Chapter 1

PROBABILITY AND MEASURE

(March 30, 2003)

1. The Texas lotto

1.1 Introduction

Texans (used to) play the lotto by selecting six different numbers between 1 and 50, which cost \$1 for each combination[1]. Twice a week, on Wednesday and Saturday at 10 PM, six ping-pong balls are released without replacement from a rotating plastic ball containing 50 ping-pong balls numbered 1 through 50. The winner of the jackpot (which occasionally accumulates to 60 or more million dollars!) is the one who has all six drawn numbers correct, where the order in which the numbers are drawn does not matter. What are the odds of winning if you play one set of six numbers only?

In order to answer this question, suppose first that the order of the numbers does matter. Then the number of *ordered* sets of 6 out of 50 numbers is: 50 possibilities for the first drawn number, times 49 possibilities for the second drawn number, times 48 possibilities for the third drawn number, times 47 possibilities for the fourth drawn number, times 46 possibilities for the fifth drawn number, times 45 possibilities for the sixth drawn number:

$$\prod_{j=0}^{5}(50 - j) = \prod_{k=45}^{50} k = \frac{\prod_{k=1}^{50} k}{\prod_{k=1}^{50-6} k} = \frac{50!}{(50 - 6)!}.$$

The notation $n!$, read: n factorial, stands for the product of the natural numbers 1 through n:

$$n! = 1\times 2\times ....\times(n-1)\times n \ \ if\ n > 0, \quad 0! = 1.$$

The reason for defining $0! = 1$ will be explained below.

Since a set of six given numbers can be permutated in 6! ways, we need to correct the

[1] In the Spring of 2000 the Texas Lottery has changed the rules: The number of balls has been increased to 54, in order to create a larger jackpot. The official reason for this change is to make playing the lotto more attractive, because a higher jackpot will make the lotto game more exciting. Of course, the actual reason is to boost the lotto revenues!

above number for the 6! replications of each unordered set of six given numbers. Therefore, the number of sets of six *unordered* numbers out of 50 is:

$$\binom{50}{6} \overset{def.}{=} \frac{50!}{6!(50-6)!} = 15{,}890{,}700.$$

Thus, the probability of winning the Texas lotto if you play only one combination of six numbers is 1/15,890,700. [2]

1.2 Binomial numbers

In general, the number of ways we can draw a set of *k* *unordered* objects out of a set of *n* objects *without* replacement is:

$$\binom{n}{k} \overset{def.}{=} \frac{n!}{k!(n-k)!}. \tag{1}$$

These (binomial) numbers[3], read as: *n* choose *k*, also appear as coefficients in the binomial expansion

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}. \tag{2}$$

The reason for defining 0! = 1 is now that the first and last coefficients in this binomial expansion are always equal to 1:

[2] Under the new rules (see footnote 1), this probability is: 1/25,827,165.

[3] These binomial numbers can be computed using the "Tools → Discrete distribution tools" menu of *EasyReg International*, the free econometrics software package developed by the author. *EasyReg International* can be downloaded from web page http://econ.la.psu.edu/~hbierens/EASYREG.HTM

$$\binom{n}{0} = \binom{n}{n} = \frac{n!}{0!n!} = \frac{1}{0!} = 1 .$$

For not too large an n the binomial numbers (1) can be computed recursively by hand, using the *Triangle of Pascal*:

$$\begin{array}{ccccccccccccc}
 & & & & & & 1 & & & & & & \\
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & \cdots & & \cdots & & \cdots & & \cdots & & \cdots & & 1
\end{array} \qquad (3)$$

Except for the 1's on the legs and top of the triangle, the entries are the sum of the adjacent numbers on the previous line, which is due to the easy equality:

$$\binom{n-1}{k-1} + \binom{n-1}{k} = \binom{n}{k} \quad for\ n \geq 2,\ k = 1,....,n-1. \qquad (4)$$

Thus, the top 1 corresponds to $n = 0$, the second row corresponds to $n = 1$, the third row corresponds to $n = 2$, etc., and for each row n+1, the entries are the binomial numbers (1) for k = 0,....,n. For example, for $n = 4$ the coefficients of $a^k b^{n-k}$ in the binomial expansion (2) can be found on row 5 of the triangle: $(a + b)^4 = 1\times a^4 + 4\times a^3 b + 6\times a^2 b^2 + 4\times ab^3 + 1\times b^4$.

1.3 Sample space

The Texas lotto is an example of a statistical experiment. The set of possible outcomes of this statistical experiment is called the *sample space*, and is usually denoted by Ω. In the Texas lotto case Ω contains N = 15,890,700 elements: $\Omega = \{\omega_1,.....,\omega_N\}$, where each element ω_j is a set itself consisting of six different numbers ranging from 1 to 50, such that for any pair ω_i, ω_j with $i \neq j$, $\omega_i \neq \omega_j$. Since in this case the elements ω_j of Ω are sets themselves, the condition $\omega_i \neq \omega_j$ for $i \neq j$ is equivalent to the condition that $\omega_i \cap \omega_j \notin \Omega$.

1.4 Algebras and $\sigma-$algebras of events

A set $\{\omega_{j_1},....,\omega_{j_k}\}$ of different number combinations you can bet on is called an *event*. The collection of all these events, denoted by $\mathscr{F}$, is a "family" of subsets of the sample space Ω. In the Texas lotto case the collection $\mathscr{F}$ consists of all subsets of Ω, including Ω itself and the empty set $\emptyset$.[4] In principle you could bet on all number combinations if you are rich enough (it will cost you \$15,890,700). Therefore, the sample space Ω itself is included in $\mathscr{F}$. You could also decide not to play at all. This event can be identified as the empty set $\emptyset$. For the sake of completeness it is included in $\mathscr{F}$ as well.

Since in the Texas lotto case the collection $\mathscr{F}$ contains all subsets of Ω, it automatically satisfies the following conditions:

$$\textit{If } A \in \mathscr{F} \textit{ then } \tilde{A} = \Omega\backslash A \in \mathscr{F}, \tag{5}$$

where $\tilde{A} = \Omega\backslash A$ is the *complement* of the set A (relative to the set Ω), i.e., the set of all elements of Ω that are not contained in A;

$$\textit{If } A,B \in \mathscr{F} \textit{ then } A\cup B \in \mathscr{F}. \tag{6}$$

By induction, the latter condition extends to any finite union of sets in $\mathscr{F}$: If $A_j \in \mathscr{F}$ for j = 1,2,...,n, then $\cup_{j=1}^{n}A_j \in \mathscr{F}$.

DEFINITION 1: *A collection $\mathscr{F}$ of subsets of a non-empty set Ω satisfying the conditions* (5) *and* (6) *is called an* ***algebra***.[5]

In the Texas lotto example the sample space Ω is finite, and therefore the collection $\mathscr{F}$ of subsets of Ω is finite as well. Consequently, in this case the condition (6) extends to:

$$\textit{If } A_j \in \mathscr{F} \textit{ for } j = 1,2,.... \textit{ then } \cup_{j=1}^{\infty}A_j \in \mathscr{F}. \tag{7}$$

[4] Note that the latter phrase is superfluous, because $\Omega \subset \Omega$ reads: every element of Ω is included in Ω, which is clearly a true statement, and $\emptyset \subset \Omega$ is true because $\emptyset \subset \emptyset\cup\Omega = \Omega$.

[5] Also called a **Field**.

However, since in this case the collection $\mathscr{F}$ of subsets of Ω is finite, there are only a finite number of distinct sets $A_j \in \mathscr{F}$. Therefore, in the Texas lotto case the countable infinite union $\cup_{j=1}^{\infty} A_j$ in (7) involves only a finite number of distinct sets A_j; the other sets are replications of these distinct sets. Thus, condition (7) does not require that all the sets $A_j \in \mathscr{F}$ are different.

DEFINITION 2: *A collection $\mathscr{F}$ of subsets of a non-empty set Ω satisfying the conditions* (5) *and* (7) *is called a* ***σ-algebra***.[6]

1.5 Probability measure

Now let us return to the Texas lotto example. The odds, or probability, of winning is $1/N$ for each valid combination ω_j of six numbers, hence if you play *n* different valid number combinations $\{\omega_{j_1},...,\omega_{j_n}\}$, the probability of winning is *n/N*: $P(\{\omega_{j_1},...,\omega_{j_n}\}) = n/N$. Thus, in the Texas lotto case the probability $P(A)$, $A \in \mathscr{F}$, is given by the number *n* of elements in the set *A*, divided by the total number *N* of elements in Ω. In particular we have $P(\Omega) = 1$, and if you do not play at all the probability of winning is zero: $P(\varnothing) = 0$.

