = 1 - \left(\frac{\sum_{i=1}^{n-1}|2F_i - 1|^\alpha}{n-1}\right)^{1/\alpha}, \quad -\infty < \alpha < \infty \tag{14}$$

where $LOV = OV_1$, and $COV = OV_2$. Furthermore, in terms of the Minkowski metric distance of order $\alpha \geq 1$ (i.e., $d_\alpha(X,Y) = \left(\sum_{i=1}^{n}|x_i - y_i|^\alpha\right)^{1/\alpha}$),

$$OV_\alpha = 1 - \frac{d_\alpha\left(F_{(n)}, F_{(n)}^{(1)}\right)}{d_\alpha\left(F_{(n)}^0, F_{(n)}^{(1)}\right)}, \quad \alpha \geq 1 \tag{15}$$

with $F_{(n)}, F_{(n)}^{(0)}$, and $F_{(n)}^{(1)}$ defined in (10). Clearly, $d_\alpha\left(F_{(n)}F_{(n)}^{(1)}\right) \leq d_\alpha\left(F_{(n)}^0, F_{(n)}^{(1)}\right)$ since $|F_i - 0.5| \leq 0.5$ for all i. Thus, $OV_\alpha \in [0,1]$, $OV_\alpha\left(F_{(n)}^0\right) = 0$, and $OV_\alpha\left(F_{(n)}^{(1)}\right) = 1$. The expressions in (14) – (15), especially (15), provide interpretations and bases for *LOV* and *COV*, with *LOV* and *COV* being based, respectively, on city-block (Hamming) distances ($\alpha = 1$) and Euclidean distances ($\alpha = 2$) (see also Blair and Lacy 1996).

Statistical Inferences

For a generic variation measure V, consider now (a) that $V(P_n)$ is the sample value based on the distribution $P_n = (p_1, \ldots, p_n)$ of sample probabilities n_i/N for $i = 1, \ldots, n$ with sample size $N = \sum_{i=1}^{n} n_i$ and (b) that $V(\Pi_n)$ is the population value based on the corresponding population distribution $\Pi_n = (\pi_1, \ldots, \pi_n)$. It may then be of interest to construct a confidence interval or test an hypothesis about $V(\Pi_n)$. This can be done using the *delta method* (Agresti 2002, Ch. 14). Accordingly, under multinomial sampling with N reasonably large, $V(P_n)$ is approximately normally distributed with mean $V(\Pi_n)$ and estimated variance

$$\hat{\sigma}_V^2 = \frac{1}{N}\left[\sum_{i=1}^{n} p_i \hat{\phi}_{V_i}^2 - \left(\sum_{i=1}^{n} p_i \hat{\phi}_{V_i}\right)^2\right], \tag{16}$$

where

$$\hat{\phi}_{V_i} = \left.\frac{\partial V(\Pi_n)}{\partial \pi_i}\right|_{\pi_i = p_i}, \quad i = 1, \ldots, n \tag{17}$$

i.e., $\hat{\phi}_{V_i}$ is the partial derivative of $V(\Pi_n)$ with respect to π_i, which is then replaced with p_i, for $i = 1, \ldots, n$.

In the case of *CNV* in (5), it follows from (17) (with $V = CNV$) that

$$\hat{\phi}_{CNV_i} = \frac{-n}{(n-1)(1-CNV)} p_i, \quad i = 1, \ldots, n,$$

so that; from (16),

$$\hat{\sigma}_{CNV}^2 = \left(\frac{1}{N}\right)\left(\frac{n}{(n-1)(1-CNV)}\right)^2\left[\sum_{i=1}^{n} p_i^3 - \left(\sum_{i=1}^{n} p_i^2\right)^2\right]. \tag{18}$$

For the case of *COV* in (13), and with $V = COV$, it is found from (17) that

$$\hat{\phi}_{COV_i} = \begin{cases} \frac{2}{(n-1)(1-COV)}\left[n - i - 2\sum_{j=1}^{n-1} F_j\right], & i = 1, \ldots, n-1 \\ 0, & i = n \end{cases} \tag{19}$$

which can then be used to compute $\hat{\sigma}_{COV}^2$ from (16).

As a numerical example, consider the respective multinomial frequencies $n_i = 20, 15, 5, 60$ so that, with $N = 100$, $P_4(0.20, 0.15, 0.05, 0.60)$. From (1) and (5), $IQV = 0.77$ and $CNV = 0.52$. From (18), with $CNV = 0.5170$, $\hat{\sigma}_{CNV}^2 = 0.0036$. Therefore, an approximate 95% confidence interval for the population measure $CNV(\Pi_4)$ becomes $0.5170 \pm 1.96\sqrt{0.0036}$ or $(0.40, 0.63)$. If the four categories are ordinal so that $F_i = 0.20, 0.35, 0.40, 1$ for $i = 1, \ldots, 4$, it follows from (13) and (19) that $COV = 0.5959$ and $\hat{\phi}_{COV_i} = 1.8148, 0.8249, 0.3300, 0$ for $i = 1, \ldots, 4$ so that, from (16), with $V = COV$, $\hat{\sigma}_{COV}^2 = 0.0051$. Therefore, an approximate 95% confidence interval for $COV(\Pi_4)$ becomes $0.5959 \pm 1.96\sqrt{0.0051}$, or $(0.46, 0.74)$.

About the Author

For biography *see* the entry ▶Entropy.

Cross References

►Association Measures for Nominal Categorical Variables
►Categorical Data Analysis
►Scales of Measurement
►Variables

References and Further Reading

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken, NJ

Blair J, Lacy MG (1996) Measures of variation for ordinal data as functions of the cumulative distribution. Percept Mot Skills 82:411–418

Kvälseth TO (1995a) Coefficients of variation for nominal and ordinal categorical data. Percept Mot Skills 80:843–847

Kvälseth TO (1995b) Comment on the coefficient of ordinal variation. Percept Mot Skills 81:621–622

Kvälseth TO (1998) On difference – based summary measures. Percept Mot Skills 87:1379–1384

Leik RK (1966) A measure of ordinal consensus. Pacific Sociol Rev 9:85–90

Marshall AW, Olkin I (1979) Inequalities: theory of majorization and its applications. Academic Press, San Deigo, CA

Weisberg HF (1992) Central tendency and variability. (Sage University Paper Series No. 07-083). Sage Publications, Newbury Park, CA

Vector Autoregressive Models

Helmut Lütkepohl
Professor of Econometrics
European University Institute, Firenze, Italy

Vector autoregressive (VAR) processes are popular in economics and other sciences because they are flexible and simple models for multivariate time series data. In econometrics they became standard tools when Sims (1980) questioned the way classical simultaneous equations models were specified and identified and advocated VAR models as alternatives. A textbook treatment of these models with details on the issues mentioned in the following introductory exposition is available in Lütkepohl (2005).

The Model Setup

The basic form of a VAR process is

$$y_t = Dd_t + A_1 y_{t-1} + \cdots + A_p y_{t-p} + u_t,$$

where $y_t = (y_{1t}, \ldots, y_{Kt})'$ (the prime denotes the transpose) is a vector of K observed time series variables, d_t is a vector of deterministic terms such as a constant, a linear trend and/or seasonal ►dummy variables, D is the associated parameter matrix, the A_i's are $(K \times K)$ parameter matrices attached to the lagged values of y_t, p is the lag order or VAR order and u_t is an error process which is assumed to be white noise with zero mean, that is, $E(u_t) = 0$, the covariance matrix, $E\left(u_t u_t'\right) = \Sigma_u$, is time invariant and the u_t's are serially uncorrelated or independent.

VAR models are useful tools for forecasting. If the u_t's are independent white noise, the minimum mean squared error (MSE) h-step forecast of y_{t+h} at time t is the conditional expectation given y_s, $s \le t$,

$$\begin{aligned} y_{t+h|t} &= E(y_{t+h}|y_t, y_{t-1}, \ldots) \\ &= Dd_{t+h} + A_1 y_{t+h-1|t} + \cdots + A_p y_{t+h-p|t}, \end{aligned}$$

where $y_{t+j|t} = y_{t+j}$ for $j \le 0$. Using this formula, the forecasts can be computed recursively for $h = 1, 2, \ldots$. The forecasts are unbiased, that is, the forecast error $y_{t+h} - y_{t+h|t}$ has mean zero and the forecast error covariance is equal to the MSE matrix. The 1-step ahead forecast errors are the u_t's.

VAR models can also be used for analyzing the relation between the variables involved. For example, Granger (1969) defined a concept of causality which specifies that a variable y_{1t} is causal for a variable y_{2t} if the information in y_{1t} is helpful for improving the forecasts of y_{2t}. If the two variables are jointly generated by a VAR process, it turns out that y_{1t} is not Granger-causal for y_{2t} if a simple set of zero restrictions for the coefficients of the VAR process are satisfied. Hence, Granger-causality is easy to check in VAR processes.

