Journal of Official Statistics
Vol. 1, No. 4, 1985, pp. 381–397
© Statistics Sweden

Variance Estimation for Complex Estimators in Sample Surveys

Keith Rust[1]

Abstract: The complex estimators frequently used in sample surveys lead to special difficulties in the estimation of sampling variance. This paper reviews current practical methods of deriving such variance estimates. Both theoretical and empirical results for these methods and comparisons between the methods are considered.

Key words: Linear estimator; Taylor series linearization; replication; random groups; balanced repeated replication (BRR); combined strata; jackknife; bootstrap.

1. Introduction

In producing estimates from sample surveys, it is a generally recommended practice that due concern be given to quantifying the extent of sampling error. Textbooks on survey sampling, such as those by Cochran (1977), Hansen, Hurwitz, and Madow (1953), and Kish (1965), provide explicit formulae for unbiased estimators of the sampling variance of simple (linear) estimators of, say, totals and means from a population of known size. This is provided for a variety of the more straightforward sample designs. However, such formulae are usually inadequate in survey practice because non-linear estimators are often used for estimating parameters of interest.

An estimator $\hat{\mu}$ of a parameter μ is linear if it can be expressed in the form

$$\hat{\mu} = \sum_{i \epsilon s} \beta_i^T X_i$$

where S denotes the sample and i the selected units, and $\underline{\beta}_i^T$ denotes the transpose of the vector $\underline{\beta}_i$. The components of $\underline{X}_i$ are the values of some characteristic (or characteristics) of the selected units, and the components of $\underline{\beta}_i$ are constants which do not depend on the sample *values*, but only upon the label of unit i. For multistage samples, $\underline{X}_i$ is an estimate of characteristics of the primary sampling unit (PSU) i, derived from the units sampled at the second and later stages. In survey practice, such linear estimators consist of linear combinations (with known coefficients) of estimators of stratum totals. Commonly encountered estimators of parameters such as ratios, regression coefficients, measures of association, and population variances are

[1] Statistical Services Branch, Australian Bureau of Statistics. The author wishes to thank Graham Kalton and Dennis Trewin for their advice and criticism in the preparation of this paper, and two referees for their comments on an earlier draft.

non-linear. Furthermore, non-linear estimators are used frequently for estimating population totals and means, as they may be more precise than the appropriate linear estimators (a ratio estimator of a total provides an example).

In traditional statistical theory, models are used and statistical inference relates to the parameters of these models. The use of such models frequently gives rise to explicit variance estimators for non-linear estimators, and permits one to ignore the mechanism by which sample observations are generated. When inferences about the parameters of a finite population are based on sample survey data without model assumptions other methods must be used to derive estimates of the variances of the parameter estimates.

Two approaches to the task of estimating the variance of a non-linear estimator can be distinguished. One is linearization, in which the non-linear estimator is approximated by a linear one for the purposes of variance estimation. The second is replication, in which several estimates of the population parameter under study are derived from different, comparable parts of the original sample. The variability of these estimates is used to estimate the variance of the parameter estimator. This paper reviews current knowledge about the properties and practice of these procedures.

2. The Taylor Series Linearization Procedure

The basis of the Taylor series linearization procedure is the approximation of the estimator of interest by a linear estimator. The variance of this linear approximation is then estimated using standard variance estimation methods available for linear estimators. The linear approximation is derived by taking a first order Taylor series approximation for the estimator $\hat{\mu}$ about the parameter μ. Some of the components of this linear approximation will involve unknown population parameters, and must be estimated from sample data. As $\hat{\mu}$ is in general biased for μ, the resulting variance estimator actually estimates the mean square error of $\hat{\mu}$.

As with linear estimators, for multistage samples it is necessary only to use estimates of PSU totals in variance estimation, provided that the assumption of with replacement sampling at the first stage is tenable. Woodruff (1971) was the first to make this clear.

Traditionally, the technique has been used for ratio estimation, and it is described specifically for this purpose in many sampling textbooks where it is often referred to as the delta method. Tepping (1968) gives a detailed description of the use of the method for more complex estimators. Shah, Holt, and Folsom (1977) evaluated the use of the technique for estimating the covariance matrices of sets of regression coefficients estimated from complex sample designs. They found that the linearization approach gave rise to generally acceptable inferences, far superior to inferences based on the assumption of simple random sampling. Other studies of the linearization method, in which its performance is compared with that of replication, are discussed in Section 7.

When using the linearization method, one encounters two major difficulties. The procedure assumes that in the Taylor series expansion of the estimator, terms beyond the linear one make a negligible contribution to the variance of the estimator. This may not be the case, particularly for small sample sizes, and the approximation can fail badly in practice.

The second difficulty lies in the derivation of the linear substitute appropriate for a given estimator. This is straightforward for estimators such as a ratio, $r = x/y$, used to estimate a population parameter $R = X/Y$. However, for more complex estimators, the analytic partial differentiation needed to derive the linear substitute has generally been found to be intractable. Woodruff and Causey (1976) describe a solution to this problem that uses a numerical procedure to obtain the necessary partial derivatives. A computer program, GENVAR, has been developed for implementing the procedure. The program requires that the user supply a subroutine for calculating the within stratum variance, appropriate to the design, of a linear estimator. The program is not actively maintained and is not being developed further (Francis (1981, Ch. 5)).

Recent work by Binder (1983) provides a general approach to the analytic derivation of variance estimators for first order (linear) Taylor series approximations for a wide class of estimators. In particular, variance estimators can be derived for estimators of implicitly defined parameters estimated by iterative procedures. Binder gives examples of the application of the procedure to variance estimation for estimators of ordinary least squares regression coefficients, logistic regression coefficients, and coefficients for log-linear models for categorical data. The method has wide applicability, and it appears that a solution is available for many of the estimators likely to be used in survey practice. However, a separate variance estimator must be derived for each estimator used, and correspondingly a specific computer subroutine is required in each case. The resulting variance estimation formulae can easily become complex. An explicit variance (and covariance) estimator that is appropriate for linear estimators is required in all cases.

Consequently, it is not always clear to what extent this approach provides a more easily implemented and less costly alternative to the replicated procedures described later. In these replicated procedures, the variance estimator does not explicitly vary with the estimator being considered.