The function $P(A)$, $A \in \mathscr{F}$, is called a probability measure: it assigns a number $P(A) \in [0,1]$ to each set $A \in \mathscr{F}$. Not every function which assigns numbers in [0,1] to the sets in $\mathscr{F}$ is a probability measure, though:

DEFINITION 3: *A mapping P: $\mathscr{F} \to [0,1]$ from a σ-algebra $\mathscr{F}$ of subsets of a set Ω into the unit interval is a probability measure on $\{\Omega, \mathscr{F}\}$ if it satisfies the following three conditions:*

$$\textit{If } A \in \mathscr{F} \textit{ then } P(A) \geq 0, \tag{8}$$

$$P(\Omega) = 1, \tag{9}$$

$$\textit{For disjoint sets } A_j \in \mathscr{F},\ P(\cup_{j=1}^{\infty} A_j) = \textstyle\sum_{j=1}^{\infty} P(A_j). \tag{10}$$

[6] Also called a **σ-Field**, or a **Borel Field**.

Recall that sets are *disjoint* if they have no elements in common: their intersections are the empty set.

The conditions (8) and (9) are clearly satisfied for the case of the Texas lotto. On the other hand, in the case under review the collection $\mathscr{F}$ of events contains only a finite number of sets, so that any countably infinite sequence of sets in $\mathscr{F}$ must contain sets that are the same. At first sight this seems to conflict with the implicit assumption that there always exist countably infinite sequences of **disjoint** sets for which (10) holds. It is true indeed that any countably infinite sequence of disjoint sets in a finite collection $\mathscr{F}$ of sets can only contain a finite number of non-empty sets. This is no problem though, because all the other sets are then equal to the empty set $\emptyset$. The empty set is disjoint with itself: $\emptyset \cap \emptyset = \emptyset$, and with any other set: $A \cap \emptyset = \emptyset$. Therefore, if $\mathscr{F}$ is finite then any countable infinite sequence of disjoint sets consists of a finite number of non-empty sets, and an infinite number of replications of the empty set. Consequently, if $\mathscr{F}$ is finite then it is sufficient for the verification of condition (10) to verify that for any pair of disjoint sets A_1, A_2 in $\mathscr{F}$, $P(A_1 \cup A_2) = P(A_1) + P(A_2)$. Since in the Texas lotto case $P(A_1 \cup A_2) = (n_1 + n_2)/N$, $P(A_1) = n_1/N$, and $P(A_2) = n_2/N$, where n_1 is the number of elements of A_1 and n_2 is the number of elements of A_2, the latter condition is satisfied, and so is condition (10).

The statistical experiment is now completely described by the triple $\{\Omega, \mathscr{F}, P\}$, called the *probability space*, consisting of the sample space Ω, i.e., the set of all possible outcomes of the statistical experiment involved, a $\sigma-$algebra $\mathscr{F}$ of events, i.e., a collection of subsets of the sample space Ω such that the conditions (5) and (7) are satisfied, and a probability measure P: $\mathscr{F} \rightarrow [0,1]$ satisfying the conditions (8), (9), and (10).

In the Texas lotto case the collection $\mathscr{F}$ of events is an algebra, but because $\mathscr{F}$ is finite it is automatically a $\sigma-$algebra.

2. *Quality control*

2.2 *Sampling without replacement*

As a second example, consider the following case. Suppose you are in charge of quality control in a light bulb factory. Each day N light bulbs are produced. But before they are shipped out to the retailers, the bulbs need to meet a minimum quality standard, say: no more than R out of N bulbs are allowed to be defective. The only way to verify this exactly is to try all the N bulbs out, but that will be too costly. Therefore, the way quality control is conducted in practice is to draw randomly n bulbs *without* replacement, and to check how many bulbs in this sample are defective.

Similarly to the Texas lotto case, the number M of different samples s_j of size n you can draw out of a set of N elements without replacement is:

$$M = \binom{N}{n}.$$

Each sample s_j is characterized by a number k_j of defective bulbs in the sample involved. Let K be the actual number of defective bulbs. Then $k_j \in \{0,1,...,\min(n,K)\}$.

Let $\Omega = \{0,1,....,n\}$, an let the $\sigma-$algebra $\mathscr{F}$ be the collection of all subsets of Ω. The number of samples s_j with $k_j = k \le \min(n,K)$ defective bulbs is:

$$\binom{K}{k}\binom{N-K}{n-k},$$

because there are "*K* choose *k* " ways to draw *k* unordered numbers out of *K* numbers without replacement, and "*N-K* choose *n-k*" ways to draw *n* - *k* unordered numbers out of *N* - *K* numbers without replacement. Of course, in the case that *n* > *K* the number of samples s_j with $k_j = k >$ min (*n,K*) defective bulbs is zero. Therefore, let:

$$P(\{k\}) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \quad \textit{if}\ 0 \le k \le \min(n,K), \quad P(\{k\}) = 0\ \textit{elsewhere}, \tag{11}$$

and let for each set $A = \{k_1,......,k_m\} \in \mathscr{F}$, $P(A) = \sum_{j=1}^{m}P(\{k_j\})$. (*Exercise*: Verify that this function P satisfies all the requirements of a probability measure.) The triple $\{\Omega,\mathscr{F},P\}$ is now the probability space corresponding to this statistical experiment .

The probabilities (11) are known as the *Hypergeometric*(*N,K,n*) probabilities.

2.2 Quality control in practice[7]

The problem in applying this result in quality control is that K is unknown. Therefore, in practice the following decision rule as to whether $K \le R$ or not is followed. Given a particular number $r \le n$, to be determined below, assume that the set of N bulbs meets the minimum quality requirement $K \le R$ if the number k of defective bulbs in the sample is less or equal to r. Then the set $A(r) = \{0,1,...,r\}$ corresponds to the assumption that the set of N bulbs meets the minimum quality requirement $K \le R$, hereafter indicated by "accept", with probability

$$P(A(r)) = \Sigma_{k=0}^{r} P(\{k\}) = p_r(n,K),$$

say, whereas its complement $\tilde{A}(r) = \{r+1,....,n\}$ corresponds to the assumption that this set of N bulbs does not meet this quality requirement, hereafter indicated by "reject", with corresponding probability

$$P(\tilde{A}(r)) = 1 - p_r(n,K).$$

Given r, this decision rule yields two types of errors, a type I error with probability $1 - p_r(n,K)$ if you reject while in reality $K \le R$, and a type II error with probability $p_r(K,n)$ if you accept while in reality $K > R$. The probability of a type I error has upper bound:

$$p_1(r,n) = 1 - \min_{K \le R} p_r(n,K), \tag{12}$$

say, and the probability of a type II error has upper bound

$$p_2(r,n) = \max_{K > R} p_r(n,K), \tag{13}$$

say.

In order to be able to choose r, one has to restrict either $p_1(r,n)$ or $p_2(r,n)$, or both. Usually it is former which is restricted, because a type I error may cause the whole stock of N bulbs to be trashed. Thus, allow the probability of a type I error to be maximal α, say $\alpha = 0.05$. Then r should be chosen such that $p_1(r,n) \le \alpha$. Since $p_1(r,n)$ decreases if we increase r we could in principle choose r arbitrarily large. But since $p_2(r,n)$ increases with r, we should not choose r unnecessarily large. Therefore, choose $r = r(n|\alpha)$, where $r(n|\alpha)$ is the minimum value of

[7] This section may be skipped.

r for which $p_1(r,n) \leq \alpha$. Moreover, if we allow the type II error to be maximal β, we have to choose the sample size n such that $p_2(r(n|\alpha),n) \leq \beta$.

As we will see later, this decision rule is an example of a statistical test, where H_0: $K \leq R$ is called the null hypothesis to be tested at the $\alpha \times 100\%$ significance level, against the alternative hypothesis H_1: $K > R$. The number $r(n|\alpha)$ is called the critical value of the test, and the number k of defective bulbs in the sample is called the test statistic.

2.3 Sampling with replacement

As a third example, consider the quality control example in the previous section, except that now the light bulbs are sampled *with* replacement: After testing a bulb, it is put back in the stock of N bulbs, even if the bulb involved proves to be defective. The rationale for this behavior may be that the customers will accept maximal a fraction R/N of defective bulbs, so that they will not complain as long as the actual fraction K/N of defective bulbs does not exceed R/N. In other words, why not selling defective light bulbs if it is OK with the customers?