Impulse responses offer another possibility for analyzing the relation between the variables of a VAR process by tracing the responses of the variables to impulses hitting the system. If the VAR process is stable and stationary, it has a moving average representation of the form

$$y_t = D^* d_t + \sum_{j=0}^{\infty} \Phi_j u_{t-j},$$

where the Φ_j's are $(K \times K)$ coefficient matrices which can be computed from the VAR coefficient matrices A_i with $\Phi_0 = I_K$, the $(K \times K)$ identity matrix. This representation can be used for tracing the effect of a specific forecast error through the system. For example, if $u_t = (1, 0, \ldots, 0)'$, the coefficients of the first columns of the Φ_j matrices represent the marginal reactions of the y_t's. Unfortunately, these so-called *forecast error impulse responses* are often not of interest for economists because they may not reflect properly what actually happens in a system of variables. Given that the components of u_t are typically instantaneously correlated, such shocks or impulses are not likely to appear in isolation. Impulses or shocks of interest for economists are usually instantaneously uncorrelated. They are obtained from the forecast errors, the u_t's, by some transformation,

for example, $\varepsilon_t = Bu_t$ may be a vector of shocks of interest if the $(K \times K)$ matrix B is such that $\varepsilon_t \sim (0, \Sigma_\varepsilon)$ has a diagonal covariance matrix Σ_ε. The corresponding moving average representation in terms of the ε_t's becomes

$$y_t = D^* d_t + \sum_{j=0}^{\infty} \Theta_j \varepsilon_{t-j},$$

where $\Theta_j = \Phi_j B^{-1}$.

There are many B matrices with the property that Bu_t is a random vector with diagonal covariance matrix. Hence, there are many shocks ε_t of potential interest. Finding those which are interesting from an economic point of view is the subject of *structural VAR analysis*.

Estimation and Model Specification

In practice the process which has generated the time series under investigation is usually unknown. In that case, if VAR models are regarded as suitable, the lag order has to be specified and the parameters have to be estimated. For a given VAR order p, estimation can be conveniently done by equationwise ordinary ►least squares (OLS). For a sample of size T, $y_1, \ldots, y_T$, and assuming that in addition presample values $y_{-p+1}, \ldots, y_0$ are also available, the OLS estimator of the parameters $B = [D, A_1, \ldots, A_p]$ can be written as

$$\hat{B} = \left(\sum_{t=1}^{T} y_t Z_{t-1}' \right) \left(\sum_{t=1}^{T} Z_{t-1} Z_{t-1}' \right)^{-1},$$

where $Z_{t-1}' = \left(d_t', y_{t-1}', \ldots, y_{t-p}'\right)$. Under standard assumptions the estimator is consistent and asymptotically normally distributed. In fact, if the residuals and, hence, the y_t's are normally distributed, that is, $u_t \sim$ i.i.d.$\mathcal{N}(0, \Sigma_u)$, the OLS estimator is equal to the maximum likelihood (ML) estimator with the usual asymptotic optimality properties. If the dimension K of the process is large, then the number of parameters is also large and estimation precision may be low if a sample of typical size in macroeconomic studies is available for estimation. In that case it may be useful to exclude redundant lags of some of the variables from some of the equations and fit so-called subset VAR models. In general, if zero or other restrictions are imposed on the parameter matrices, other estimation methods may be more efficient.

VAR order selection is usually done by sequential tests or model selection criteria (see ►Model Selection). ►*Akaike's information criterion* (AIC) is, for instance, a popular model selection criterion (Akaike, 1973). It has the form

$$\mathrm{AIC}(m) = \log \det(\hat{\Sigma}_m) + 2mK^2/T,$$

where $\hat{\Sigma}_m = T^{-1} \sum_{t=1}^{T} \hat{u}_t \hat{u}_t'$ is the residual covariance matrix of a VAR(m) model estimated by OLS. The criterion consists of the determinant of the residual covariance matrix which tends to decline with increasing VAR order whereas the penalty term $2mK^2/T$, which involves the number of parameters, grows with m. The VAR order is chosen which optimally balances both terms. In other words, models of orders $m = 0, \ldots, p_{\max}$ are estimated and the order p is chosen such that it minimizes the value of AIC.

Once a model is estimated it should be checked that it represents the data features adequately. For this purpose a rich toolkit is available. For example, descriptive tools such as plotting the residuals and residual autocorrelations may help to detect model deficiencies. In addition, more formal methods such as tests for residual autocorrelation, conditional heteroskedasticity, nonnormality and structural stability or tests for parameter redundancy may be applied.

Extensions

If some of the time series variables to be modeled with a VAR have stochastic trends, that is, they behave similarly to a ►random walk, then another model setup may be more useful for analyzing especially the trending properties of the variables. Stochastic trends in some of the variables are generated by models with unit roots in the VAR operator, that is, $\det(I_K - A_1 z - \cdots - A_p z^p) = 0$ for $z = 1$. Variables with such trends are nonstationary and not stable. They are often called integrated. They can be made stationary by differencing. Moreover, they are called cointegrated if stationary linear combinations exist or, in other words, if some variables are driven by the same stochastic trend. Cointegration relations are often of particular interest in economic studies. In that case, reparameterizing the standard VAR model such that the cointegration relations appear directly may be useful. The so-called *vector error correction model* (VECM) of the form

$$\Delta y_t = D d_t + \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \cdots + \Gamma_{p-1} \Delta y_{t-p+1} + u_t$$

is a simple example of such a reparametrization, where Δ denotes the differencing operator defined such that $\Delta y_t = y_t - y_{t-1}$, $\Pi = -(I_K - A_1 - \cdots - A_p)$ and $\Gamma_i = -(A_{i+1} + \cdots + A_p)$ for $i = 1, \ldots, p-1$. This parametrization is obtained by subtracting y_{t-1} from both sides of the standard VAR representation and rearranging terms. Its advantage is that Π can be decomposed such that the cointegration relations are directly present in the model. More precisely, if all variables are stationary after differencing once, and there are $K - r$ common trends, then the matrix Π has rank r and can be decomposed as $\Pi = \alpha\beta'$, where α and β are

$(K \times r)$ matrices of rank r and β contains the cointegration relations. A detailed statistical analysis of this model is presented in Johansen (1995) (see also Part II of Lütkepohl (2005)).

There are also other extensions of the basic VAR model which are often useful and have been discussed extensively in the associated literature. For instance, in the standard model all observed variables are treated as endogenous, that is, they are jointly generated. This setup often leads to heavily parameterized models, imprecise estimates and poor forecasts. Depending on the context, it may be possible to classify some of the variables as exogenous and consider partial models which condition on some of the variables. The latter variables remain unmodeled.

One may also question the focus on finite order VAR models and allow for an infinite order. This can be done by either augmenting a finite order VAR by a finite order MA term or by accounting explicitly for the fact that the finite order VAR approximates some more general model. Details on these and other extensions are provided, e.g., by Hannan and Deistler (1988) and Lütkepohl (2005).

About the Author

Professor Lütkepohl was Dean of the School of Economics and Business Administration, Humboldt University, Berlin (1998–2000); Head of the Economics Department, European University Institute, Florence (2006–2008).

Cross References

►Akaike's Information Criterion
►Akaike's Information Criterion: Background, Derivation, Properties, and Refinements
►Asymptotic Normality
►Econometrics: A Failed Science?
►Forecasting: An Overview
►Likelihood
►Random Walk
►Residuals
►Seasonal Integration and Cointegration in Economic Time Series
►Time Series

References and Further Reading

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov B, Csáki F (eds) 2nd International Symposium on Information Theory, Académiai Kiadó, Budapest, pp 267–281

Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37:424–438

Hannan EJ, Deistler M (1988) The statistical theory of linear systems. Wiley, New York

Johansen S (1995) Likelihood-based inference in cointegrated vector autoregressive models. Oxford University Press, Oxford

Lütkepohl H (2005) New introduction to multiple time series analysis. Springer-Verlag, Berlin

Sims CA (1980) Macroeconomics and reality. Econometrica 48:1–48

W

Weak Convergence of Probability Measures

Milan Merkle
Professor, Faculty of Electrical Engineering
University of Belgrade, Belgrade, Serbia

Weak Convergence of Probability Measures on $\mathbb{R}^d$

Among several concepts of convergence that are being used in Probability theory, the weak convergence has a special role, as it is related not to values of random variables, but to their probability distributions. In a simplest case of a sequence $\{X_n\}$ of real valued random variables (or vectors with values in $\mathbb{R}^d$, $d \geq 1$) defined on probability spaces $(\Omega_n, \mathcal{F}_n, P_n)$, we say that a sequence $\{X_n\}$ converges *weakly* (or *in law*) to a random variable X if

$$\lim_{n\to+\infty} F_n(x) = F(x) \tag{1}$$

for each $x \in \mathbb{R}$ where the function F is continuous. Here F_n and F are distribution functions of X_n and X, respectively. The notation for this kind of convergence is $X_n \Rightarrow X$, or $X_n \xrightarrow{\mathcal{L}} X$. The convergence defined by (1) can be as well thought of as being a convergence of corresponding distributions, i.e., probability measures defined on $(\mathbb{R}^d, \mathcal{B}^d)$ by $\mu_n(B) = P_n(\{\omega \in \Omega_n \mid X_n(\omega) \in B\})$, where $B \in \mathcal{B}$ and $\mathcal{B}$ is a Borel sigma-field on $\mathbb{R}$. Hence we say also that a sequence of probability measures μ_n on $(\mathbb{R}^d, \mathcal{B}^d)$ converges weakly to μ, in notation $\mu_n \Rightarrow \mu$.