Empirical evidence has shown that the variance estimates obtained from the linearization procedure are generally of adequate accuracy (see Section 7). It is also frequently an inexpensive procedure relative to the alternative replicated procedures. Consequently, Shah (1978), in many circumstances recommends the use of linearization rather than replication because of its ease of computation and generality of application.

3. Replicated Variance Estimation Procedures

There are a number of commonly used replication procedures for variance estimation. Their origins are found in the works of Mahalanobis (1944), and Deming (1956), who have advocated the use of interpenetrating or replicated samples. If a given sample design is used to select a sample, and then repeated *r* times, the final sample consists of *r* independent replicates, each with identical design. The resulting sample is called a simple replicated sample. Each replicate sample then provides an estimate of the parameter of interest. The variability among the *r* replicate estimates then gives a measure of the variance of the overall sample estimator, which is the simple average of the *r* replicate estimators. In his work, Mahalanobis advocated the use of four replicates, while Deming proposed the use of ten.

There are two reasons that simple replicated sampling has failed to find general favour. First, the constraint that the sample design includes *r* replicates often leads to less precise estimation than otherwise would be the case. Fewer gains in precision are possible through the use of techniques such as stratification and systematic selection, since the sample size of each replicate is only one *r*th of the full sample. This loss in precision can be considerable, particularly if *r* is large (more than five). Second, unless *r* is reasonably large (at least ten), the precision of the simple replication variance estimator is poor. For the most part, this negates the benefits of having a procedure that gives unbiased variance estimates.

The attractive concept of using the variability among estimates derived from subsample replicates to provide an overall estimator of variance has been adapted to derive a number of pseudoreplication procedures. These do not impose restrictions on the design of the overall sample as simple replication does, and are designed to give variance estimators of low bias and adequate precision. Three methods are in current use for large scale sample surveys: the random groups method, balanced (half-sample) repeated replication, and jackknifing. These are discussed in the next three sections.

4. The Random Groups Method

The random groups procedure, which was first developed at the U.S. Bureau of the Census, is described by Hansen, Hurwitz, and Madow (1953, Vol. 1, Section 10.16). Of the three commonly used pseudoreplication procedures, it is the closest to the original concept of replicated sampling. The major difference between the random groups method and the original simple replicated sampling is that the replicates are not formed independently.

The original sample is divided into a number of groups. An estimate of the population parameter of interest is formed from each group (replicate), and the variability among these estimates is used to obtain an estimate of the variance of the parameter estimator computed from the full sample. To ensure that the variance estimator has small bias, the groups must be formed so that each reflects the design of the full sample.

The random groups procedure attempts to impart a simple replicated design after the sample has been selected. Consequently, it faces the same problems as simple replicated sampling with regard to the choice of the number of groups to be used.

The larger the number of groups, the greater the departure from true replication. This gives rise to increased bias in the estimation of variance. In particular, if the number of groups exceeds the sample size of one or more strata, then not all groups can contain a sampled PSU from each stratum. In practice, this can be handled by collapsing strata for the purpose of forming replicates, but then the groups do not constitute true replicates, and bias in variance estimation results.

If the number of random groups is kept small to minimize the bias of variance estimation, as with simple replication, the precision of variance estimation is low. Thus, for many designs the random groups estimator of variance is poor, having either high bias, low precision, or a combination of these.

The random groups method is most useful in surveys using a large sample of PSUs, where either many PSUs are selected per stratum, or few gains are believed to result from the finer levels of stratification. The Retail Trade Survey of the U.S. Bureau of the Census fulfils these requirements, and 16 random groups are used for variance estimation (see Wolter et al. (1976)).

5. Balanced Repeated Replication

For surveys with a large number of strata, but only few sampled PSUs per stratum, the random groups method described in Section 4 is unsatisfactory. The method of balanced repeated replication (BRR) has been developed for use with such designs, and has proved most useful in the common case where two PSUs are selected per stratum.

During the 1960s, a number of statisticians developed the idea of using repeated replications based on half-samples to estimate variances. McCarthy (1966, 1969) discovered a means of deriving balanced half-sample repeated replications. This method limits the number of replicates required while it achieves high precision.

For designs where two PSUs are selected per stratum, the procedure is as follows. Let μ

be the population parameter of interest, estimated from the sample by $\hat{\mu}$. Two groups of PSUs are formed. For each stratum h, randomly denote one of the two sampled PSUs as belonging to group 1, the other to group 2. Let $\hat{\mu}'$ be the estimate of μ derived from the half-sample of PSUs from group 1, and $\hat{\mu}''$ the estimate of μ derived from the complementary half-sample of PSUs, taken from group 2. The operation of forming random half-samples can be repeated T times, with the half-sample estimates from the tth repetition being denoted by μ_t' and μ_t''. An estimator of variance is given by

$$v_T(\hat{\mu}) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\mu}_t'-\hat{\mu})^2.$$

For linear $\hat{\mu}$, this estimator can be expressed in an illuminating form. Let $\alpha_{th} = +1$ if in stratum h the unit from group 1 is chosen in the half-sample (rather than the complement) on the tth replication, and $\alpha_{th} = -1$ if the other unit in stratum h is chosen. Then

$$v_T(\hat{\mu}) = \frac{1}{4}\sum_{h}(\hat{\mu}_h' - \hat{\mu}_h'')^2$$

$$+\frac{1}{2T}\sum_{t}\sum_{h<k}\sum\alpha_{th}\alpha_{tk}\ (\hat{\mu}_h-\hat{\mu}_h'')(\hat{\mu}_k'-\hat{\mu}_k''),$$

where μ_h' is the estimate derived from the PSU from group 1 and stratum h, and $\hat{\mu}_h''$ is derived from the PSU from group 2 and stratum h.

The first term is the standard explicit unbiased variance estimator when two PSUs are selected with replacement per stratum; the second term reflects variability around the standard estimator, and represents the lack of precision of the replicated estimator relative to the standard form. This second term has an expected value of zero.

If the α_{th}'s are chosen so that

$$\sum_{t=1}^{T}\alpha_{th}\alpha_{tk} = 0 \qquad (1)$$

for all pairs of strata h and k, the second term in the above variance formula disappears. BRR variance estimation is defined as half-sample replication for which condition (1) holds true.