The sample space Ω and the $\sigma-$ algebra $\mathscr{F}$ are the same as in the case of sampling without replacement, but the probability measure P is different. Consider again a sample s_j of size n containing k defective light bulbs. Since the light bulbs are put back in the stock after being tested, there are K^k ways of drawing a an *ordered* set of k defective bulbs, and $(N - K)^{n-k}$ ways of drawing an *ordered* set of n-k working bulbs. Thus the number of ways we can draw, with replacement, an ordered set of n light bulbs containing k defective bulbs is $K^k(N - K)^{n-k}$. Moreover, similarly to the Texas lotto case it follows that the number of *unordered* sets of k defective bulbs and n-k working bulbs is: n choose k. Thus, the total number of ways we can choose a sample with replacement containing k defective bulbs and n-k working bulbs in any order is:

$$\binom{n}{k} K^k(N - K)^{n-k}.$$

Moreover, the number of ways we can choose a sample of size n with replacement is N^n. Therefore,

$$P(\{k\}) = \binom{n}{k}\frac{K^k(N-K)^{n-k}}{N^n} = \binom{n}{k}p^k(1-p)^{n-k}, \quad k = 0,1,2,....,n, \ where\ p = K/N, \qquad (14)$$

and again for each set $A = \{k_1,......,k_m\} \in \mathscr{F}$, $P(A) = \sum_{j=1}^m P(\{k_j\})$. Of course, replacing $P(\{k\})$ in (11) by (14) the argument in section 2.2 still applies.

The probabilities (14) are known as the *Binomial*(n,p) probabilities.

2.4 Limits of the hypergeometric and binomial probabilities

Note that if N and K are large relative to n, the hypergeometric probability (11) and the binomial probability (14) will be almost the same. This follows from the fact that for fixed k and n:

$$P(\{k\}) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = \frac{\frac{K!(N-K)!}{k!(K-k)!(n-k)!(N-K-n+k)!}}{\frac{N!}{n!(N-n)!}}$$

$$= \frac{n!}{k!(n-k)!}\times\frac{\frac{K!(N-K)!}{(K-k)!(N-K-n+k)!}}{\frac{N!}{(N-n)!}} = \binom{n}{k}\times\frac{\frac{K!}{(K-k)!}\times\frac{(N-K)!}{(N-K-n+k)!}}{\frac{N!}{(N-n)!}} \qquad (15)$$

$$= \binom{n}{k}\times\frac{\left(\prod_{j=1}^k(K-k+j)\right)\times\left(\prod_{j=1}^{n-k}(N-K-n+k+j)\right)}{\prod_{j=1}^n(N-k+j)}$$

$$= \binom{n}{k}\times\frac{\left[\prod_{j=1}^k\left(\frac{K}{N}-\frac{k}{N}+\frac{j}{N}\right)\right]\times\left[\prod_{j=1}^{n-k}\left(1-\frac{K}{N}-\frac{n}{N}+\frac{k}{N}+\frac{j}{N}\right)\right]}{\prod_{j=1}^n\left(1-\frac{k}{N}+\frac{j}{N}\right)}$$

$$\to \binom{n}{k}p^k(1-p)^{n-k}\ \ if\ N \to \infty\ \ and\ \ K/N \to p.$$

Thus, the binomial probabilities also arise as limits of the hypergeometric probabilities.

Moreover, if in the case of the binomial probability (14) p is very small and n is very large, the probability (14) can be approximated quite well by the Poisson(λ) probability:

$$P(\{k\}) = \exp(-\lambda)\frac{\lambda^k}{k!},\ k = 0,1,2,.........., \tag{16}$$

where $\lambda = np$. This follows from (14) by choosing $p = \lambda/n$ for $n > \lambda$, with $\lambda > 0$ fixed, and letting $n \to \infty$ while keeping k fixed:

$$\begin{aligned} P(\{k\}) &= \binom{n}{k} p^k (1 - p)^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \times \frac{n!}{n^k(n-k)!} \times \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k} \to \exp(-\lambda)\frac{\lambda^k}{k!}\ \mathit{for}\ n \to \infty, \end{aligned} \tag{17}$$

because

$$\frac{n!}{n^k(n-k)!} = \frac{\prod_{j=1}^k (n-k+j)}{n^k} = \prod_{j=1}^k \left(1 - \frac{k}{n} + \frac{j}{n}\right) \to \prod_{j=1}^k 1 = 1\ \mathit{for}\ n \to \infty, \tag{18}$$

$$\left(1 - \lambda/n\right)^k \to 1\ \mathit{for}\ n \to \infty, \tag{19}$$

and

$$\left(1 - \lambda/n\right)^n \to \exp(-\lambda)\ \mathit{for}\ n \to \infty. \tag{20}$$

Since (16) is the limit of (14) for $p = \lambda/n \downarrow 0$ as $n \to \infty$, the Poisson probabilities (16) are often used to model the occurrence of *rare* events.

Note that the sample space corresponding to the Poisson probabilities is $\Omega = \{0,1,2,....\}$, and the $\sigma-$algebra $\mathscr{F}$ of events involved can be chosen to be the collection of *all* subsets of Ω, because any non-empty subset A of Ω is either countable infinite or finite. If such a subset A is countable infinite, it takes the form $A = \{k_1, k_2, k_3, ...\}$, where the k_j's are distinct non-negative integers, hence $P(A) = \sum_{j=1}^\infty P(\{k_j\})$ is well-defined. The same applies of course if A is finite: if $A = \{k_1, ..., k_m\}$ then $P(A) = \sum_{j=1}^m P(\{k_j\})$. This probability measure clearly satisfies the conditions (8), (9), and (10).

3. *Why do we need sigma-algebras of events*?

In principle we could define a probability measure on an algebra $\mathscr{F}$ of subsets of the sample space, rather than on a σ−algebra. We only need to change condition (10) to: For disjoint sets $A_j \in \mathscr{F}$ such that $\cup_{j=1}^{\infty} A_j \in \mathscr{F}$, $P(\cup_{j=1}^{\infty} A_j) = \Sigma_{j=1}^{\infty} P(A_j)$. By letting all but a finite number of these sets be equal to the empty set, this condition then reads: For disjoint sets $A_j \in \mathscr{F}$, j = 1,2,...,$n < \infty$, $P(\cup_{j=1}^{n} A_j) = \Sigma_{j=1}^{n} P(A_j)$. However, if we would confine a probability measure to an algebra all kind of useful results will no longer apply. One of these results is the so-called strong law of large numbers. See Chapter 6.

Consider the following game. Toss a fair coin infinitely many times, and assume that after each tossing you will get one dollar if the outcome it head, and nothing if the outcome is tail. The sample space Ω in this case can be expressed in terms of the winnings, i.e., each element ω of Ω takes the form of a string of infinitely many zeros and ones, for example $\omega = (1,1,0,1,0,1......)$. Now consider the event: "After n tosses the winning is k dollars". This event corresponds to the set $A_{k,n}$ of elements ω of Ω for which the sum of the first n elements in the string involved is equal to k. For example, the set $A_{1,2}$ consists of all ω of the type (1,0,......) and (0,1,......). Similarly to the example in Section 2.3 it can be shown that

$$P(A_{k,n}) = \binom{n}{k}(1/2)^n \textit{ for } k = 0,1,2,....,n, \quad P(A_{k,n}) = 0 \textit{ for } k > n \textit{ or } k < 0. \tag{21}$$

Next, for q = 1,2,.... consider the events: "After n tosses the average winning k/n is contained in the interval [0.5−1/q, 0.5+1/q]". These events correspond to the sets $B_{q,n} = \cup_{k=[n/2-n/q)]+1}^{[n/2+n/q]} A_{k,n}$, where $[x]$ denotes the smallest integer ≥ x. Then the set $\cap_{m=n}^{\infty} B_{q,m}$ corresponds to the event: "From the n-th tossing onwards the average winning will stay in the interval [0.5−1/q, 0.5+1/q]", and the set $\cup_{n=1}^{\infty}\cap_{m=n}^{\infty} B_{q,m}$ corresponds to the event: "There exists an n (possibly depending on ω) such that from the n-th tossing onwards the average winning will stay in the interval [0.5−1/q, 0.5+1/q]". Finally, the set $\cap_{q=1}^{\infty}\cup_{n=1}^{\infty}\cap_{m=n}^{\infty} B_{q,m}$ corresponds to the event: "The average winning converges to ½ as n converges to infinity". Now the strong law of large numbers states that the latter event has probability 1: $P[\cap_{q=1}^{\infty}\cup_{n=1}^{\infty}\cap_{m=n}^{\infty} B_{q,m}] = 1$. However, this probability is only defined if $\bigcap_{q=1}^{\infty}\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty}B_{q,m} \in \mathscr{F}$. In order to guarantee this, we need to require that $\mathscr{F}$ is a σ-algebra.

4. Properties of algebras and $\sigma-$algebras

4.1 General properties

In this section I will review the most important results regarding algebras, $\sigma-$algebras, and probability measures.