The following result is known as Lévy-Cramér Continuity Theorem.

Theorem 1 *Let μ_n be a sequence of probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$. Then $\mu_n \Rightarrow \mu$ if and only if the corresponding ►characteristic functions converge pointwise:*

$$\lim_{n\to+\infty} \mathrm{E}\, e^{i\langle t, X_n\rangle} = \mathrm{E}\, e^{i\langle t, X\rangle} \quad \textit{for every } t \in \mathbb{R}^d,$$

where X_n, X are random vectors with distributions μ_n and μ, respectively.

In $d = 1$, the Lévy's metric d_L (see Lévy 1937) is defined as a distance between two univariate distribution functions

$$d_L(F,G) = \inf\{\varepsilon > 0 \mid F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon \quad \text{for all } x \in \mathbb{R}\}.$$

If φ_F and φ_G are characteristic functions that correspond to F, G, respectively, then

$$d_L(F,G) \leq \frac{1}{\pi}\int_0^T |\varphi_F(t) - \varphi_G(t)|\frac{\mathrm{d}t}{t} + 2e\frac{\log T}{T}, \quad T > e.$$

The weak convergence $X_n \Rightarrow X$ is implied by convergence in probability, and consequently with all stronger notions of convergence (with probability one and in the pth mean, $p \geq 1$). To see that the weak convergence does not imply nearness of values of corresponding random variables, we may recall that for any symmetric random variable ($\mathcal{N}(0,1)$, say), X and $-X$ have the same distribution. However, for any given sequence μ_n of d-dimensional distributions such that $\mu_n \Rightarrow \mu$, there exists a probability space $(\Omega, \mathcal{F}, P)$ and random mappings X_n and X from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathcal{B}^d)$ such that μ_n and μ are distributions of X_n and X and also $\lim_{n\to+\infty} X_n = X$ almost surely. This result on separable metric spaces is obtained by Skorohod and later generalized to nets by Wichura (1970).

Some Typical Roles of Weak Convergence in Probability

The weak convergence appears in Probability chiefly in the following classes of problems.

- Knowing that $\mu_n \Rightarrow \mu$, we may replace μ_n by μ for n large enough. A typical example is the Central Limit Theorem (any of its versions), which enables us to conclude that the properly normalized sum of random variables has approximately a unit Gaussian law.
- Conversely, if $\mu_n \Rightarrow \mu$ then we may approximate μ with μ_n, for n large enough. A typical example of this sort is the approximation of Dirac's delta function (understood as a density of a point mass at zero) by, say triangle-shaped functions.
- It is not always easy to construct a measure with specified properties. If we need to show just its existence, sometimes we are able to construct a sequence

Miodrag Lovric (ed.), *International Encyclopedia of Statistical Science*, DOI 10.1007/978-3-642-04898-2,

(or a net) of measures which can be proved to be weakly convergent and that its limit satisfies the desired properties. For example, this procedure is usually applied to show the existence of the Wiener measure (the measure induced by Brownian Motion process (see ▶Brownian Motion and Diffusions) on the space of continuous functions).

The last mentioned example is related to measures in infinitely dimensional spaces, the case which usually arises in the context of assigning measures to the set of trajectories of a stochastic process. In fact, what is called weak convergence in Probability theory, is inherited from so called weak-star convergence in Topology, where it can be defined in duals of arbitrary topological spaces. In Probability theory we do not need such a generality, as we are interested only in spaces of measures. Since the spaces of measures always appear as dual spaces of continuous functions, the most general definition of weak convergence of probability measures is the following.

Definition 1 Let $\mathcal{X}$ be a topological space. Let $\mu_n (n = 1, 2, \ldots)$ and μ be probability measures defined on the Borel sigma field generated by open subsets of $\mathcal{X}$. We say that the sequence $\{\mu_n\}$ converges *weakly* to μ, in notation $\mu_n \Rightarrow \mu$ if

$$\lim_{n\to+\infty} \int_{\mathcal{X}} f(x)\, \mathrm{d}\mu_n(x) = \int_{\mathcal{X}} f(x)\, \mathrm{d}\mu(x)\,,$$

for every continuous and bounded real valued function $f : \mathcal{X} \mapsto \mathbb{R}$. The set of these functions is denoted by $\boldsymbol{C}(\mathcal{X})$. □

In terms of random variables, let X_n $(n = 1, 2, \ldots)$ and X be $\mathcal{X}$-valued random variables and let μ_n and μ be corresponding distributions. Then we say that the sequence X_n converges weakly to X and write $X_n \Rightarrow X$ if and only if $\mu_n \Rightarrow \mu$. A setup that yields infinitely dimensional spaces $\mathcal{X}$ is when X_n is a sequence of random processes and $\mathcal{X}$ is a space of functions where paths of X_n belong. Finally, in a general situation, we may think of nets $\{X_d\}$ and $\{\mu_d\}$ instead of sequences.

Weak Convergence of Measures on Metric Spaces

Let now $\mathcal{X}$ be a metric space and let $\mathcal{B}$ be the sigma field of Borel subsets of $\mathcal{X}$. Let $\mathcal{M}_1(\mathcal{X})$ be the set of all probability measures on $\mathcal{X}$.

Theorem 2 *Let μ_d be a net of probability measures on $\mathcal{X}$ and let μ_0 be a probability measure on $\mathcal{X}$. The following statements are equivalent* (Billingsley 1999; Stroock 1993):

(i) $\mu_d \Rightarrow \mu_0$, i.e., $\lim_d \int f\, \mathrm{d}\mu_d = \int f\, \mathrm{d}\mu_0$, *for each* $f \in \mathbf{C}(\mathcal{X})$.

(ii) $\lim_d \int f\, \mathrm{d}\mu_d = \int f\, \mathrm{d}\mu_0$ *for each* $f \in \mathbf{C}_u(\mathcal{X})$ (*uniformly continuous and bounded functions*).

(iii) $\overline{\lim}\, \mu_d(F) \le \mu_0(F)$ *for any closed set* $F \subset \mathcal{X}$.

(iv) $\underline{\lim}\, \mu_d(G) \ge \mu_0(G)$ *for each open set* $G \subset \mathcal{X}$.

(v) $\lim \mu_d(A) = \mu_0(A)$ *for each continuity set* A *for* μ_0 (that is, $\mu_0(\partial A) = 0$, *where* ∂A *is the boundary* of A).

(vi) $\overline{\lim} \int f\, \mathrm{d}\mu_d \le \int f\, \mathrm{d}\mu_0$ *for each upper semi-continuous and bounded from above function* $f : \mathcal{X} \mapsto \mathbb{R}$.

(vii) $\underline{\lim} \int f\, \mathrm{d}\mu_d \ge \int f\, \mathrm{d}\mu_0$ *for each lower semi-continuous and bounded from below function* $f : \mathcal{X} \mapsto \mathbb{R}$.

(viii) $\lim \int f\, \mathrm{d}\mu_d = \int f\, \mathrm{d}\mu_0$ *for each* μ_0*–a.e. continuous function* $f : \mathcal{X} \mapsto \mathbb{R}$.

In concrete metric spaces, the conditions can be checked to hold only for some special families of sets, so called *convergence determining families*. For example, a convergence determining family in $\mathbb{R}$ is a family of sets of the form $(-\infty, b]$, $b \in \mathbb{R}$, and using this family in the condition (v), we get the standard definition from the beginning of section "▶Weak Convergence of Probability Measures on $\mathbb{R}^d$". Similarly, it can suffice to check condition (i) only for special families of functions – Theorem 1 gives an example of such a family.