The usefulness of BRR is manifested when used for non-linear estimators. For linear estimators, BRR does not differ from the standard explicit form, and there is no point in using BRR. We assume that, since BRR is unbiased and relatively precise for linear estimators obtained by sampling with replacement of PSUs, it will have low bias and be relatively precise for non-linear estimators.

For designs in which without replacement sampling of PSUs is used, the variance estimator $V_T(\hat{\mu})$ is positively biased. Either a separate adjustment is introduced to account for this bias, or, more commonly, this bias is ignored, since it is deemed unsubstantial.

Plackett and Burman (1946) developed a method for constructing *mxm* orthogonal matrices with entries of +1 and –1, where m is a multiple of four. These can be used directly to obtain values of α_{th} satisfying (1). If there are H strata, each with two sampled PSUs, the orthogonal matrix of size T, where T is the multiple of four between H and H+3, can be used, dropping the last T–H columns. The entries of the matrix, being +1 and –1 values, can be successively substituted as the αth values, one row for each replicate and one column for each stratum. Thus, a number of replicates in the range of H to H+3 is required.

The question of what use to make of the complementary half-samples arises in the case of non-linear $\hat{\mu}$. There are four possible variations of BRR variance estimators:

$$v_{\text{BRR-S}}(\hat{\mu}) = \frac{1}{2T}\sum_{t}\{(\hat{\mu}_t'-\hat{\mu})^2 + (\hat{\mu}_t''-\hat{\mu})^2\},$$

$$v_{\text{BRR-H}}(\hat{\mu}) = \frac{1}{T}\sum_{t}(\hat{\mu}_t'-\hat{\mu})^2,$$

$$v_{\mathrm{BRR-C}}(\hat{\mu}) = \frac{1}{T}\sum_t (\hat{\mu}_t'' - \hat{\mu})^2 ,$$

$$v_{\mathrm{BRR-D}}(\hat{\mu}) = \frac{1}{4T}\sum_t (\hat{\mu}_t' - \hat{\mu}_t'')^2 . \qquad (2)$$

Since $v_{\mathrm{BRR-S}}$ is the average of $v_{\mathrm{BRR-H}}$ and $v_{\mathrm{BRR-C}}$ (which are of equivalent form, one using the half-sample and the other the complementary half-sample), it is at least as precise as the others, and equally biased. However, $v_{\mathrm{BRR-S}}$ requires the computation of half-sample estimates from both half-samples for each replicate. Thus, it is more costly than $v_{\mathrm{BRR-H}}$, and perhaps significantly so when many estimates are produced.

Another set of variance estimators for non-linear estimators can be obtained using $\hat{\hat{\mu}} = \sum_t \mu_t'/T$ in place of $\hat{\mu}$ in the estimators (2). Such estimators are unbiased for linear $\hat{\mu}$ only if the number of half-samples T is strictly greater than the number of strata H. If H is a multiple of four, $T=H+4$ half-samples must be used to retain this unbiasedness (see Lemeshow and Epp (1977)).

The estimators that use $\hat{\hat{\mu}}$ are generally not preferred to those that use $\hat{\mu}$, since they give smaller and less conservative estimates of the mean square error, as they fail to include a component for the bias of $\hat{\mu}$.

Although BRR is an effective means of maximizing the precision of variance estimation without the use of an excessive number of replicates, it can easily occur that for samples with many strata, the cost and complexity of using BRR become prohibitive when many variance estimates are required. Two adaptations of BRR have been proposed that require fewer replicates, and are correspondingly less precise, but no more biased, than the full BRR. One method is to use the combined strata technique (see for example Kalton (1977)). With this approach, strata are combined into sets. For each replicate t, all strata h in a set g are assigned the same value for α_{th}. The constraint $\sum_{t=1}^{T'} \alpha_{th}\alpha_{tk} = 0$ is imposed for pairs h and k of strata which are not within the same combined stratum g. This constraint means that, if G combined strata are formed, the number of replicates T' required is the multiple of four which will fall in the range of G to $G+3$. The orthogonal matrix of size T' is used to derive the values of α_{th}. This combined strata BRR method is somewhat less precise than the full BRR, since it uses fewer replicates.

The second procedure for reducing the number of replicates required for variance estimation, discussed by McCarthy (1966) and developed by Lee (1972, 1973), is the method of partially balanced repeated replication (PBRR). In this procedure, the strata are divided into groups, and full balancing is applied to the strata within each group. If G replicates are required for H strata, S $(=H/G)$ groups are formed with G strata in each. A GxG orthogonal matrix is used to ensure a full balance within each group. As noted by Rust (1984, Ch. 2), PBRR as proposed by Lee is equivalent to a special form of combined strata BRR, in which all combined strata have the same number of component strata.

Lee (1972, 1973) suggests methods of implementing PBRR to minimize the loss in precision over fully balanced BRR. The methods proposed require an ordering of strata by the size of their contribution to the total sampling variance. Rust (1984, Ch. 2) generalizes these findings, making use of the equivalence of PBRR and combined strata BRR. It is suggested that, given reasonable knowledge of the relative contributions of each stratum to the total sampling variance, the use of about 30 replicates will probably suffice for most purposes.

Procedures for extending BRR to the case of more than two selections per stratum have been developed. Gurney and Jewett (1975) describe a procedure that can be used for any design having a constant, prime, number of PSUs selected per stratum. Few designs meet such a requirement (other than two per stratum), and a large number of replicates may be required. These facts have limited the use of this procedure.

For more general designs, in which stratum sample sizes may vary, BRR can be implemented by randomly dividing the PSUs in each stratum into two groups of equal size, and then using BRR by treating these groups as units (see Kish and Frankel (1970)). In this case, the precision of the BRR variance estimator for linear estimators is somewhat less than that of an appropriate explicit variance estimator.

6. Jackknife Variance Estimation Procedures

The jackknife procedure was originated by Quenouille (1956) as a means of reducing the bias of parameter estimates. Tukey (1958) conjectured that the technique could be adapted to produce variance estimates for a large class of estimators, and this included finite population samples. Miller (1964, 1974) has reviewed the possible uses of the jackknife procedure in a range of statistical applications. For variance estimation, the technique consists of splitting the total sample into a set of L equal-sized, disjoint, exhaustive subsamples, dropping out each of the subsamples in turn, and estimating the population parameter of interest from the remaining units each time. The variability among these L estimates can then be used to estimate the variance of the original sample estimator.