Our first result is trivial:

THEOREM 1 : *If an algebra contains only a finite number of sets then it is a σ-algebra. Consequently, an algebra of subsets of a finite set Ω is a $\sigma-$algebra.*

However, an algebra of subsets of an *infinite* set Ω is not necessarily a $\sigma-$algebra. A counter example is the collection $\mathscr{F}_*$ of all subsets of $\Omega = (0,1]$ of the type $(a,b]$, where $a < b$ are *rational* numbers in $[0,1]$, together with their *finite* unions and the empty set $\varnothing$. Verify that $\mathscr{F}_*$ is an algebra. Next, let $p_n = [10^n\pi]/10^n$ and $a_n = 1/p_n$, where $[x]$ means truncation to the nearest integer $\leq x$. Note that $p_n \uparrow \pi$, hence $a_n \downarrow \pi^{-1}$ as $n \to \infty$. Then for n = 1,2,3,...., $(a_n,1] \in \mathscr{F}_*$, but $\bigcup_{n=1}^{\infty}(a_n,1] = (\pi^{-1},1] \notin \mathscr{F}_*$, because π^{-1} is irrational. Thus $\mathscr{F}_*$ is *not* a $\sigma-$algebra.

THEOREM 2: *If $\mathscr{F}$ is an algebra, then $A,B \in \mathscr{F}$ implies $A\cap B \in \mathscr{F}$, hence by induction, $A_j \in \mathscr{F}$ for $j = 1,...,n < \infty$ imply $\bigcap_{j=1}^{n}A_j \in \mathscr{F}$. A collection $\mathscr{F}$ of subsets of a nonempty set Ω is an algebra if it satisfies condition* (5) *and the condition that for any pair $A,B \in \mathscr{F}$, $A\cap B \in \mathscr{F}$.*

Proof: Exercise.

Similarly, we have

THEOREM 3: *If $\mathscr{F}$ is a $\sigma-$algebra, then for any countable sequence of sets $A_j \in \mathscr{F}$, $\cap_{j=1}^{\infty} A_j \in \mathscr{F}$. A collection $\mathscr{F}$ of subsets of a nonempty set Ω is a $\sigma-$algebra if it satisfies condition* (5) *and the condition that for any countable sequence of sets $A_j \in \mathscr{F}$, $\cap_{j=1}^{\infty} A_j \in \mathscr{F}$.*

These results will be convenient in cases where it is easier to prove that (countable) intersections are included in $\mathscr{F}$ than to prove that (countable) unions are included

If $\mathscr{F}$ is already an algebra, then condition (7) alone would make it a $\sigma-$algebra. However, the condition in the following theorem is easier to verify than (7):

THEOREM 4: *If $\mathscr{F}$ is an algebra and A_j, j =1,2,3,... is a countable sequence of sets in $\mathscr{F}$, then there exists a countable sequence of* ***disjoint*** *sets B_j in $\mathscr{F}$ such that $\cup_{j=1}^{\infty} A_j = \cup_{j=1}^{\infty} B_j$. Consequently, an algebra $\mathscr{F}$ is also a $\sigma-$ algebra if for any sequence of disjoint sets B_j in $\mathscr{F}$, $\cup_{j=1}^{\infty} B_j \in \mathscr{F}$.*

Proof: Let $A_j \in \mathscr{F}$. Denote $B_1 = A_1$, $B_{n+1} = A_{n+1} \backslash (\cup_{j=1}^{n} A_j) = A_{n+1} \cap (\cap_{j=1}^{n} \tilde{A}_j)$. It follows from the properties of an algebra (see Theorem 2) that all the B_j 's are sets in $\mathscr{F}$. Moreover, it is easy to verify that the B_j's are disjoint, and that $\cup_{j=1}^{\infty} A_j = \cup_{j=1}^{\infty} B_j$. Thus, if $\cup_{j=1}^{\infty} B_j \in \mathscr{F}$ then $\cup_{j=1}^{\infty} A_j \in \mathscr{F}$. Q.E.D.

THEOREM 5: *Let $\mathscr{F}_\theta$, $\theta \in \Theta$, be a collection of $\sigma-$algebras of subsets of a given set Ω, where Θ is a possibly uncountable index set. Then $\mathscr{F} = \cap_{\theta \in \Theta} \mathscr{F}_\theta$ is a $\sigma-$algebra.*

Proof: Exercise.

Theorem 5 is important, because it guarantees that for any collection $\mathfrak{C}$ of subsets of Ω there exists a smallest $\sigma-$algebra containing $\mathfrak{C}$. By adding complements and countable unions it is possible to extend $\mathfrak{C}$ to a $\sigma-$algebra. This can always be done, because $\mathfrak{C}$ is contained in the $\sigma-$algebra of all subsets of Ω, but there is often no unique way of doing this, except in the case where $\mathfrak{C}$ is finite. Thus, let $\mathscr{F}_\theta$, $\theta \in \Theta$, be the collection of all $\sigma-$algebras containing $\mathfrak{C}$. Then $\mathscr{F} = \cap_{\theta \in \Theta} \mathscr{F}_\theta$ is the smallest $\sigma-$algebra containing $\mathfrak{C}$.

DEFINITION 4: *The smallest* $\sigma-$*algebra containing a given collection* $\mathfrak{C}$ *of sets is called the* $\sigma-$*algebra generated by* $\mathfrak{C}$, *and is usually denoted by* $\sigma(\mathfrak{C})$.

Note that $\mathscr{F} = \cup_{\theta\in\Theta}\mathscr{F}_\theta$ is not always a $\sigma-$algebra. For example, let $\Omega = [0,1]$, and let for $n \geq 1$, $\mathscr{F}_n = \{[0,1],\emptyset,[0,1-n^{-1}],(1-n^{-1},1]\}$. Then $A_n = [0,1-n^{-1}] \in \mathscr{F}_n \subset \cup_{n=1}^{\infty}\mathscr{F}_n$, but the interval $[0,1) = \cup_{n=1}^{\infty}A_n$ is not contained in any of the $\sigma-$algebras $\mathscr{F}_n$, hence $\cup_{n=1}^{\infty}A_n \notin \cup_{n=1}^{\infty}\mathscr{F}_n$.

However, it is always possible to extend $\cup_{\theta\in\Theta}\mathscr{F}_\theta$ to a $\sigma-$algebra, often in various ways, by augmenting it with the missing sets. The smallest $\sigma-$algebra containing $\cup_{\theta\in\Theta}\mathscr{F}_\theta$ is usually denoted by

$$\vee_{\theta\in\Theta}\mathscr{F}_\theta \overset{def.}{=} \sigma\left(\cup_{\theta\in\Theta}\mathscr{F}_\theta\right). \tag{22}$$

The notion of smallest σ-algebra of subsets of Ω is always relative to a given collection $\mathfrak{C}$ of subsets of Ω. Without reference to such a given collection $\mathfrak{C}$ the smallest σ-algebra of subsets of Ω is $\{\Omega,\emptyset\}$, which is called the *trivial* σ-algebra.

Moreover, similarly to Definition 4 we can define the smallest algebra of subsets of Ω containing a given collection $\mathfrak{C}$ of subsets of Ω, which we will denote by $\alpha(\mathfrak{C})$.

For example, let $\Omega = (0,1]$, and let $\mathfrak{C}$ be the collection of all intervals of the type $(a,b]$ with $0 \leq a < b \leq 1$. Then $\alpha(\mathfrak{C})$consists of the sets in $\mathfrak{C}$ together with the empty set $\emptyset$, and all finite unions of disjoint sets in $\mathfrak{C}$. To see this, check first that this collection $\alpha(\mathfrak{C})$ is an algebra, as follows.

(a) The complement of $(a,b]$ in $\mathfrak{C}$ is $(0,a]\cup(b,1]$.If $a = 0$ then $(0,a] = (0,0] = \emptyset$, and if $b = 1$ then $(b,1] = (1,1] = \emptyset$, hence $(0,a]\cup(b,1]$ is a set in $\mathfrak{C}$ or a finite union of disjoint sets in $\mathfrak{C}$.

(b) Let $(a,b]$ in $\mathfrak{C}$ and $(c,d]$ in $\mathfrak{C}$, where without loss of generality we may assume that $a \leq c$. If $b < c$ then $(a,b]\cup(c,d]$ is a union of disjoint sets in $\mathfrak{C}$. If $c \leq b \leq d$ then $(a,b]\cup(c,d] = (a,d]$ is a set in $\mathfrak{C}$ itself, and if $b > d$ then $(a,b]\cup(c,d] = (a,b]$ is a set in $\mathfrak{C}$ itself. Thus, finite unions of sets in $\mathfrak{C}$ are either sets in $\mathfrak{C}$ itself or finite unions of disjoint sets in $\mathfrak{C}$.

(c) Let $A = \cup_{j=1}^{n}(a_j,b_j]$, where $0 \leq a_1 < b_1 < a_2 < b_2 <......< a_n < b_n \leq 1$. Then

$\tilde{A} = \cup_{j=0}^{n}(b_j,a_{j+1}]$, where $b_0 = 0$ and $a_{n+1} = 1$, which is a finite union of disjoint sets in $\mathfrak{C}$ itself. Moreover, similarly to part (b) it is easy to verify that finite unions of sets of the type A can be written as finite unions of disjoint sets in $\mathfrak{C}$.