If $\mathcal{X}$ is a separable metric space, the topology of weak convergence of probability measures is metrizable by the metric

$$d(P, Q) = \inf\{\varepsilon > 0 \mid Q(B) \le P(B^\varepsilon) + \varepsilon,\ P(B) \le Q(B^\varepsilon) + \varepsilon, \quad B \in \mathcal{B}\},$$

where $B^\varepsilon = \{x \in \mathcal{X} \mid d(x, B) < \varepsilon\}$, and $\mathcal{B}$ is the Borel sigma algebra on $\mathcal{X}$. This metric is called Prohorov's metric, and it is a generalization of Lévy's metric from section "▶Weak Convergence of Probability Measures on $\mathbb{R}^d$". There are metrics which are known to be equivalent to Prohorov's metrics (see, for example, [Stroock 1993, p.117]).

Relative Compactness, Tightness and Prohorov's Theorem

Let $\mathcal{X}$ be a metric space, $\mathcal{B}$ a Borel sigma-algebra generated by open subsets of $\mathcal{X}$. In infinitely dimensional metric spaces, the weak convergence of finite dimensional distributions alone is not sufficient condition for weak convergence of measures. The additional condition is relative compactness.

Definition 2 We say that a set $\mathcal{P}$ of probability measures on $(\mathcal{X}, \mathcal{B})$ is relatively compact if any sequence of probability measures $P_n \in \mathcal{P}$ contains a subsequence P_{n_k} which converges weakly to a probability measure in $\mathcal{M}_1(X)$.

Theorem 3 *Let $\{\mu_n\}$ be a relatively compact sequence of probability measures on $\mathcal{X}$. If all finite-dimensional distributions converge weakly to corresponding finite-dimensional distributions of a measure μ, then $\mu_n \Rightarrow \mu$.*

Hence, an usual procedure to show weak convergence on a metric space is to first show convergence of finite dimensional distributions (via ►characteristic functions), and then to prove relative compactness. If $\mathcal{X}$ is compact, then any set $\mathcal{P}$ of probability measures is relatively compact. Otherwise, we need some conditions which are easier to check, a convenient tool is the notion of *tightness*.

Definition 3 Let $\mathcal{P}$ be a set of probability measures on $(\mathcal{X}, \mathcal{B})$. We say that $\mathcal{P}$ is *tight* if for any $\varepsilon > 0$ there is a compact set $K \subset \mathcal{X}$ such that $\mu(K') \le \varepsilon$ for any $\mu \in \mathcal{P}$. □

Next theorem links tightness with relative compactness.

Theorem 4 (Prohorov 1956) *(a) Any tight set of measures in arbitrary metric space is relatively compact. (b) If X is a complete separable metric space, then any relatively compact set of probability measures is tight.*

In particular metric spaces, it is useful to have simpler equivalent conditions for tightness. For example, observe the metric space $C[0,1]$ of continuous functions defined on $[0,1]$, with the metric of uniform convergence, $d(x,y) = \sup_{t\in[0,1]} |x(t) - y(t)|$. Then a sequence $\{\mu_n\}$ of probability measures (defined on Borel sets of this metric space) is tight if an only if

$$\lim_{\substack{K\to+\infty \\ n\to+\infty}} \mu_n\{x \in C[0,1] \mid |x(0)| \ge K\} = 0 \qquad \text{and}$$

$$\lim_{\delta\to 0} \limsup_{n\to+\infty} \mu_n\{x \in C[0,1] \mid w_x(\delta) \ge \varepsilon\} = 0,$$

for each $\varepsilon > 0$,

where w_x is defined as

$$w_x(\delta) = \sup_{|s-t|\le\delta} |x(s) - x(t)|, \quad 0 < \delta \le 1$$

(modulus of continuity of x).

Similar conditions exist in the space $D[0,1]$ of all right - continuous functions with left limits (càdlàg functions), equipped with Skorohod's metric (Billingsley 1999).

Finally, let us mention that in a Hilbert space H with an inner product $\langle \cdot, \cdot \rangle$, we may define characteristic function of a random variable X with a probability distribution μ, in the same way as in the finite dimensional spaces:

$$\varphi(x) = \int_H e^{i\langle x,y\rangle}\, \mathrm{d}\mu(y), \qquad x \in H.$$

Theorem 5 *Let $\{P_n\}$ be a sequence of probability measures on H and let φ_n be the corresponding characteristic functions. Let P and φ be a probability measure and its characteristic function. If $P_n \Rightarrow P$ then $\lim_n \varphi_n(x) = \varphi(x)$ for all $x \in H$.*

Conversely, if a sequence P_n of probability measures on H is relatively compact and $\lim_n \varphi_n(x) = \varphi(x)$ for all $x \in H$, then there exists a probability measure P such that φ is its characteristic function and $P_n \Rightarrow P$.

Cross References

►Almost Sure Convergence of Random Variables
►Bootstrap Asymptotics
►Central Limit Theorems
►Convergence of Random Variables
►Measure Theory in Probability
►Nonparametric Estimation Based on Incomplete Observations
►Role of Statistics in Advancing Quantitative Education

References and Further Reading

Billingsley P (1999) Convergence of probability measures. Wiley, New York
Lévy P (1937) Théorie de l'addition des variables aléatoires. Gauthier-Villars, Paris
Parthasarathy KR (1967) Probability measures on metric spaces. Academic Press, New York
Prohorov YuV (1956) Shodimost slučainih processov i predelynie teoremi teorii veroyatnostei. Teor ver primenen 1:177–238
Stroock D (1993) Probability theory, an analytic view. Cambridge University Press, Cambridge
Wichura MJ (1970) On the construction of almost uniformly convergent random variables with given weakly convergent image laws. Ann Stat 41:284–291

Weibull Distribution

William J. Padgett
Distinguished Professor Emeritus of Statistics
University of South Carolina, Columbia, SC, USA

The Weibull family of probability distributions (*see* also ►Generalized Weibull Distributions) is one the most widely used parametric families of distributions for

modeling failure times or lifetimes. This is especially true in engineering and science applications (as suggested originally by Weibull 1951) and is mainly due to the variety of shapes of its density function and the behaviors of its failure rate function. Literally thousands of references to the Weibull distribution can be found in the scientific literature. See Johnson et al. (1994) or a more recent treatment by Rinne (2008) for a detailed comprehensive overview of this family of distributions.

Let T denote a random variable (rv) representing the failure time or lifetime of an item under study. This rv has a Weibull distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$ if its probability density function (pdf) is $f(t) = \alpha t^{\alpha-1}\beta^{-\alpha}\exp[-(t/\beta)^{\alpha}]$ for $t \geq 0$. The cumulative distribution function (cdf) is then $F(t) = 1 - \exp[-(t/\beta)^{\alpha}]$, $t \geq 0$, and the survival or reliability function is $R(t) = 1 - F(t)$. Then the failure rate (or hazard rate) function is $h(t) = f(t)/R(t) = \alpha\beta^{-\alpha}t^{\alpha-1}$. For shape parameter $\alpha = 1$, the Weibull reduces to the *exponential* distribution with scale β, and when $\alpha = 2$ the resulting Weibull distribution is referred to as the *Rayleigh* distribution (*see* "Generalized Rayleigh distribution").

A major reason that the Weibull distributions are so useful is that the failure rate function can be increasing if the shape $\alpha > 1$, decreasing if $\alpha < 1$, or constant for $\alpha = 1$. An increasing failure rate function corresponds to the common assumption that the item whose lifetime is under study fails due to wearout over time, that is, an "ageing process" occurs where failure becomes more likely as time increases. The case of decreasing failure rate is less common but sometimes holds for types of items that tend to fail early due to defects or low quality and that tend to last longer if no defects are present, perhaps with very slow ageing. The constant failure rate corresponds to random failures occurring over time, which is the "memoryless" property of the exponential distribution. That is, there is no ageing process so that an item is always as good as new over time. Although the ageless property might seem to be unrealistic, some high-quality electronic items often approximately satisfy such an assumption for a period of time. So, Weibull distributions provide good models over a wide variety of "ageing" scenarios.

For integer $r > 0$ the r^{th} moment of a Weibull rv T is $E(T^r) = \beta^r\Gamma(1 + r/\alpha)$, where $\Gamma(c) = \int_0^{\infty} x^{c-1}e^{-x}dx$ is the gamma function. Therefore, the mean of the Weibull distribution is $\mu = E(T) = \beta\Gamma(1 + 1/\alpha)$ and the variance is $\sigma^2 = E(T^2) - \mu^2 = \beta^2\Gamma(1 + 2/\alpha) - \beta^2\Gamma^2(1 + 1/\alpha)$. These expressions are generally not very easy to use, but they can be obtained by computing approximate values of the gamma function. Since calculation of the mean lifetime is not very user-friendly, the value of the scale parameter itself is often used as a measure of "typical" lifetime, referred to as the *characteristic life* of the item. Since $R(\beta) = \exp(-1) \approx 0.37$, the characteristic life β is approximately the 63rd percentile of the distribution. Also, the variance is proportional to the square of the characteristic life.