Unlike BRR, jackknifing is not restricted to designs with two (or any other constant number) selections per stratum in order to attain full efficiency, and thus can be applied in situations not well-suited to BRR.

For a stratified design with H strata, the sample within each stratum is subdivided randomly into ℓ_h disjoint groups of equal size, giving a total of $L = \sum_{h=1}^{H} \ell_h$ dropout groups.

Let μ be the parameter of interest, $\hat{\mu}$ the estimator of μ based on the whole sample, and $\hat{\mu}_{(ih)}$ the estimator of μ based on the sample with the ith subsample from stratum h omitted. If $\hat{\mu}$ is linear, and the sample of PSUs is drawn with replacement, then

$$v_{\mathrm{GJ}}(\hat{\mu}) = \sum_{h=1}^{H} \frac{(\ell_h-1)}{\ell_h} \sum_{i=1}^{\ell_h} (\hat{\mu}_{(ih)}-\hat{\mu})^2 \qquad (3)$$

is an unbiased variance estimator for $\hat{\mu}$. For without replacement designs and non-linear estimators, the bias of $v_{\mathrm{GJ}}(\hat{\mu})$ is assumed to be relatively small. The estimator v_{GJ} in (3) requires the formation of L replicates.

The jackknife variance estimator can be generalized by dropping out only a random selection of g_h of the ℓ_h subsamples in stratum h, giving $G = \sum_{h=1}^{H} g_h$ replicates. The appropriate variance estimator, unbiased for linear $\hat{\mu}$ under with replacement sampling of PSUs, is

$$v_{\mathrm{SGJ}}(\hat{\mu}) = \sum_{h=1}^{H} \frac{(\ell_h-1)}{g_h} \sum_{i=1}^{g_h} (\hat{\mu}_{(ih)}-\hat{\mu})^2.$$

In the case where $\ell_h=2$ for every stratum, with each subsample dropped out, an alternative estimator, equivalent to (3) for linear $\hat{\mu}$, is

$$v_{\mathrm{JD}}(\hat{\mu}) = \frac{1}{4} \sum_{h=1}^{H} (\hat{\mu}_{(1h)} - \hat{\mu}_{(2h)})^2.$$

The precision of the jackknife variance estimator is maximized when each dropout group is of size one, and each unit is dropped out once. If the sample in stratum h consists of n_h PSUs, $n = \sum_{h=1}^{H} n_h$ replicates are required to give maximum precision. The variance estimator in this case is (3) with ℓ_h replaced by n_h.

Chakrabarty and Rao (1967) considered the use of the jackknife for estimating the variance of a ratio estimator. They showed that under a particular regression model the bias of the jackknife variance estimator is minimized for dropout groups of size one, with each unit dropped out once.

The jackknife variance estimation procedure can be used with combined strata also. Replicates can be formed by dropping out PSUs from several strata at the same time. As with BRR, combining strata does not add bias to the jackknife variance estimation procedure. Rust (1984, Ch. 3 and 4) discusses strategies for determining the most suitable form of jackknife variance estimator for a given survey, subject to constraints on the number of replicates to be formed.

7. Comparisons Among Variance Estimation Procedures

A number of investigators have compared the performances of the BRR, jackknife, and linearization procedures, both analytically and empirically. The empirical investigations have been of two distinct types: those using actual survey data which comparisons are based on, and those based on randomly generated data sets with known properties.

A pioneering and central investigation has been the work of Kish and Frankel (1968, 1970, 1974), and Frankel (1971)). This work includes a major study using data from the 1967 Current Population Survey conducted by the U.S. Bureau of the Census. The performances of the linearization, BRR, and jackknife procedures were compared for two PSU per stratum designs featuring 6, 12, and 30 strata. A large range of estimators were considered: ratio means, differences of ratio means, correlation coefficients, regression coefficients, partial correlation coefficients, and multiple correlation coefficients. The three variance estimation methods were compared with regard to bias, mean square error, and the coverage of confidence intervals based upon the variance estimates and parameter estimates using the appropriate coefficient from a t distribution.

There was good evidence that all three methods performed satisfactorily and with similar accuracy for ratio means and differences between them, correlation coefficients, and regression coefficients. Although there was variation across estimators and sample sizes, consistent evidence was found that linearization gave the smallest mean square error for the variance estimates, and BRR gave the largest. As expected, the bias of each technique decreased as the number of strata increased. Although linearization generally had the smallest mean square error, it did not consistently have the smallest absolute relative bias.

For this same group of estimators, all three methods showed that the quality of coverage of the t-intervals increased as the number of strata increased. BRR consistently gave the best coverage, and linearization the poorest. The intervals slightly overstated the true confidence level for the linearization and jackknife techniques, but no such consistent over- or understatement of the true level of confidence resulted from BRR-based confidence intervals.

For the BRR and jackknife methods, the performances of the different alternative forms of variance estimator were considered. The forms $v_{\text{BRR-S}}$, $v_{\text{BRR-H}}$, and $v_{\text{BRR-D}}$ (described in Section 5) were compared for BRR, and v_{GJ}, v_{SGJ}, and v_{JD}, (described in Section 6) were compared for jackknifing. Frankel (1971) recommended $v_{\text{BRR-S}}$ over $v_{\text{BRR-D}}$, and correspondingly v_{GJ} over v_{JD}. These recommendations were based on the quality of the coverage of confidence intervals derived from the variance estimates. However, the question of whether it is cost effective to compute estimates from the complementary replicates for two PSUs per stratum designs cannot be answered unequivocally. Complementary replicate estimates are required for $v_{\mathrm{BRR\text{-}S}}$, $v_{\mathrm{BRR\text{-}D}}$, v_{GJ}, and v_{JD}, but not for the estimators $v_{\mathrm{BRR\text{-}H}}$ and v_{SGJ}. The relative performances of the different variance estimators vary both with the parameter and the number of PSUs in the sample.