Thus, the sets in $\mathfrak{C}$ together with the empty set $\varnothing$ and all finite unions of disjoint sets in $\mathfrak{C}$ form an algebra of subsets of $\Omega = (0,1]$.

In order to verify that this is the smallest algebra containing $\mathfrak{C}$, remove one of the sets in this algebra that does not belong to $\mathfrak{C}$ itself. Since all sets in the algebra are of the type A in part (c), let us remove this particular set A. But then $\cup_{j=1}^{n}(a_j,b_j]$ is no longer included in the collection, hence we have to remove each of the intervals $(a_j,b_j]$ as well, which however is not allowed because they belong to $\mathfrak{C}$.

Note that the algebra $\alpha(\mathfrak{C})$ is not a σ-algebra, because countable infinite unions are not always included in $\alpha(\mathfrak{C})$. For example, $\cup_{n=1}^{\infty}(0,1-n^{-1}] = (0,1)$ is a countable union of sets in $\alpha(\mathfrak{C})$ which itself is not included in $\alpha(\mathfrak{C})$. However, it follows from Theorem 4 that we can extend $\alpha(\mathfrak{C})$ to a σ-algebra by adding all countable unions of disjoint sets in $\alpha(\mathfrak{C})$, and such an extension is actually the smallest σ-algebra containing $\alpha(\mathfrak{C})$, which in its turn coincide with $\sigma(\mathfrak{C})$.

4.2 Borel sets

An important special case of Definition 4 is where $\Omega = \mathbb{R}$, and $\mathfrak{C}$ is the collection of all open intervals:

$$\mathfrak{C} = \{(a,b): \forall\, a < b,\ a,b \in \mathbb{R}\}. \tag{23}$$

DEFINITION 5: *The* $\sigma-$*algebra generated by the collection* (23) *of all open intervals in* $\mathbb{R}$ *is called the Euclidean Borel field, denoted by* $\mathfrak{B}$, *and its members are called the Borel sets.*

Note, however, that $\mathfrak{B}$ can be defined in different ways, because the $\sigma-$algebras generated by the collections of open intervals, closed intervals: $\{[a,b]: \forall\, a \le b,\ a,b \in \mathbb{R}\}$, and half-open intervals, $\{(-\infty,a]: \forall\, a \in \mathbb{R}\}$, respectively, are all the same! We show this for one case only:

THEOREM 6: $\mathcal{B} = \sigma(\{(-\infty,a]: \forall a \in \mathbb{R}\})$.

Proof: Let

$$\mathfrak{C}_* = \{(-\infty,a]: \forall a \in \mathbb{R}\}. \tag{24}$$

(a) If the collection $\mathfrak{C}$ defined by (23) is contained in $\sigma(\mathfrak{C}_*)$, then $\sigma(\mathfrak{C}_*)$ is a $\sigma-$algebra containing $\mathfrak{C}$. But $\mathcal{B} = \sigma(\mathfrak{C})$ is the smallest $\sigma-$algebra containing $\mathfrak{C}$, hence $\mathcal{B} = \sigma(\mathfrak{C}) \subset \sigma(\mathfrak{C}_*)$.

In order to prove this, construct an arbitrary set (a,b) in $\mathfrak{C}$ out of countable unions and/or complements of sets in $\mathfrak{C}_*$, as follows. Let $A = (-\infty,a]$ and $B = (-\infty,b]$, where $a < b$ are arbitrary real numbers. Then $A, B \in \mathfrak{C}_*$, hence $A, \tilde{B} \in \sigma(\mathfrak{C}_*)$, and thus

$$\sim(a,b] = (-\infty,a]\cup(b,\infty) = A\cup\tilde{B} \in \sigma(\mathfrak{C}_*).$$

This implies that $\sigma(\mathfrak{C}_*)$ contains all sets of the type $(a,b]$, hence $(a,b) = \cup_{n=1}^{\infty}(a, b - (b-a)/n] \in \sigma(\mathfrak{C}_*)$. Thus, $\mathfrak{C} \subset \sigma(\mathfrak{C}_*)$.

(b) If the collection $\mathfrak{C}_*$ defined by (24) is contained in $\mathcal{B} = \sigma(\mathfrak{C})$, then $\sigma(\mathfrak{C})$ is a $\sigma-$algebra containing $\mathfrak{C}_*$. But $\sigma(\mathfrak{C}_*)$ is the smallest $\sigma-$algebra containing $\mathfrak{C}_*$, hence $\sigma(\mathfrak{C}_*) \subset \sigma(\mathfrak{C}) = \mathcal{B}$.

In order to prove the latter, observe that for $m = 1,2,....$, $A_m = \cup_{n=1}^{\infty}(a-n, a+m^{-1})$ is a countable union of sets in $\mathfrak{C}$, hence $\tilde{A}_m \in \sigma(\mathfrak{C})$, and consequently $(-\infty,a] = \cap_{m=1}^{\infty}A_m = \sim(\cup_{m=1}^{\infty}\tilde{A}_m) \in \sigma(\mathfrak{C})$. Thus, $\mathfrak{C}_* \subset \sigma(\mathfrak{C}) = \mathcal{B}$.

We have shown now that $\mathcal{B} = \sigma(\mathfrak{C}) \subset \sigma(\mathfrak{C}_*)$ and $\sigma(\mathfrak{C}_*) \subset \sigma(\mathfrak{C}) = \mathcal{B}$. Thus, $\mathcal{B}$ and $\sigma(\mathfrak{C}_*)$ are the same. Q.E.D.[8]

The notion of Borel set extends to higher dimensions as well:

DEFINITION 6: $\mathcal{B}^k = \sigma(\{\times_{j=1}^{k}(a_j,b_j): \forall a_j < b_j,\ a_j, b_j \in \mathbb{R}\})$ *is the k-dimensional Euclidean Borel field. Its members are also called Borel sets* (*in* $\mathbb{R}^k$).

[8] See also Appendix C.

Also this is only one of the ways to define higher-dimensional Borel sets. In particular, similarly to Theorem 6 we have:

THEOREM 7: $\mathscr{B}^k = \sigma(\{\times_{j=1}^k(-\infty,a_j]: \forall\, a_j \in \mathbb{R}\})$.

5. *Properties of probability measures*

The three axioms (8), (9), and (10) imply a variety of properties of probability measures. Here we list only the most important ones.

THEOREM 8: *Let* $\{\Omega,\mathscr{F},P\}$ *be a probability space. The following hold for sets in* $\mathscr{F}$:
(*a*) $P(\emptyset) = 0$,
(*b*) $P(\tilde{A}) = 1 - P(A)$,
(*c*) $A \subset B$ *implies* $P(A) \le P(B)$,
(*d*) $P(A\cup B) + P(A\cap B) = P(A) + P(B)$,
(*e*) *If* $A_n \subset A_{n+1}$ *for* $n = 1,2,...,$ *then* $P(A_n) \uparrow P(\cup_{n=1}^\infty A_n)$,
(*f*) *If* $A_n \supset A_{n+1}$ *for* $n = 1,2,...,$ *then* $P(A_n) \downarrow P(\cap_{n=1}^\infty A_n)$,
(*g*) $P(\cup_{n=1}^\infty A_n) \le \Sigma_{n=1}^\infty P(A_n)$.

Proof: (*a*)-(*c*): Easy exercises.

(*d*) $A\cup B = (A\cap\tilde{B}) \cup (A\cap B) \cup (B\cap\tilde{A})$ is a union of disjoint sets, hence by axiom (10), $P(A\cup B) = P(A\cap\tilde{B}) + P(A\cap B) + P(B\cap\tilde{A})$. Moreover, $A = (A\cap\tilde{B}) \cup (A\cap B)$ is a union of disjoint sets , hence $P(A) = P(A\cap\tilde{B}) + P(A\cap B)$, and similarly, $P(B) = P(B\cap\tilde{A}) + P(A\cap B)$. Combining these results, part (*d*) follows.

(*e*) Let $B_1 = A_1$, $B_n = A_n\backslash A_{n-1}$ for $n \ge 2$. Then $A_n = \cup_{j=1}^n A_j = \cup_{j=1}^n B_j$ and $\cup_{j=1}^\infty A_j = \cup_{j=1}^\infty B_j$. Since the B_j's are disjoint, it follows from axiom (10) that

$$P(\cup_{j=1}^\infty A_j) = \Sigma_{j=1}^\infty P(B_j) = \Sigma_{j=1}^n P(B_j) + \Sigma_{j=n+1}^\infty P(B_j) = P(A_n) + \Sigma_{j=n+1}^\infty P(B_j).$$

Part (*e*) follows now from the fact that $\Sigma_{j=n+1}^\infty P(B_j) \downarrow 0$.