The Weibull distribution arises also as the limiting distribution of the first order statistic from some probability distribution, so the Weibull is a limit of *extreme-value* distributions in this sense. That is, let $X_{(1)}$ denote the first order statistic from n independent identically distributed (iid) random variables, $X_1, \ldots, X_n$, from a specified cdf. Then as $n \to \infty$, the distribution of $X_{(1)}$ approaches a Weibull distribution (see Mann et al. 1974, or Rinne 2008). In fact, the Weibull distribution satisfies the important "weakest-link" property which is another reason for its applicability. Suppose that $X_1, \ldots, X_n$ are n iid random variables each with the Weibull cdf $F(t)$ and reliability function $R(t) = \exp[-(t/\beta)^{\alpha}]$, $t \geq 0$. Then the reliability function of the first order statistic, i.e. the "weakest" or smallest observation, $X_{(1)} = \min\{X_1, \ldots, X_n\}$, is $R_1(t) = \Pr[X_i > t \text{ for all } i = 1, \ldots, n] = [R(t)]^n = \{\exp[-(t/\beta)^{\alpha}]\}^n = \exp[-(t/\beta n^{-1/\alpha})^{\alpha}]$. Thus, $X_{(1)}$ also has the Weibull distribution with the same shape parameter α and new scale parameter $\beta n^{-1/\alpha}$. This implies that in the increasing failure rate case, $\alpha > 1$, a long chain of "links" has a higher probability of failure than a shorter chain. This idea is important in modeling the failure of materials (see, for example, Smith 1991, and Wolstenholme 1995), as well as the failure of a series system of n iid components.

Several generalizations of the Weibull distribution have been proposed, three of which will be mentioned here. A frequently used version is obtained by adding a "shift parameter," γ, also referred to as a "guarantee time." That is, the Weibull pdf and cdf are shifted from zero to γ, so $f(t) = \alpha(t-\gamma)^{\alpha-1}\beta^{-\alpha}\exp\left[-\left(\frac{t-\gamma}{\beta}\right)^{\alpha}\right]$ and $F(t) = 1 - \exp[-((t-\gamma)/\beta)^{\alpha}]$, $t \geq \gamma$, to obtain the *three-parameter Weibull* distribution. Another generalization introduced by Mudholkar and Srivastava (1993) is known as the *exponentiated Weibull* distribution which has cdf $F_{EW}(t) = \{1 - \exp[-(t/\beta)^{\alpha}]\}^{\theta}$, $t \geq 0$, with another parameter $\theta > 0$. For $\alpha = 2$ and $\beta = 1$, this exponentiated Weibull distribution reduces to the *Burr type X* distribution (see Burr 1942). The third generalization mentioned here is called the *brittle fracture* distribution (see Black et al. 1990), whose reliability function is of the form $R_{BF}(t) = \exp[-\delta t^{2\rho}\exp(-\theta/t^2)]$, $t > 0$. This distribution was found to provide good model fits specifically to observed breaking stress data for boron fibers and carbon fibers. Taking $\theta = 0$ and $2\rho = a$ yields the usual two-parameter Weibull distribution with shape parameter α.

Estimation of the parameters for fitting the Weibull distribution to observed failure data is typically accomplished by the method of maximum likelihood. This involves numerical solution since the likelihood equations yield nonlinear functions of the shape parameter. For example, see Mann et al. (1974) and Rinne (2008). *Weibull plotting* is a graphical technique that is often used for quick (not maximum likelihood) estimation of the parameters (for examples, refer to Rinne (2008) and Wolstenholme (1999)). Tests of hypotheses for the parameters, interval estimation, and other inferences for the Weibull model are discussed by Mann et al. (1974) and Rinne (2008) as well as by many other authors. There are several available statistical or engineering software packages that include Weibull modeling procedures. Among others, two dedicated software packages for Weibull analysis of lifetime data may be found at http://Weibull.ReliaSoft.com and http://www.relex.com/products/weibull.asp.

About the Author

Professor Padgett was Department Chairman (1985–1993) and (1996–2001). He was awarded the Donald S. Russell Award for Creative Research in Science and Engineering, University of South Carolina (1975) and Paul Minton Service Award, Southern Regional Council on Statistics (2003), among others. He is a Fellow of the American Statistical Association and of the Institute of Mathematical Statistics, and an Elected member of the International Statistical Institute. Professor Padgett was an Associate editor for many international journals including, *Technometrics* (1987–1992), *Journal of Nonparametric Statistics* (1989–2004), *Journal of Statistical Theory and Applications* (2001–2007), *Lifetime Data Analysis* (1994–2003), *Journal of Applied Statistical Science* (1992–2001) and *Journal of Statistical Computation and Simulation* (1980–1986). He was also Coordinating Editor of *Journal of Statistical Planning and Inference* (1995–1997).

Cross References

►Extreme Value Distributions
►Generalized Extreme Value Family of Probability Distributions
►Generalized Rayleigh Distribution
►Generalized Weibull Distributions
►Modeling Survival Data
►Multivariate Statistical Distributions
►Statistical Distributions: An Overview
►Statistics of Extremes
►Survival Data

References and Further Reading

Black CM, Durham SD, Padgett WJ (1990) Parameter estimation for a new distribution for the strength of brittle fibers: a simulation. Comm Stat Simulat Comput 19:809–825
Burr IW (1942) Cumulative frequency functions. Ann Math Stat 13:215–222
Johnson NL, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, vol 1, 2nd edn. Wiley Series in Probability and Mathematical Statistics, New York
Mann NR, Shafer RE, Singpurwalla ND (1974) Methods for statistical analysis of reliability and lifetime data. Wiley, New York
Mudholkar GS, Srivastava DK (1993) Exponentiated Weibull family for analyzing bathtub failure-rate data. IEEE Trans Reliab 42:299–302
Rinne H (2008) The Weibull distribution: a handbook. CRC Press, Boca Raton
Smith RJ (1991) Weibull regression models for reliability data. Reliab Eng Syst Saf 34:35–57
Weibull W (1951) A statistical distribution function of wide applicability. J Appl Mech 18:293–297
Wolstenholme LC (1995) A non-parametric test of the weakest-link property. Technometrics 37:169–175
Wolstenholme LC (1999) Reliability modelling: a statistical approach. Chapman & Hall/CRC, Boca Raton

Weighted Correlation

Joaquim F. Pinto da Costa
Professor
University of Porto, Porto, Portugal

Introduction

Weighted correlation is concerned with the use of weights assigned to the subjects in the calculation of a correlation coefficient (see ►Correlation Coefficient) between two variables X and Y. The weights can either be naturally available beforehand or chosen by the user to serve a specific purpose. For instance, if there is a different number of measurements on each subject, it is natural to use these numbers as weights and calculate the correlation between the subject means. On the other hand, if the variables X and Y represent, for instance, the ranks of preferences of two human beings over a set of n items, one might want to give larger weights to the first preferences, as these are more accurate. In another situation, if we want to calculate the correlation between two stocks in a stock exchange market during last year, we might want to favor (larger weight) the more recent observations, as these are more important for the present situation. Suppose that X_i and Y_i are the pair of values corresponding to observation i in each sample and w_i the weight attributed to this observation, such

that $\sum_{i=1}^{n} w_i = 1$. Then, the sample weighted correlation coefficient is given by the formula

$$r_w = \frac{\sum w_i(X_i - \overline{X}_w)(Y_i - \overline{Y}_w)}{\sqrt{\sum w_i(X_i - \overline{X}_w)^2}\sqrt{\sum w_i(Y_i - \overline{Y}_w)^2}} = \frac{\sum w_i X_i Y_i - \sum w_i X_i \sum w_i Y_i}{\sqrt{\sum w_i X_i^2 - (\sum w_i X_i)^2}\sqrt{\sum w_i Y_i^2 - (\sum w_i Y_i)^2}}, \quad (1)$$

where the sums are from $i = 1$ to n and $\overline{X}_w = \sum w_i X_i$ and $\overline{Y}_w = \sum w_i Y_i$ are the weighted means. When all the w_i are equal they cancel out, giving the usual formula for the Pearson product–moment correlation coefficient.