For estimators of partial and multiple correlation coefficients, these three methods proved much less satisfactory. The linearization procedure was not used, since Frankel found the necessary analytic differentiation intractable for these two types of estimator. The stated confidence levels of the *t*-intervals based on BRR were substantially above the true confidence levels, particularly for multiple correlation coefficients, and were even further overstated when variance estimates were derived using jackknifing. These poor performances persisted over designs with different numbers of strata.

Using numerical differentiation, Woodruff and Causey (1976) were able to consider the performance of linearization for partial and multiple correlations. They used the same set of data (but not the same samples) as Kish and Frankel. Woodruff and Causey found that, as with the other estimators considered by Kish and Frankel, linearization gave a lower mean square error for the estimates of variance for both partial and multiple correlation coefficients across 6, 12, and 30 stratum designs. The quality of coverage of symmetric *t*-intervals based on linearization was worse than for BRR and jackknifing, again consistent with the results for other estimators. This is particularly notable in view of the poor quality of confidence interval estimation shown by BRR and jackknifing, especially for multiple correlation coefficients.

No definitive reason for the poor performance of all three variance estimation methods for partial and multiple correlation coefficients in these two studies has been found. Kalton (1974) points out that a simple correlation is a special case of a multiple correlation, and yet the performance of these methods of estimating the variances of simple correlations was found to be most satisfactory.

Bean (1975) undertook an empirical study using data from the 1969 Health Interview Survey of the U.S. National Center for Health Statistics. The performances of the linearization variance estimator and BRR were considered. For BRR, the variance estimators $v_{\mathrm{BRR\text{-}S}}$ and $v_{\mathrm{BRR\text{-}H}}$ were used. Ratio estimates using poststratification were studied for several two-PSUs-per-stratum designs. The findings were similar to those of Kish and Frankel. The bias, the mean square error, and the confidence interval performance were compared for five characteristics and three different self-weighting two-stage designs, varying only in their overall sampling rate. This variation was introduced to assess the effect of ignoring the without replacement aspect of PSU selection.

The three variance estimators were found to have comparable and negligible bias for all sample designs and estimators. Treating the PSUs as if they had been selected with replacement did not introduce a substantial bias. The linearization estimator showed greatest consistency. The tendency for bias to decrease as sample size increased was strongest for this method. The two BRR variance estimators showed similar bias, as expected, with $v_{\mathrm{BRR\text{-}S}}$ having smaller variance than $v_{\mathrm{BRR\text{-}H}}$. The linearization estimator showed smaller mean square error than BRR, but this effect was notable only for one estimator.

The analysis of the performances of confidence intervals based on these variance estimators showed patterns very similar to those found by Kish and Frankel. Both linearization and BRR gave good results, with the true confidence levels close to the stated levels. The performance of BRR was consistently a little better than that of linearization. The findings from this study for estimates of ratios were thus consistent with the earlier findings of Kish and Frankel.

Campbell and Meyer (1978) also conducted a study to compare linearization, BRR, and jackknifing for variance estimation. They studied the performance of these methods for designs with two PSUs per stratum, but used randomly generated data, so that they could compare the performances over different population conditions. The main criterion used to judge the performance of the variance estimators was the quality of coverage of symmetric 90 % and 95 % confidence intervals, based on the *t* distribution. The parameters estimated were means, ratios, population variances and logarithmically transformed population variances, regression coefficients, simple correlation coefficients, and the Fisher *z* transform of a simple correlation coefficient. Campbell and Meyer found that, averaged across parameters, BRR gave better performance across a variety of populations than either of the other two procedures. The study also indicated that the performance of all methods was poor for the populations and sample sizes studied when the parameter estimated was a population variance or its logarithmic transform. The study's findings strengthen the impression that BRR gives superior confidence interval statements in comparison to the other two procedures, and that the performances of all three methods (but not their relative performances) vary markedly with the parameter being estimated.

Another finding common to the studies of Kish and Frankel (1970), Bean (1975), and a study by Simmons and Baird (1968) is that there is little loss in accuracy in using the original whole sample weights (needed to account for the effects of poststratification and nonresponse) for units in each half-sample of each replicate. The alternative is to derive new weights each time that are appropriate to the particular half-sample. This finding has important implications for the practical use of replicated techniques, since the procedure of using the same weights is simpler and cheaper. It appears that in the applications considered, sampling variability in the unit weights did not make an important contribution to the sampling variance.

Lemeshow (1976, 1979), using randomly generated data, studied the effects of these alternative weighting procedures for half-samples, particularly when estimates of variance are required for subclass estimates, a common requirement in practice. Both BRR (Lemeshow (1976, 1979)) and jackknifing (Lemeshow (1976)) were considered. Lemeshow found that the use of constant weights throughout proved far from satisfactory, for the situations considered, giving both greater bias and lower precision of variance estimation. Lemeshow concluded that, especially when subclass estimates are important, the use of the reweighting method is preferable to the use of a single set of weights. These findings were considered not to be inconsistent with those of Kish and Frankel and Simmons and Baird, who found a single set of weights to be satisfactory. These earlier investigations did not consider subclass estimates, and there were not large differences among stratum means in these studies, an important factor contributing to Lemeshow's findings.

Lemeshow's findings suggest that rather than using fully balanced BRR with fixed weights, new weights for each half-sample should be used. However, it might prove necessary to use fewer replicates. Lemeshow's findings are particularly useful when subclass estimates are important. The cost of performing the reweighting can be balanced by the reduction in cost resulting from the use of fewer replicates. This argument also applies to the use of jackknifing.

Other studies have used randomly generated data to study the properties of variance estimators for complex estimators. Lemeshow, Hosmer, and Hislop (1980) conducted a Monte Carlo study comparing the performance of linearization, BRR, and the jackknife for estimating the variance of a combined ratio estimate in the presence of non-normality of the underlying distributions. They found that the estimators performed badly when only three strata with small sample sizes were used. However, with 16 strata or a reasonably large sample of second stage units in each stratum all three techniques performed satisfactorily and were comparable over a variety of distributions. Lemeshow and Levy (1979) had previously studied the performance of the methods for combined ratio estimation when the underlying variables are normally distributed, again using a Monte Carlo approach. In this case, they found all three methods to be satisfactory. It would seem that in actual survey practice, at least for the combined ratio estimator, these techniques are robust to the underlying population distribution, and that they differ little on the basis of this criterion.