(*f*) This part follows from part (*e*), using complements.

(*g*) Exercise

6. The uniform probability measure

6.1 Introduction

Fill a bowl with 10 balls numbered from 0 to 9. Draw randomly a ball from this bowl, and write down the corresponding number as the first decimal digit of a number between zero and one. For example, if the first drawn number is 4, then write down 0.4. Put the ball back in the bowl, and repeat this experiment. If for example the second ball corresponds to the number 9, then this number becomes the second decimal digit: 0.49. Repeating this experiment infinitely many times yields a random number between zero and one. Clearly, the sample space involved is the unit interval: $\Omega = [0,1]$.

For a given number $x \in [0,1]$ the probability that this random number is less or equal to x is: x. To see this, suppose that you only draw two balls, and that $x = 0.58$. If the first ball has a number less than 5, it does not matter what the second number is. There are 5 ways to draw a first number less or equal to 4, and 10 ways to draw the second number. Thus, there are 50 ways to draw a number with a first digit less or equal to 4. There is only one way to draw a first number equal to 5, and 9 ways to draw a second number less or equal to 8. Thus, the total number of ways we can generate a number less or equal to 0.58 is 59, and the total number of ways we can draw two numbers with replacement is 100. Therefore, if we only draw two balls with replacement, and use the numbers involved as the first and second decimal digit, the probability that we get a number less or equal to 0.58 is: 0.59. Similarly, if we draw 10 balls with replacement, the probability that we get a number less or equal to, say, 0.5831420385 is: 0.5831420386. In the limit the difference between x and the corresponding probability disappears. Thus, for $x \in [0,1]$ we have: $P([0,x]) = x$. By the same argument it follows that for $x \in [0,1]$, $P(\{x\}) = P([x,x]) = 0$, i.e., the probability that the random number involved will be exactly equal to a given number x is zero. Therefore, for given $x \in [0,1]$, $P((0,x]) = P([0,x)) = P((0,x)) = x$. More generally, for any interval in [0,1] the corresponding probability is the length of the interval involved, regardless as to whether the endpoints are included or not: Thus, for $0 \leq a < b \leq 1$ we have $P([a,b]) = P((a,b]) = P([a,b)) = P((a,b))$ $= b-a$. Any finite union of intervals can be written as a finite union of disjoint intervals by cutting out the overlap. Therefore, this probability measure extends to finite unions of intervals, simply by adding up the lengths of the disjoint intervals involved. Moreover, observe that the collection of all finite unions of sub-intervals in [0,1], including [0,1] itself and the empty set, is closed under the formation of complements and finite unions. Thus, we have derived the probability measure P corresponding the statistical experiment under review for an *algebra* $\mathscr{F}_0$ of subsets of [0,1], namely

$$\mathscr{F}_0 = \{(a,b),[a,b],(a,b],[a,b),\ \forall a,b\in[0,1],\ a\leq b,\ \textit{and their finite unions}\}, \tag{25}$$

where $[a,a]$ is the singleton $\{a\}$, and each of the sets (a,a), $(a,a]$ and $[a,a)$ should be interpreted as the empty set $\emptyset$. This probability measure is a special case of the Lebesgue measure, which assigns to each interval its length.

If you are only interested in making probability statements about the sets in the algebra (25), then your are done. However, although the algebra (25) contains a large number of sets, we cannot yet make probability statements involving arbitrary Borel sets in [0,1], because not all the Borel sets in [0,1] are included in (25). In particular, for a countable sequence of sets $A_j \in \mathscr{F}_0$ the probability $P(\cup_{j=1}^{\infty}A_j)$ is not always defined, because there is no guarantee that $\cup_{j=1}^{\infty}A_j \in \mathscr{F}_0$. Therefore, if you want to make probability statements about arbitrary Borel set in [0,1], you need to extend the probability measure P on $\mathscr{F}_0$ to a probability measure defined on the Borel sets in [0,1]. The standard approach to do this is to use the *outer measure*:

6.2 Outer measure

Any subset A of [0,1] can always be completely covered by a finite or countably infinite union of sets in the algebra $\mathscr{F}_0$: $A \subset \cup_{j=1}^{\infty}A_j$, where $A_j \in \mathscr{F}_0$, hence the "probability" of A is bounded from above by $\sum_{j=1}^{\infty}P(A_j)$. The smallest upper bound is called the *outer measure*:

DEFINITION 7: *Let $\mathscr{F}_0$ be an algebra of subsets of Ω. The outer measure of an arbitrary subset A of Ω is*:

$$P^*(A) = \inf_{A\subset\cup_{j=1}^{\infty}A_j,\ A_j\in\mathscr{F}_0} \sum_{j=1}^{\infty}P(A_j). \tag{26}$$

Clearly, if $A \in \mathscr{F}_0$ then $P^*(A) = P(A)$. (*Exercise*: Why?) . Moreover, it follows from Theorem 4 that the sets $A_j \in \mathscr{F}_0$ can be chosen disjoint. The question now arises for which other subsets of Ω the outer measure is a probability measure. Note that the conditions (8) and (9) are satisfied for the outer measure P^* (*Exercise*: Why?), but in general condition (10) does not hold for arbitrary sets. See for example Royden (1968, pp. 63-64). Nevertheless, it is possible to extend the outer measure to a probability measure on a σ-algebra $\mathscr{F}$ containing $\mathscr{F}_0$:

THEOREM 9: *Let P be a probability measure on* $\{\Omega, \mathscr{F}_0\}$, *where* $\mathscr{F}_0$ *is an algebra, and let* $\mathscr{F} = \sigma(\mathscr{F}_0)$ *be the smallest* $\sigma-$*algebra containing the algebra* $\mathscr{F}_0$. *Then the outer measure* P^* *is a* ***unique*** *probability measure on* $\{\Omega, \mathscr{F}\}$ *which coincides with P on* $\mathscr{F}_0$.

Partial proof: See Appendix D.

Consequently, for the statistical experiment under review there exists a $\sigma-$algebra $\mathscr{F}$ of subsets of $\Omega = [0,1]$, containing the algebra $\mathscr{F}_0$ defined in (25), for which the outer measure $P^*: \mathscr{F} \to [0,1]$ is a unique probability measure. This probability measure assigns in this case to each interval in [0,1] its length as probability. It is called the *uniform* probability measure.

It is not hard to verify that the $\sigma-$algebra $\mathscr{F}$ involved contains all the Borel subsets of [0,1]:

$$\{[0,1] \cap B, \textit{ for all Borel sets } B\} \subset \mathscr{F}. \tag{27}$$

(*Exercise*: Why?) This collection of Borel subsets of [0,1] is usually denoted by $[0,1] \cap \mathscr{B}$, and is a $\sigma-$algebra itself (*Exercise*: Why?). Therefore, we could also describe the probability space of this statistical experiment by the probability space $\{[0,1], [0,1] \cap \mathscr{B}, P^*\}$, where P^* is the same as before. Moreover, defining the probability measure μ on $\mathscr{B}$ as:

$$\mu(B) = P^*([0,1] \cap B), \tag{28}$$

we could describe this statistical experiment also by the probability space $\{\mathbb{R}, \mathscr{B}, \mu\}$, where in particular

$$\mu((-\infty,x]) = 0 \textit{ if } x \leq 0,\ \mu((-\infty,x]) = x \textit{ if } 0 < x \leq 1,\ \mu((-\infty,x]) = 1 \textit{ if } x > 1, \tag{29}$$

and more generally for intervals with endpoints $a < b$,

$$\mu((a,b)) = \mu([a,b]) = \mu([a,b)) = \mu((a,b]) = \mu((-\infty,b]) - \mu((-\infty,a]), \tag{30}$$

whereas for all other Borel sets B,

$$\mu(B) = \inf_{B \subset \cup_{j=1}^{\infty}(a_j,b_j)} \sum_{j=1}^{\infty}\mu((a_j,b_j)). \tag{31}$$

7. *Lebesgue measure and Lebesgue integral*

7.1 *Lebesgue measure*

Along similar lines as in the construction of the uniform probability measure we can define the Lebesgue measure, as follows. Consider a function λ which assigns to each open interval (a,b) its length:

$$\lambda((a,b)) = b - a, \tag{32}$$

and define for all other Borel sets B in $\mathbb{R}$,

$$\lambda(B) = \inf_{B \subset \cup_{j=1}^{\infty}(a_j,b_j)} \sum_{j=1}^{\infty}\lambda((a_j,b_j)) = \inf_{B \subset \cup_{j=1}^{\infty}(a_j,b_j)} \sum_{j=1}^{\infty}(b_j - a_j). \tag{33}$$

This function λ is called the Lebesgue measure on $\mathbb{R}$, which measures the total "length" of a Borel set, where the measurement is taken from the outside.