Weighted Rank Correlation

Rank correlation coefficients are nonparametric statistics that are less restrictive than others (e.g., Pearson's correlation coefficient), because they do not try to fit one particular kind of relationship, linear or other, to the data. Their objective is to assess the degree of monotonicity between two series of paired data. Common rank correlation coefficients are Spearman's and Kendall's (Neave and Worthington 1992). One interesting fact about rank correlation is that, contrary to other correlation methods, it can be used not only on numerical data but on any data that can be ranked.

Blest (2000) proposed an alternative weighted measure of rank correlation that gives more importance to the first ranks but has some drawbacks because it is not a symmetric function of the two vectors of ranks. Later, Pinto da Costa and Soares (Pinto da Costa and Soares 2005; Soares et al. 2001) presented a new weighted rank correlation coefficient that gives larger weight to the first ranks and does not have the problems of Blest's coefficient. This coefficient is

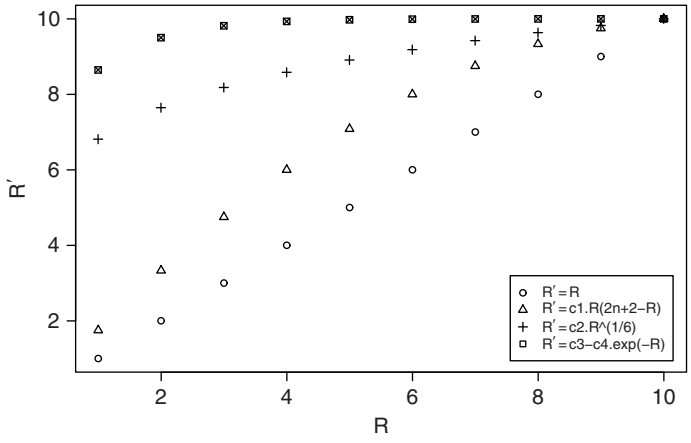
$$r_W = 1 - \frac{6\sum_{i=1}^{n}(R_i - Q_i)^2(2n + 2 - R_i - Q_i)}{n^4 + n^3 - n^2 - n}, \quad (2)$$

where R_i is the rank corresponding to the ith observation of the first variable, X, and Q_i is the rank corresponding to the ith observation of the second variable, Y. r_W, which yields values between −1 and +1, uses a linear weight function: $2n + 2 - R_i - Q_i$. Some properties of the distribution of the statistic r_W, including its sample distribution, are analyzed in Pinto da Costa and Soares (2005) and Pinto da Costa and Roque (2006); in particular, the expected value of this statistic is zero when the two variables are independent, and its sampling distribution converges to the Gaussian when the sample size increases. Later, Pinto da Costa and Soares (2007) introduced a new weighted rank correlation coefficient that uses a quadratic weight function:

$$r_{W2} = 1 - \frac{90\sum_{i=1}^{n}(R_i - Q_i)^2(2n + 2 - R_i - Q_i)^2}{n(n-1)(n+1)(2n+1)(8n+11)}. \quad (3)$$

A New Way of Developing Weighted Correlation Coefficients

It can be proved that the coefficient r_{W2} is equal to the Pearson's correlation coefficient of the transformed ranks $R'_i = R_i(2n + 2 - R_i)$ and $Q'_i = Q_i(2n + 2 - Q_i)$ and this suggests a new and easy way of developing weighted correlation coefficients. In fact, by applying a transformation to the ranks so that the first ones are favored and then computing the Pearson's correlation coefficient of the transformed ranks, we can define many new measures of weighted correlation (Pinto da Costa and Soares 2007). In Fig. 1 we can see four different cases. The first, when $R' = R$, corresponds to Spearman's coefficient and so it



Weighted Correlation. Fig. 1 Scatterplot for four different rank transformations

does not correspond to a weighted measure; when $R' = R(2n + 2 - R)$ we have r_{W2}. We can now use other functions such as $R' = R^{1/6}$ or $R' = -e^{-R}$. In order to be able to represent the four cases in the same diagram some of the transformations had to be multiplied by a constant and in the last case another constant was also added, but these operations do not change the value of Pearson's correlation. Thus, the importance given to the first ranks is larger when $R' = -e^{-R}$ and smaller when $R' = R$. This means that the ranks that are in a flatter region are given smaller weight.

From this perspective, and in case we want to give larger weight to the first ranks, all that is needed is that the transformation is monotone and the last ranks are more flattened by the transformation compared with the first ranks. However, if we want to give larger weight to other ranks, not the first, we just have to find an appropriate transformation to do that; one that is less flat where the weights are to be larger. This in turn has two additional advantages. First, we can use different transformations to each variable and so we are not obliged to give the same set of weights to the two variables. Secondly, this strategy can be used with the original data, not only ranks, and so many new measures of weighted correlation can be developed.

About the Author

Joaquim Costa, University of Porto, is an elected member of the International Statistical Institute. He is also a member of the Portuguese Statistical Society and of the Portuguese Classification Society. He completed his PhD in 1996 under the supervision of Israel Lerman from University of Rennes 1. He has published in statistical and machine learning journals and his main contributions are in weighted measures of correlation, weighted principal component analyses, and also new methods of supervised classification for ordered classes.

Cross References

- ►Correlation Coefficient
- ►Kendall's Tau
- ►Measures of Dependence
- ►Rank Transformations

References and Further Reading

Blest D (2000) Rank correlation – an alternative measure. Aust N Z J Stat 42(1):101–111

Neave H, Worthington P (1992) Distribution-free tests. Routledge, London

Pinto da Costa J, Roque L (2006) Limit distribution for the weighted rank correlation coefficient, rW. REVSTAT – Stat J 4(3):189–200

Pinto da Costa J, Soares C (2005) A weighted rank measure of correlation. Aust N Z J Stat 47(4):515–529

Pinto da Costa J, Soares C (2007) Rejoinder to letter to the editor from C. Genest and J-F. Plante concerning Pinto da Costa J & Soares C (2005) A weighted rank measure of correlation. Aust N Z J Stat 49(2):205–207

Soares C, Pinto da Costa J, Brazdil P (2001) Improved statistical support for matchmaking: rank correlation taking rank importance into account. In: JOCLAD 2001: VII Jornadas de Classificação e Análise de Dados, Porto, Portugal, pp 72–75

Weighted *U*-Statistics

Aluísio de Souza Pinheiro
Professor, Head of Department of Statistics
University of Campinas, Campinas, SP, Brazil

Following the seminal papers of Hoeffding (1948, 1961), let T_n be a linear combination defined by

$$T_n = \sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\cdots i_m}\phi(\mathbf{X}_{i_1},\ldots,\mathbf{X}_{i_m}), \qquad (1)$$

such that: (a) $\eta_{n,i_1\cdots i_m}$ are weight functions, (b) $\sum_{i_1,\ldots,i_m}^{1,n}$ is taken on all strictly ordered permutations of $1,\ldots,n$, (c) $\phi(\cdot,\ldots,\cdot)$ is a kernel of degree m, stationary of order r $(1 \le r \le m)$, for which we let $\theta = \mathrm{E}\phi(X_1,\ldots,X_m)$, and (d) $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are i.i.d. random vectors of dimension K, not necessarily quantitative in nature (Pinheiro et al. 2009).

Some configurations of $\eta_{n,i_1\cdots i_m}$ lead to special classes of (generalized) ►*U*-statistics, as follows: If $\eta_{n,i_1\cdots i_m} \equiv \binom{n}{m}^{-1}$ and $r \ge 1$, T_n is a degenerate U-statistics of degree m whose projection variances are such that $0 = \sigma_1^2 = \cdots = \sigma_r^2 < \sigma_{r+1}^2$; then T_n has a degeneracy of order r and $n^{(r+1)/2}(T_n - \theta)$ converges to a (possibly) infinite linear combination of independent random variables, each distributed accordingly to a $(r+1)$-dimensional Wiener integral (Dynkin and Mandelbaum 1983).

If $K = 1$ and the $\eta_{n,i_1\cdots i_m}$ assume 0 or 1 values only, T_n is said to be an *incomplete* U-statistic (Janson 1984). Asymptotic distribution of T_n will be either a linear combination of independent Wiener integrals or a mixture of such a distribution with an independent normal r.v., under suitable sampling conditions (Janson 1984). For a class of *conditional* U-statistics, where the weights can be decomposed as $\eta_{n,i_1\cdots i_m} = e(i_1)\cdots e(i_m)$, $e(\cdot)$ being the marginal weight function, ►asymptotic normality follows from Stute (1991). Moreover, the conditional nature of the class derives from the fact that weights are defined as random functions of another set of r.v.'s.