A number of authors have considered these variance estimation techniques analytically. The results obtained have concentrated on particular aspects and properties of the variance estimators, whereas much of the empirical analysis has concentrated on overall measures of performance.

Krewski and Rao (1981) showed that for two-PSUs-per-stratum designs, using with replacement selection of PSUs, as the number of strata increases, the variance estimators obtained via the linearization, jackknife, and BRR techniques are consistent, under reasonably limiting conditions. This result holds for all variations of BRR and jackknife variance estimators. This result, while suggesting small differences in the quality of the three procedures for samples with many strata, does not indicate the relative performances of the three methods in practice. In the same paper, Krewski and Rao consider the performances of the three methods for combined ratio estimation. Exact analytic results on the bias and precision of each of the three methods were sought under a general regression model. The results are for stratified single stage sampling with proportional allocation. The relativities when comparing the size of the biases of the three methods depend upon the heteroscedasticity of the error distribution of the regression model in each stratum, the intercepts of the regression models, the stratum sizes, and the number of strata. The bias for jackknifing and BRR depend upon the exact form of the variance estimator used.

Even for a single estimator under restricted population model and sample design conditions, general relationships are not discernible in either the signs or the magnitudes of the biases of the three methods. When the mean square error (MSE) of the methods was considered, it was found that the MSE of the BRR estimator $v_{\text{BRR-H}}$ exceeds that of the linearization procedure. However, the size of the excess is not generalizable across different model conditions. The MSEs for jackknifing and the other forms of the BRR estimator proved intractable (as was even the bias for some forms of jackknife estimator).

The results of Krewski and Rao's work illustrate two aspects of the non-asymptotic analytic examination of these variance estimators. Results can be difficult to derive, especially considering the range of possible variance estimators that can be used with BRR and jackknifing. When the results are derivable, they tend to be of limited generality, and hence of little use for discerning which technique is the most suitable for a range of estimators, and for a population whose characteristics are at best only broadly known.

Mellor (1973) compared various replicated variance estimators both empirically and analytically. He showed that among a class of replicated variance estimators that he called "balanced" (essentially estimators in which each sample unit appears in an equal number of replicates), the jackknife procedure is optimal for linear estimators (for which it is equivalent to BRR for two-PSUs-per-stratum designs).

Mellor also considered the large-sample properties of some replicated variance procedures for non-linear estimators. He generalized the results of Brillinger (1964) to show that a variety of replication estimators, including jackknifing, result in the quantity $(y-Y)/\sqrt{v}$ being distributed asymptotically as a t distribution with b–1 degrees of freedom, where b is the number of replicates used, y is the estimate of Y, and v is the variance estimate. This asymptotic result holds as the sample size increases and b remains fixed. This large-sample result gives reason for confidence in the use of replicated procedures for variance estimation, and in particular for estimating confidence intervals.

It seems clear that in order to discriminate analytically between variance estimation procedures for non-linear estimators, higher order analyses are required. Recently, Rao and Wu (1983a) have used second order analyses to compare BRR, jackknifing, and the linearization method. In particular, the asymptotic relative biases of the three procedures were considered. The authors show that for designs with two PSUs selected per stratum with replacement the BRR (using $v_{\text{BRR-D}}$) and jackknife (using v_{JD}) procedures are identical to the linearization procedure for *quadratic* estimators. Results are also derived showing that seven variations of the jackknife variance estimator are equivalent to higher order terms. However, the three variations of the BRR estimator considered do not show this equivalence to the same order. The jackknife and linearization estimators are equivalent to higher order terms. The BRR estimators $v_{\text{BRR-S}}$ and $v_{\text{BRR-D}}$, in which complementary half-samples are used, show closer relation to the linearization procedure than $v_{\text{BRR-H}}$, which does not make use of complementary half-samples values.

These results, and in particular the equivalence of the linearization procedure and all seven variations of the jackknife procedure considered to stochastic order n^{-3}, suggest that for designs where the sample size of PSUs is large there is very little difference among these estimators, and that practical matters should be considered when choosing among them.

Rao and Wu (1983a) also considered these variance estimators non-asymptotically, specifically for use with combined ratio estimation. They found that the biases of both the jackknife and the BRR variance estimators exceeded that of the linearization estimator. The result for jackknife estimators is based on the assumption that the squared residuals about the ratio slope are positively correlated with the auxiliary variable, a common occurrence in practice.

Little research appears to have been undertaken to compare the properties of the random groups method with those of linearization, BRR and jackknifing. This is perhaps because the random groups method is a useful procedure for relatively few survey designs. Also, the method of forming the random groups depends upon the design. This means that it is difficult to consider the general properties of this method.

8. Other Properties of Replicated Techniques

A number of studies have shown replication to be a useful tool in analytic surveys, particularly for the estimation of covariance matrices. Koch and Lemeshow (1972) used a replication estimator of a covariance matrix in analyzing differences between domain means in the U.S. Health Examination Survey. Freeman (1975) undertook an empirical investigation of replicated estimates of covariance matrices, and Koch, Freeman, and Freeman (1975) discuss the use of replication methods for use in univariate and multivariate comparisons among cross-classified domains. Chapman

(1966) and Nathan (1975) present an approximate test for independence in contingency tables using replicated estimates of the covariance matrices.

In addition to the findings of Kish and Frankel (1970, 1974), which showed that both BRR and jackknifing could perform poorly when used with estimates of partial and multiple correlation coefficients, there is further evidence of the poor performance of replicated procedures in special circumstances. Brillinger (1977) shows analytically that the jackknife variance estimator will have a severe downward bias when used for an estimate of the median of a population from a simple random sample. The variance estimate derived using the jackknife estimator with dropout groups of size one is shown to be asymptotically for only 25 % of the true asymptotic variance. Brillinger (1964) suggests that the poor approximation of the sample median by a linear estimator accounts for this phenomenon. This in turn suggests that the jackknife variance estimator is likely to perform poorly for other estimators not well approximated by a linear form, and that this argument should extend to other replicated variance estimation techniques.