Similarly, let now

$$\lambda\left(\times_{i=1}^{k}(a_i,b_i)\right) = \Pi_{i=1}^{k}(b_i - a_i). \tag{34}$$

and define for all other Borel sets B in $\mathbb{R}^k$,

$$\lambda(B) = \inf_{B \subset \cup_{j=1}^{\infty}\{\times_{i=1}^{k}(a_{i,j},b_{i,j})\}} \sum_{j=1}^{\infty}\lambda\left(\times_{i=1}^{k}(a_{i,j},b_{i,j})\right) = \inf_{B \subset \cup_{j=1}^{\infty}\{\times_{i=1}^{k}(a_{i,j},b_{i,j})\}} \sum_{j=1}^{\infty}\left\{\Pi_{i=1}^{k}(b_{i,j} - a_{i,j})\right\}. \tag{35}$$

This is the Lebesgue measure on $\mathbb{R}^k$, which measures the area (in the case $k = 2$) or the volume (in the case $k \geq 3$) of a Borel set in $\mathbb{R}^k$, where again the measurement is taken from the outside.

Note that in general Lebesgue measures are not probability measures, because the Lebesgue measure can be infinite. In particular, $\lambda(\mathbb{R}^k) = \infty$. However, if confined to a set with Lebesgue measure 1 it becomes the uniform probability measure. More generally, for any Borel set $A \in \mathbb{R}^k$ with positive and finite Lebesgue measure, $\mu(B) = \lambda(A \cap B)/\lambda(A)$ is the uniform probability measure on $\mathscr{B}^k \cap A$.

7.2 Lebesgue integral

The Lebesgue measure gives rise to a generalization of the Riemann integral. Recall that the Riemann integral of a non-negative function $f(x)$ over a finite interval $(a,b]$ is defined as

$$\int_a^b f(x)dx = \sup\sum_{m=1}^{n}\left(\inf_{x\in I_m} f(x)\right)\lambda(I_m) \tag{36}$$

where the I_m are intervals forming a finite partition of $(a,b]$, i.e., they are disjoint, and their union is $(a,b]$: $(a,b] = \cup_{m=1}^{n} I_m$, $\lambda(I_m)$ is the length of I_m, hence $\lambda(I_m)$ is the Lebesgue measure of I_m, and the supremum is taken over all finite partitions of $(a,b]$. Mimicking this definition, the Lebesgue integral of a non-negative function $f(x)$ over a Borel set A can be defined as

$$\int_A f(x)dx = \sup\sum_{m=1}^{n}\left(\inf_{x\in B_m} f(x)\right)\lambda(B_m) \tag{37}$$

where now the B_m 's are Borel sets forming a finite partition of $A,$ and the supremum is taken over all such partitions.

If the function $f(x)$ is not non-negative, we can always write it as the difference of two non-negative functions:

$$f(x) = f_+(x) - f_-(x),\ where\ f_+(x) = \max[0,f(x)],\ f_-(x) = \max[0,-f(x)].$$

Then the Lebesgue integral over a Borel set A is defined as

$$\int_A f(x)dx = \int_A f_+(x)dx - \int_A f_-(x)dx, \tag{38}$$

provided that at least one of the right hand side integrals is finite.

Finally, note that if A is an interval and $f(x)$ is Riemann integrable over A, then the

Riemann integral and the Lebesgue integral coincide.

8. Random variables and their distributions

8.1 Random variables and vectors

Loosely speaking, a random variable is a numerical translation of the outcomes of a statistical experiment. For example, flip a fair coin once. Then the sample space is $\Omega = \{H,T\}$, where *H* stands for Head, and *T* stands for Tail. The σ-algebra of events is $\mathscr{F} = \{\Omega, \varnothing, \{H\}, \{T\}\}$, and the corresponding probability measure is defined by $P(\{H\}) = P(\{T\}) = 1/2$. Now define the function $X(\omega) = 1$ if $\omega = H$, $X(\omega) = 0$ if $\omega = T$. Then X is a random variable which takes the value 1 with probability ½ and the value 0 with probability ½:

$$\begin{aligned} P(X = 1) &\overset{(short\text{-}hand\ notation)}{=} P(\{\omega\in\Omega: X(\omega) = 1\} = P(\{H\}) = 1/2, \\ P(X = 0) &\overset{(short\text{-}hand\ notation)}{=} P(\{\omega\in\Omega: X(\omega) = 0\} = P(\{T\}) = 1/2. \end{aligned} \tag{39}$$

Moreover, for an arbitrary Borel set *B* we have

$$P(X \in B) = P(\{\omega\in\Omega: X(\omega) \in B\}) \begin{cases} = P(\{H\}) & = 1/2 \ \ if\ 1 \in B\ \ and\ \ 0 \notin B, \\ = P(\{T\}) & = 1/2 \ \ if\ 1 \notin B\ \ and\ \ 0 \in B, \\ = P(\{H,T\}) & = \ 1 \ \ \ if\ 1 \in B\ \ and\ \ 0 \in B, \\ = P(\varnothing) & = \ 0 \ \ \ if\ 1 \notin B\ \ and\ \ 0 \notin B, \end{cases} \tag{40}$$

where again $P(X \in B)$ is a short-hand notation[9] for $P(\{\omega\in\Omega: X(\omega) \in B\})$.

In this particular case the set $\{\omega\in\Omega: X(\omega) \in B\}$ is automatically equal to one of the elements of $\mathscr{F}$, and therefore the probability $P(X \in B) = P(\{\omega\in\Omega: X(\omega) \in B\})$ is well-defined. In general, however, we need to confine the mappings $X: \Omega \rightarrow \mathbb{R}$ to those for which we

[9] In the sequel we will denote the probability of an event involving random variables or vectors X as *P*("expression involving *X*"), without referring to the corresponding set in $\mathscr{F}$. For example, for random variables *X* and *Y* defined on a common probability space $\{\Omega, \mathscr{F}, P\}$ the short-hand notation $P(X > Y)$ should be interpreted as $P(\{\omega\in\Omega: X(\omega) > Y(\omega)\})$.

can make probability statements about events of the type $\{\omega\in\Omega: X(\omega) \in B\}$, where B is an arbitrary Borel set, which is only possible if these sets are members of $\mathscr{F}$:

DEFINITION 8: *Let $\{\Omega,\mathscr{F},P\}$ be a probability space. A mapping X: $\Omega \to \mathbb{R}$ is called a* ***random variable*** *defined on $\{\Omega,\mathscr{F},P\}$ if X is* ***measurable*** *$\mathscr{F}$, which means that for every Borel set B, $\{\omega\in\Omega: X(\omega) \in B\} \in \mathscr{F}$. Similarly, a mapping X: $\Omega \to \mathbb{R}^k$ is called a k-dimensional* ***random vector*** *defined on $\{\Omega,\mathscr{F},P\}$ if X is* ***measurable*** *$\mathscr{F}$, in the sense that for every Borel set B in $\mathscr{B}^k$, $\{\omega\in\Omega: X(\omega) \in B\} \in \mathscr{F}$.*

In verifying that a real function$X$: $\Omega \to \mathbb{R}$ is measurable$\mathscr{F}$, it is not necessary to verify that for *all* Borel sets B, $\{\omega\in\Omega: X(\omega) \in B\} \in \mathscr{F}$, but only that this property holds for Borel sets of the type $(-\infty,x]$:

THEOREM 10: *A mapping X: $\Omega \to \mathbb{R}$ is measurable$\mathscr{F}$ (hence X is a random variable) if and only if for all $x \in \mathbb{R}$ the sets $\{\omega\in\Omega: X(\omega) \le x\}$ are members of $\mathscr{F}$. Similarly, a mapping X: $\Omega \to \mathbb{R}^k$ is measurable$\mathscr{F}$ (hence X is a random vector of dimension k) if and only if for all $x = (x_1,.....,x_k)^T \in \mathbb{R}^k$ the sets*

$$\cap_{j=1}^{k}\{\omega\in\Omega: X_j(\omega) \le x_j\} = \{\omega\in\Omega: X(\omega) \in \times_{j=1}^{k}(-\infty,x_j]\}$$

are members of $\mathscr{F}$, where the X_j's are the components of X.