For $K = 1$, O'Neil and Redner (1993) and Major (1994) present asymptotic results in a more general setup for

the class of *weighted* U-statistics, defined by (1). The case $m = 2$ using moment matching techniques to determine the asymptotic distribution of T_n is discussed in O'Neil and Redner (1993). Under some regularity conditions on $\eta_{n,i_1\cdots i_m}$, a non-normal limit is proven for either $r = 1$ or $r = 0$. For $r = 0$, a class of weighted U-statistics is proved to be asymptotically normal under a second set of conditions on weights. ►Asymptotic normality is also established for $r = 1$ and *incomplete designs*. The common idea behind all weight-designs is the orthogonality on the set of (possibly random) weights. Major (1994) points out that the aforementioned approach cannot be adapted for $m \geq 3$; Poisson approximation is then used to pursue asymptotic behavior of T_n.

A class of *quasi U-statistics* having the novelty that it can be applied to any i.i.d. random vectors of arbitrary (and even increasing) dimension K, is introduced in Pinheiro et al. (2009). The proposed class is constructed in such a way that, although ϕ can be degenerate, the chosen weights lead to a contrast, i.e., such that $\sum_{i_1,\ldots,i_m}^{1,n} \eta_{n,i_1\cdots i_m} = 0$, providing asymptotically normal distributions. For the quasi U-statistics, the aforementioned contrast condition is an essential requirement. Otherwise, for degenerate U-statistics the asymptotic distribution is non-normal.

About the Author

Dr. Aluisio Pinheiro has been Chair of the Department of Statistics at the University of Campinas since 2007. He has coauthored several papers on U statistics (with P. K. Sen and H. P. Pinheiro), and a book (with H. P. Pinheiro) on the use of Quasi-statistics in Genetic Data published by the Brazilian Society of Mathematics in 2007.

Cross References

►Asymptotic Normality
►U-Statistics

References and Further Reading

Dynkin EB, Mandelbaum A (1983) Symmetric statistics, Poisson point processes and multiple Wiener integrals. Ann Stat 11(3):739–745
Hoeffding W (1948) A class of statistics with asymptotically normal distribution. Ann Math Stat 19(3):293–325
Hoeffding W (1961) The strong law of large numbers for U-statistics. Institute of Statistics Mimeo Series No. 302. University of North Carolina, Chapel Hill
Janson S (1984) The asymptotic distribution of incomplete U-statistics. Z Wahrsch Verw Geb 66(4):495–505
Major P (1994) Asymptotic distributions for weighted U-statistics. Ann Prob 22(3):1514–1535
O'Neil KA, Redner RA (1993) Asymptotic distributions of weighted U-statistics of degree 2. Ann Probab 21(2):1159–1169
Pinheiro A, Sen PK, Pinheiro HP (2009) Decomposability of high-dimensional diversity measures: quasi U-statistics, martingales and nonstandard asymptotics. J Multivar Anal 100:1645–1656
Stute W (1991) Conditional U-statistics. Ann Probab 19(2):812–823

Wilcoxon–Mann–Whitney Test

Markus Neuhäuser
Professor
Koblenz University of Applied Sciences, Remagen, Germany

The Wilcoxon–Mann–Whitney (WMW) test was proposed by Frank Wilcoxon in 1945 ("Wilcoxon rank sum test") and by Henry Mann and Donald Whitney in 1947 ("Mann–Whitney U test"). However, the test is older: Gustav Deuchler introduced it in 1914 (see Kruskal 1957). Nowadays, this test is a commonly used nonparametric test for the two-sample location problem. As with many other nonparametric tests, this is based on ranks rather than on the original observations.

The sample sizes of the two groups or random samples are denoted by n and m. The observations within each sample are independent and identically distributed, and we assume independence between the two samples. The null hypothesis, H_0, is one of no difference between the two groups.

Let F and G be the distribution functions corresponding to the two samples. Then we have the null hypothesis $H_0 : F(t) = G(t)$ for every t. Under the two-sided alternative there is a difference between F and G. Often, it is assumed that F and G are identical except a possible shift in location (location-shift model), i.e., $F(t) = G(t - \theta)$ for every t. Then, the null hypothesis states $\theta = 0$, and the two-sided alternative is $H_1 : \theta \neq 0$. Of course, one-sided alternatives are possible, too.

Let $V_i = 1$ when the ith smallest of the $N = n + m$ observations is from the first sample and $V_i = 0$ otherwise. The Wilcoxon rank sum is a linear rank statistic defined by $W = \sum_{i=1}^{N} i \cdot V_i$. Hence, W is the sum of the n ranks of group 1; the ranks are determined based on the pooled sample of all N values.

The Mann–Whitney statistic U is defined as $U = \sum_{i=1}^{n}\sum_{j=1}^{m}\psi(X_i, Y_j)$ where X_i (Y_j) is an observation from

group 1 (group 2), and

$$\psi(X_i, Y_j) = \begin{cases} 1 & \text{if} \quad X_i > Y_j \\ 0.5 & \text{if} \quad X_i = Y_j \\ 0 & \text{if} \quad X_i < Y_j. \end{cases}$$

Because of $W = U + \frac{n}{2}(n+1)$, the tests based on W and U are equivalent.

The standardized statistic Z_W can be computed as $Z_W = \frac{W - E_0(W)}{\sqrt{Var_0(W)}}$ with $E_0(W) = \frac{n(N+1)}{2}$ and $Var_0(W) = \frac{nm(N+1)}{12}$. In the presence of ties mean ranks can be recommended for tied observations. Then, the variance changes, in this case we have

$$Var_0(W) = \frac{nm}{12}\left(N + 1 - \frac{\sum_{i=1}^{g}(t_i - 1)t_i(t_i + 1)}{N(N-1)}\right),$$

where g is the number of tied groups and t_i the number of observations within the ith tied group. An untied value is regarded as a tied group with $t_i = 1$ (Hollander and Wolfe 1999, p. 109).

Under H_0, the standardized Wilcoxon statistic asymptotically follows a standard normal distribution. This result can be used to carry out the test and to calculate an asymptotic p-value. According to Brunner and Munzel (2002, p. 63) the normal approximation is acceptable in case of $\min(n, m) \geq 7$, if there were no ties. The two-sided asymptotic WMW test can reject H_0 if $|Z_W| \geq z_{1-\alpha/2}$, the corresponding p-value can be computed as $2(1 - \Phi(|Z_W|))$, where $z_{1-\alpha/2}$ and Φ denote the $(1 - \alpha/2)$-quantile and the distribution function, respectively, of the standard normal distribution.

Alternatively, the exact permutation null distribution of W can be determined and used for inference (see the chapter about ►permutation tests). Some monographs include tables of critical values for the permutation test, but these tables can only be used if there were no ties. A permutation test, however, is also possible in the presence of ties, because the exact conditional distribution of W can be obtained.

As a rank test the WMW test does not use all the available information; despite this, it is quite powerful. If the normal distribution is a reasonable assumption, little is lost by using the Wilcoxon test instead of the parametric t test. On the other hand, when the assumption of normality is not satisfied, the nonparametric Wilcoxon test may have considerable advantages in terms of efficiency. To be precise, the asymptotic relative efficiency (ARE) of the WMW test in comparison to Student's t test (see ►Student's t-Tests) cannot be smaller than 0.864. However, there is no upper limit. If the data follow a normal distribution the ARE is $3/\pi = 0.955$ (Hodges and Lehmann 1956).

The two-sided WMW test is consistent against all alternatives with $\Pr(X_i < Y_j) \neq 0.5$. However, the WMW test can give a significant result for a test at the 5% level with much more than 5% probability when the population medians are identical, but the population variances differ. A generalization exists that can be applied for testing a difference in location irrespective of a possible difference in variability (Brunner and Munzel 2000).

About the Author

Dr Markus Neuhäuser is a Professor, Department of Mathematics and Technique, Koblenz University of Applied Sciences, Remagen, Germany. He was Senior Lecturer at the Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand (2002–2004). He has authored and co-authored more than 100 papers and 2 books, including *Computer-intensive und nichtparametrische statistische Tests* (Oldenbourg, 2010). Currently, he is an Associate Editor for the *Journal of Statistical Computation and Simulation*, *Communications in Statistics - Theory and Methods*, and *Communications in Statistics – Simulation and Computation*.