The theoretical justifications for both BRR and jackknifing are based on the assumption that PSUs are sampled independently (i.e., with replacement). To the extent that PSU selection is not independent, and the sampling of first stage units contributes to the overall sampling variance, replicated procedures will be biased. In empirical studies, this source of bias has not been found to have a substantial adverse effect on the performance of replicated variance estimation procedures. However, one must be aware of this potential source of bias when making use of these techniques.

Replicated variance estimation methods occupy a firmly established position in the repertoire of techniques used by practising survey statisticians. Their generality, their ability to provide variance estimators for parameter estimators for which it is difficult to do so otherwise, and their relative robustness have given them their appeal. The major drawbacks have been their relatively high cost, and the lack of theoretical results concerning their accuracy in realistic cases. Asymptotic properties are generally highly favorable. Small sample Monte Carlo studies raise some doubts as to their accuracy. Both theoretical and Monte Carlo results are lacking for samples of the size and design for which the techniques are typically used. Their accuracy when used with certain types of estimators, particularly highly non-linear estimators, remains very much in doubt.

9. Concluding Remarks

Recent theoretical developments have been discussed for Taylor series linearization, increasing its generality of application (Binder (1983)). Methods of implementing the BRR and jackknifing procedures that will give adequate precision using a modest number of replicates have been proposed (see Rust (1984, Ch. 2 to 4)). Furthermore, the costs of large scale numerical manipulation via computer have decreased substantially, and it is likely that these costs will keep decreasing for some time. These developments suggest that both linearization and replication procedures for estimating sampling variances are becoming increasingly readily available to sampling practitioners.

A number of computer programs are currently available for variance estimation for complex estimators, either as stand-alone programs or as routines in statistical packages. Examples are the programs SUPER CARP (Iowa State University) and SURREGR (Research Triangle Institute, North Carolina), which use the linearization approach, and the REPERR routine of the OSIRIS IV package (Institute for Social Research, the

University of Michigan), which provides both the BRR and jackknife procedures. Details of many of the available programs are contained in Francis (1981).

At present, on theoretical grounds there appears to be little to choose among the established procedures of linearization, BRR, and jackknifing. Given that the complexities of the sample design are still frequently ignored in the analysis of survey data, it would appear that the availability, cost, and ease of use of suitable software should be the main considerations in the practical choice among these methods. All three methods offer substantial improvement over the use of traditional methods based on an assumption of simple random sampling.

In the future, the important research on these methods seems likely to be focussed on two areas: the examination of the properties of the different methods when these methods are used for making statistical inferences about population parameters, and the development of alternative methods for use in situations where none of the currently practised methods has proved successful.

Further research is needed, no doubt both theoretical and empirical, to establish the properties of linearization, BRR, and jackknifing when the variance estimator is used to derive a confidence interval for the parameter of interest. The empirical works of Frankel (1971) and Campbell and Meyer (1978) need to be followed up in an attempt to determine the desirable properties of a variance estimator when used for such a purpose, and to establish a means of discriminating among methods.

The method known as the bootstrap, introduced by Efron (1982), appears to offer a potential alternative in cases where current methods are unsatisfactory. The expense of this method seems likely to restrict its usefulness in other circumstances. A systematic evaluation of this procedure, as applied to sample surveys, is still in its early stages. The work of Rao and Wu (1983b, 1984) provides the first detailed insight into the methods for applying the bootstrap to variance estimation for complex sample designs, and the asymptotic properties of the resulting variance estimator and confidence intervals. Recently, McCarthy and Snowden (1985) have also undertaken an evaluation of bootstrapping for finite population sampling that uses simulations based on five artificial populations.

10. References

Bean, J. A. (1975): Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples. An Empirical Comparison. Vital and Health Statistics, Series 2, No. 65. U.S. Department of Health, Education, and Welfare. Washington: U.S. Government Printing Office.

Binder, D. A. (1983): On the Variances of Asymptotically Normal Estimators From Complex Surveys. International Statistical Review, 51, pp. 279–292.

Brillinger, D. R. (1964): The Asymptotic Behavior of Tukey's General Method of Setting Approximate Confidence Intervals (the Jackknife) When Applied to Maximum Likelihood Estimates. Review of the International Statistical Institute, 32, pp. 202–206.

Brillinger, D. R. (1977): Approximate Estimation of the Standard Errors of Complex Statistics Based on Sample Surveys. New Zealand Statistician, 11(2), pp. 35–41.

Campbell, C. and Meyer, M. (1978): Some Properties of T Confidence Intervals for Survey Data. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 437–442.

Chakrabarty, R. P. and Rao, J. N. K. (1967): The Bias and Stability of the Jackknife Variance Estimator in Ratio Estimation. American Statistical Association, Proceedings of the Social Statistics Section, pp. 326–331.

Chapman, D. W. (1966): An Approximate Test for Independence Based on Replications of a Complex Sample Survey Design. M. S. thesis, Cornell University.

Cochran, W. G. (1977): Sampling Techniques. Third Edition. New York: Wiley.

Deming, W. E. (1956): On Simplifications of Sampling Design Through Replication With Equal Probabilities and Without Stages. Journal of the American Statistical Association, 51, pp. 24–53.

Efron, B. (1982): The Jackknife, the Bootstrap and Other Resampling Plans. SIAM, Philadelphia.

Francis, I. (1981): Statistical Software: A Comparative Review. New York: North Holland.

Frankel, M. R. (1971): Inference From Survey Samples. Ann Arbor: Institute for Social Research, University of Michigan.

Freeman, D. H. (1975): The Regression Analysis of Data From Complex Sample Surveys: An Empirical Investigation of Covariance Matrix Estimation. Ph. D. thesis, University of North Carolina.

Gurney, M. and Jewett, R. S. (1975): Constructing Orthogonal Replications for Variance Estimation. Journal of the Americal Statistical Association, 70, pp. 819–821.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953): Sample Survey Methods and Theory. Vol. I: Methods and Applications. Vol. II: Theory. New York: Wiley.

Kalton, G. (1974): Discussion of: Inference From Complex Samples, by L. Kish and M. R. Frankel. Journal of the Royal Statistical Society, Series B, 36, pp. 1–37.

Kalton, G. (1977): Practical Methods for Estimating Survey Sampling Errors. Bulletin of the International Statistical Institute, 47(3), pp. 495–514.