Cross References

►Asymptotic Relative Efficiency in Testing
►Continuity Correction
►Explaining Paradoxes in Nonparametric Statistics
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Presentation of Statistical Testimony
►Rank Transformations
►Ranks
►Scales of Measurement and Choice of Statistical Methods
►Sequential Ranks
►Statistical Fallacies: Misconceptions, and Myths
►Student's t-Tests
►Wilcoxon-Signed-Rank Test

References and Further Reading

Brunner E, Munzel U (2000) The nonparametric Behrens-Fisher problem: asymptotic theory and a small sample approximation. Biom J 42:17–25
Brunner E, Munzel U (2002) Nichtparametrische Datenanalyse. Springer, Berlin
Hodges JL, Lehmann EL (1956). The efficiency of some nonparametric competitors of the t-test. Ann Math Stat 27:324–335

Hollander M, Wolfe DA (1999) Nonparametric statistical methods, 2nd edn. Wiley, New York
Kruskal WH (1957) Historical notes on the Wilcoxon unpaired two-sample test. J Am Stat Assoc 52:356–360

Wilcoxon-Signed-Rank Test

Denise Rey[1], Markus Neuhäuser[2]
[1]Director
Rey Analytical Research, Hürth, Germany
[2]Professor
Koblenz University of Applied Sciences, Remagen, Germany

The Wilcoxon-signed-rank test was proposed together with the Wilcoxon-rank-sum test (see ►Wilcoxon–Mann–Whitney Test) in the same paper by Frank Wilcoxon in 1945 (Wilcoxon 1945) and is a nonparametric test for the one-sample location problem. The test is usually applied to the comparison of locations of two dependent samples. Other applications are also possible, e.g., to test the hypothesis that the median of a symmetrical distribution equals a given constant. As with many nonparametric tests, the distribution-free test is based on ranks.

To introduce the classical Wilcoxon-signed-rank test and also important further developments of it we denote by $D_i = Y_i - X_i, i = 1, \ldots, N$ the difference between two paired random variables. The classical Wilcoxon-signed-rank test assumes that the differences D_i are mutually independent and D_i, $i = 1, \ldots, N$ comes from a continuous distribution F that is symmetric about a median θ. The continuity assumption on the distribution of the differences implies that differences which are equal in absolute value may not occur, i.e., the classical Wilcoxon-signed-rank test assumes no ties in the differences $|d_i| \neq |d_j|$ for $i \neq j$ and $1 \leq i, j \leq N$. Moreover, it is assumed that the sample is free of zero differences, i.e., $d_i \neq 0, \forall i = 1, \ldots, N$. We further denote by N_0 and M the number of zero and the number of non-zero differences in the sample, respectively. It follows $N = N_0 + M$ with $N_0 = 0$ for the classical Wilcoxon-signed-rank test.

The null hypothesis states that $H_0 : \theta = 0$, i.e., the distibution of the differences is symmetric about zero corresponding to no difference in location between the two samples. The two-sided alternative is $H_1 : \theta \neq 0$. One-sided alternatives are also possible.

The Wilcoxon-signed-rank test statistic is the linear rank statistic $R_+ = \sum_{i=1}^{N} R_i V_i$ where $V_i = 1_{D_i > 0}$, is the indicator for the sign of the difference and R_i is the rank of $|D_i|$, $i = 1, \ldots, N$. Therefore, the test statistic represents the sum of the positive signed ranks. (The test statistic could also be build in terms of the sum of negative signed ranks, R_- or the difference of both $R = R_+ - R_-$. The three statistics are equivalent. For theoretical investigations is R often more suitable. Nevertheless, in literature R_+ and R_- are widespread.) The critical values w_α for the exact distribution of R_+ are tabulated. Reject the null hypothesis at the α level of significance if $R_+ \geq w_{\alpha/2}$ or $R_+ \leq \frac{N(N+1)}{2} - w_{\alpha/2}$.

Nowadays, the exact distribution can be determined by generating all 2^N sign permutations of the ranked differences. For each permutation, the value of the test statistic has to be calculated. The proportion of permutations that give a value as or more extreme than observed, is the p-value of the resulting exact test. Hence, in terms of p-values and due to the symmetry of the distribution, we reject the null hypothesis if the p-value $p = 2P(R_+ \geq r_+) \leq \alpha$ where r_+ is the observed value of the test statistic.

A large-sample approximation uses the asymptotic normal distribution of R_+. Under the null hypothesis we have

$$E_0(R_+) = \frac{N(N+1)}{4}, \quad Var_0(R_+) = \frac{N(N+1)(2N+1)}{4}$$

and the standardized version of R_+ is asymptotically

$$R_+^* = \frac{R_+ - E_0(R_+)}{\sqrt{Var_0(R_+)}} \overset{H_0}{\sim} N(0,1).$$

Reject the null hypothesis if $|R_+^*| \geq z_{1-\alpha/2}$.

In applications, the assumptions of the classical Wilcoxon-signed-rank test of non-zero differences and no ties in the sample are often not fulfilled.

We still assume that zero values are not possible but allow ties among the non-zero differences (the continuity assumption on the distribution of the differences is relaxed). Then one can apply the classical Wilcoxon-signed-rank test on the mean ranks that are associated with the tied group. In the case of ties among the non-zero differences, a conditional test based on the exact conditional distribution of the Wilcoxon signed-rank statistic given the set of tied ranks and by means of mean ranks is possible (Hollander and Wolfe 1999 p. 46).

For the large-sample approximation in the case of non-zero differences but existing tied observations among the non-zero differences, the variance of the test statistic changes to

$$Var_0(R_+) = \frac{1}{24}\left(N(N+1)(2N+1) - \frac{1}{2}\sum_{i=1}^{C} T_i(T_i-1)(T_i+1)\right) \quad (1)$$

where we have denoted by C the number of groups with ties and by $T_i \geq 1$, $i = 1, \ldots, C$ the number of observations within tie group i. It holds for this case $N = M = \sum_{i=1}^{C} T_i$. An untied observation is then considered to be a group of size 1 (Hollander and Wolfe 1999 p. 38). We remark that the classical Wilcoxon-signed-rank test assumes $T_i = 1$, $\forall i = 1, \ldots, C$. The test statistic R_+^* adapted by equation (1) is then computed with respect to mean ranks. Under the null hypothesis it is asymptotically normal distributed and corresponding tests can be applied.

In applications zero differences do often exist. Wilcoxon suggested dropping the zeros from the initial data and go on with the test on the reduced data.

Another method for handling zero differences was given by Pratt (Pratt 1959). Pratt suggested to rank all observations, including the zeros, from smallest to largest in absolute magnitude and afterwards drop the ranks of the zeros without changing the ranks of the non-zero values and proceed with the testing. In this case we have $N_0 > 0$ and ranks start by $N_0 + 1, \ldots, N$. Pratt motivated this procedure by showing that contradictory test decisions could occur when zero differences are ignored. More exactly, he showed that dropping the zeros before ranking fails to satisfy a monotonicity requirement: The probability under a test based on the signed rank statistic and randomized to have exact α level of calling a sample significantly positive should be a nondecreasing function of the observations (Pratt 1959, p. 659). Tables of critical values for a conditional exact Pratt test where a certain number of zero differences are allowed and mean ranks for ties are involved are computed by Buck (Buck 1979). Analogously, running through all 2^N sign permutations allows the computation of the exact distribution of the test statistic independent of tabulated values.

Asymptotically, the standardized test statistic where expectation and variance is properly adapted for the modification of Pratt is under the null hypothesis normal distributed and corresponding tests can be applied (Buck 1979).

Conover (Conover 1973, p. 985) showed that there are cases (e.g., the uniform distribution) for which the Wilcoxon test with the Pratt modification for handling zero differences and mean ranks for the non-zero differences has a greater asymptotic efficiency than the classical Wilcoxon test. Moreover he showed that there are also cases (e.g., the ►binomial distribution) for which the Wilcoxon test with non-zero differences and mean ranks for the non-zero differences gives better asymptotic efficiency than the Pratt method.

Another well-known test for the one-sample location problem is the ►sign test. Compared to the sign test, the Wilcoxon-signed-rank test has the additional assumption of the symmetry of the distribution but uses the ordering of the differences as additional information. In literature we find that there are advantageous cases with respect to the asymptotic efficiency for both the sign test and the Wilcoxon-signed-rank test (Higgins 2004).

About the Author

For biography *see* the entry ►Wilcoxon-Mann-Whitney test.

Cross References

►Asymptotic Normality
►Asymptotic Relative Efficiency in Testing
►Nonparametric Rank Tests
►Nonparametric Statistical Inference
►Sign Test
►Student's *t*-Tests
►Wilcoxon–Mann–Whitney Test

References and Further Reading

Buck W (1979) Signed-rank tests in presence of ties (with extended tables). Biom J 21(6):501–526
Conover WJ (1973) On methods of handling ties in the Wilcoxon signed-rank test. J Am Stat Assoc 68(344):985–988
Higgins JJ (2004) An introduction to modern nonparametric statistics. Brooks/Cole, Pacific Grove
Hollander M, Wolfe DA (1999) Nonparametric statistical methods. 2nd edn. Wiley, New York
Pratt JW (1959) Remarks on zeros and ties in the Wilcoxon signed rank procedures. J Am Stat Assoc 54:655–667
Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics 1:80–83

List of Entries