Kish, L. (1965): Survey Sampling. New York: Wiley.

Kish, L. and Frankel, M. R. (1968): Balanced Repeated Replication for Analytical Statistics. American Statistical Association, Proceedings of the Social Statistics Section, pp. 2–11.

Kish, L. and Frankel, M. R. (1970): Balanced Repeated Replication for Standard Errors. Journal of the American Statistical Association, 65, pp. 1071–1094.

Kish, L. and Frankel, M. R. (1974): Inference From Complex Samples. Journal of the Royal Statistical Society, Series B, 36, pp. 1–37.

Koch, G. G., Freeman, D. H. and Freeman, J. L. (1975): Strategies in the Multivariate Analysis of Data From Complex Surveys. International Statistical Review, 43, pp. 55–74.

Koch, G. G. and Lemeshow, S. (1972): An Application of Multivariate Analysis to Complex Sample Survey Data. Journal of the American Statistical Association, 67, pp. 780–782.

Krewski, D. and Rao, J. N. K. (1981): Inference From Stratified Samples: Properties of the Linearization, Jackknife, and Balanced Repeated Replication Methods. Annals of Statistics, 9(5), pp. 1010–1019.

Lee, K–H. (1972): Partially Balanced Designs for Half Sample Replication Method of Variance Estimation. Journal of the American Statistical Association, 67, pp. 324–334.

Lee, K–H. (1973): Using Partially Balanced Designs for the Half Sample Replication Method of Variance Estimation. Journal of the American Statistical Association, 68, pp. 612–614.

Lemeshow, S. (1976): An Evaluation of the Performance of the Balanced Half-sample and Jackknife Variance Estimation Techniques. Ph. D. thesis, University of California, Los Angeles.

Lemeshow, S. (1979): The Use of Unique Statistical Weights for Estimating Variances With the Balanced Half-sample Technique. Journal of Statistical Planning and Inference, 3, pp. 315–323.

Lemeshow, S. and Epp, R. (1977): Properties of the Balanced Half-sample and Jackknife Variance Estimation Techniques in the Linear Case. Communications in Statistics, Series A 6(13), pp. 1259–1274.

Lemeshow, S., Hosmer, D. W. and Hislop, D. (1980): The Effect of Non-normality on Estimating the Variance of the Combined Ratio Estimate in Complex Surveys. Communications in Statistics, Series B 9(4), pp. 371–387.

Lemeshow, S. and Levy, P. S. (1979): Estimating the Variance of Ratio Estimates in Complex Sample Surveys With Two Primary Sampling Units Per Stratum – A Comparison of Balanced Repeated Replication and Jackknife Techniques. Journal of Statistical Computing and Simulation, 8, pp. 191–205.

McCarthy, P. J. (1966): Replication. An Approach to the Analysis of Data From Complex Surveys. Vital and Health Statistics, Series 2, No. 14, U. S. Department of Health, Education, and Welfare. Washington: U. S. Government Printing Office.

McCarthy, P. J. (1969): Pseudo-replication: Half-samples. International Statistical Review, 37, pp. 239–264.

McCarthy, P. J. and Snowden, C. B. (1985): The Bootstrap and Finite Population Sampling. Data Evaluation and Methods Research, Series 2, No. 95. U. S. Department of Health and Human Services. Washington: U. S. Government Printing Office.

Mahalanobis, P. C. (1944): On Large Scale Sample Surveys. Philosophical Transactions of the Royal Society, B 231, pp. 329–451.

Mellor, R. W. (1973): Subsample Replication Variance Estimators. Ph. D. thesis, Harvard University.

Miller, R. G. Jr. (1964): A Trustworthy Jackknife. Annals of Mathematical Statistics, 35, pp. 1594–1605.

Miller, R. G. Jr. (1974): The Jackknife – A Review. Biometrika, 61, pp. 1–15.

Nathan, G. (1975): Tests of Independence in Contingency Tables From Stratified Proportional Samples. Sankhyā, C 37, pp. 77–87.

Plackett, R. L. and Burman, P. J. (1946): The Design of Optimum Factorial Experiments. Biometrika, 33, pp. 305–325.

Quenouille, M. H. (1956): Notes on Bias in Estimation. Biometrika, 43, pp. 353–360.

Rao, J. N. K. and Wu, C. F. J. (1983a): Inference From Stratified Samples: Second Order Analysis of Three Methods for Nonlinear Statistics. Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, No. 7.

Rao, J. N. K. and Wu, C. F. J. (1983b): Bootstrap Inference With Stratified Samples. Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University, Ottawa, No. 19.

Rao, J. N. K. and Wu, C. F. J. (1984): Bootstrap Inference for Sample Surveys. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 106–112.

Rust, K. F. (1984): Techniques for Estimating Variances for Sample Surveys. Ph. D. thesis, University of Michigan.

Shah, B. V. (1978): Variance Estimates for Complex Statistics From Multi-Stage Sample Surveys. Chapter 3 in Namboodiri, K. N. (Ed.), Survey Sampling and Measurement. New York: Academic Press.

Shah, B. V., Holt, M. M. and Folsom, R. E. (1977): Inference About Regression Models From Sample Survey Data. Bulletin of the International Statistical Institute, 47(3), pp. 43–57.

Simmons, W. R. and Baird, J. T. (1968): Pseudo-replication in the NCHS Health Examination Survey. American Statistical

Association, Proceedings of the Social Statistics Section, pp. 19–32.

Tepping, B. J. (1968): Variance Estimation in Complex Surveys. American Statistical Association, Proceedings of the Social Statistics Section, pp. 11–18.

Tukey, J. W. (1958): Bias and Confidence in Not-quite Large Samples. Abstract. Annuals of Mathematical Statistics, 29, p. 614.

Wolter, K. M., Isaki, C. T., Tyler, R. S., Monsour, N. J. and Meyes, F. M. (1976): Sample Selection and Estimation Aspects of the Census Bureau's Monthly Business Surveys. American Statistical Association, Proceedings of the Business and Economics Statistics Section, pp. 99–109.

Woodruff, R. S. (1971): A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association, 66, pp. 411–414.

Woodruff, R. S. and Causey, B. D. (1976): Computerized Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association, 71, pp. 315–321.

Received Mars 1985
Revised August 1985