Proof: Consider the case $k = 1$. Suppose that $\{\omega\in\Omega: X(\omega) \in (-\infty,x]\} \in \mathscr{F}$, $\forall x \in \mathbb{R}$. Let $\mathscr{D}$ be the collection of all Borel sets B for which $\{\omega\in\Omega: X(\omega) \in B\} \in \mathscr{F}$. Then $\mathscr{D} \subset \mathscr{B}$, and $\mathscr{D}$ contains the collection of half-open intervals $(-\infty,x]$, $x \in \mathbb{R}$. If $\mathscr{D}$ is a $\sigma-$algebra itself, it is a $\sigma-$algebra containing the half-open intervals. But $\mathscr{B}$ is the smallest $\sigma-$algebra containing the half-open intervals (see Theorem 6), so that then $\mathscr{B} \subset \mathscr{D}$, hence $\mathscr{D} = \mathscr{B}$. Therefore, it suffices to prove that $\mathscr{D}$ is a $\sigma-$algebra:

(a) Let $B \in \mathscr{D}$. Then $\{\omega\in\Omega: X(\omega) \in B\} \in \mathscr{F}$, hence

$$\sim\{\omega\in\Omega: X(\omega) \in B\} = \{\omega\in\Omega: X(\omega) \in \tilde{B}\} \in \mathscr{F}$$

and thus $\tilde{B} \in \mathscr{D}$.

(b) Next, let $B_j \in \mathscr{D}$ for j = 1,2,.... Then $\{\omega\in\Omega: X(\omega) \in B_j\} \in \mathscr{F}$, hence

$$\cup_{j=1}^{\infty}\{\omega\in\Omega: X(\omega) \in B_j\} = \{\omega\in\Omega: X(\omega) \in \cup_{j=1}^{\infty}B_j\} \in \mathscr{F}$$

and thus $\cup_{j=1}^{\infty}B_j \in \mathscr{D}$.

The proof of the case $k > 1$ is similar. Q.E.D.[10]

The sets $\{\omega\in\Omega: X(\omega) \in B\}$ are usually denoted by $X^{-1}(B)$:

$$X^{-1}(B) \overset{def.}{=} \{\omega\in\Omega: X(\omega) \in B\}. \tag{41}$$

The collection $\mathscr{F}_X = \{X^{-1}(B), \forall B \in \mathscr{B}\}$ is a $\sigma-$algebra itself (*Exercise*: Why?), and is called the $\sigma-$algebra *generated* by the random variable X. More generally:

DEFINITION 9: *Let X be a random variable* (k=1) *or a random vector* ($k > 1$). *The* $\sigma-$*algebra* $\mathscr{F}_X = \{X^{-1}(B), \forall B \in \mathscr{B}^k\}$ *is called the* $\sigma-$*algebra* ***generated*** *by X.*

In the coin tossing case, the mapping X is one-to-one, and therefore in that case $\mathscr{F}_X$ is the same as $\mathscr{F}$, but in general $\mathscr{F}_X$ will be smaller than $\mathscr{F}$. For example, roll a dice, and let $X = 1$ if the outcome is even, and $X = 0$ if the outcome is odd. Then

$$\mathscr{F}_X = \{\{1,2,3,4,5,6\}, \{2,4,6\}, \{1,3,5\}, \varnothing\},$$

whereas $\mathscr{F}$ in this case consists of *all* subsets of $\Omega = \{1,2,3,4,5,6\}$.

Given a k dimensional random vector X, or a random variable X (the case k=1), define for arbitrary Borel sets $B \in \mathscr{B}^k$:

$$\mu_X(B) = P\left(X^{-1}(B)\right) = P\left(\{\omega\in\Omega: X(\omega) \in B\}\right). \tag{42}$$

Then $\mu_X(\cdot)$ is a probability measure on $\{\mathbb{R}^k, \mathscr{B}^k\}$:

(a) for all $B \in \mathscr{B}^k$, $\mu_X(B) \geq 0$,

(b) $\mu_X(\mathbb{R}^k) = 1$,

[10] See also Appendix C.

(c) for all disjoint $B_j \in \mathscr{B}^k$, $\mu_X\left(\cup_{j=1}^{\infty} B_j\right) = \Sigma_{j=1}^{\infty}\mu_X(B_j)$.

Thus, the random variable X maps the probability space $\{\Omega, \mathscr{F}, P\}$ into a new probability space, $\{\mathbb{R}, \mathscr{B}, \mu_X\}$, which in its turn is mapped back by X^{-1} into the (possibly smaller) probability space $\{\Omega, \mathscr{F}_X, P\}$. Similarly for random vectors.

DEFINITION 10: *The probability measure* $\mu_X(\cdot)$ *defined by* (42) *is called the probability measure* ***induced*** *by X.*

8.2 Distribution functions

For Borel sets of the type $(-\infty, x]$, or $\times_{j=1}^{k}(-\infty, x_j]$ in the multivariate case, the value of the induced probability measure μ_X is called the distribution function:

DEFINITION 11: *Let X be a random variable* (k=1) *or a random vector* (k>1) *with induced probability measure* μ_X. *The function* $F(x) = \mu_X(\times_{j=1}^{k}(-\infty, x_j])$, $x = (x_1, \ldots., x_k)^T$ $\in \mathbb{R}^k$, *is called the* ***distribution function*** *of X.*

It follows from these definitions, and Theorem 8 that

THEOREM 11: *A distribution function of a random* ***variable*** *is always right continuous*:

$$\forall x \in \mathbb{R},\ \lim_{\delta \downarrow 0} F(x + \delta) = F(x), \tag{43}$$

and monotonic non-decreasing: $F(x_1) \leq F(x_2)$ *if* $x_1 < x_2$, *with*

$$\lim_{x \downarrow -\infty} F(x) = 0, \quad \lim_{x \uparrow \infty} F(x) = 1. \tag{44}$$

Proof: Exercise.

However, a distribution function is not always left continuous. As a counter example, consider the distribution function of the Binomial (*n,p*) distribution in section 2.2. Recall that the corresponding probability space consists of sample space $\Omega = \{0,1,2,...,n\}$, the σ-algebra $\mathscr{F}$

of all subsets of Ω, and probability measure $P(\{k\})$ defined by (14) . The random variable X involved is defined as $X(k) = k$, with distribution function

$$\begin{aligned} F(x) &= 0 \ \mathit{for} \ x < 0, \\ F(x) &= \Sigma_{k \le x} P(\{k\}) \ \mathit{for} \ x \in [0,n], \\ F(x) &= 1 \ \mathit{for} \ x > n, \end{aligned} \tag{45}$$

Now let for example $x = 1$. Then for $0 < \delta < 1$, $F(1 - \delta) = F(0)$, and $F(1 + \delta) = F(1)$, hence $\lim_{\delta \downarrow 0} F(1 + \delta) = F(1)$, but $\lim_{\delta \downarrow 0} F(1 - \delta) = F(0) < F(1)$.

The left limit of a distribution function F in x is usually denoted by $F(x-)$:

$$F(x-) \overset{def.}{=} \lim_{\delta \downarrow 0} F(x - \delta). \tag{46}$$

Thus if x is a continuity point then $F(x\text{-}) = F(x)$, and if x is a discontinuity point then $F(x\text{-}) < F(x)$.

The Binomial distribution involved is an example of a *discrete* distribution. The uniform distribution on [0,1] derived in section 5 is an example of a *continuous* distribution, with distribution function

$$\begin{aligned} F(x) &= 0 \ \mathit{for} \ x < 0, \\ F(x) &= x \ \mathit{for} \ x \in [0,1], \\ F(x) &= 1 \ \mathit{for} \ x > 1. \end{aligned} \tag{47}$$

In the case of the Binomial distribution (14) the number of discontinuity points of F is finite, and in the case of the Poisson distribution (16) the number of discontinuity points of F is countable infinite. In general we have:

THEOREM 12: *The set of discontinuity points of a distribution function of a random* ***variable*** *is countable.*

Proof: Let D be the set of all discontinuity points of the distribution function $F(x)$. Every point x in D is associated with an non-empty open interval $(F(x\text{-}),F(x)) = (a,b)$, *say,* which is contained in [0,1]. For each of these open intervals (a,b) there exists a rational number q such

$a < q < b$, hence the number of open intervals (a,b) involved is countable, because the rational numbers are countable. Therefore, D is countable. Q.E.D.

The results of Theorems 11-12 only hold for distribution functions of random *variables*, though. It is possible to generalize these results to distribution functions of random vectors, but this generalization is far from trivial and therefore omitted.

As follows from Definition 11, a distribution function of a random variable or vector X is completely determined by the corresponding induced probability measure $\mu_X(\cdot)$. But what about the other way around, i.e., given a distribution function $F(x)$, is the corresponding induced probability measure $\mu_X(\cdot)$ unique? The answer is yes, but we prove the result only for the univariate case:

THEOREM 13: *Given the distribution function F of a random vector $X \in \mathbb{R}^k$, there exists a* ***unique*** *probability measure μ on $\{\mathbb{R}^k, \mathscr{B}^k\}$ such that for $x = (x_1,....,x_k)^T \in \mathbb{R}^k$, $F(x) =$ $\mu\left(\times_{i=1}^{k}(-\infty,x_i]\right)$.*

Proof: Let $k = 1$ and let $\Im_0$ be the collection of all intervals of the type

$$(a,b),[a,b],(a,b],[a,b),(-\infty,a),(\infty,a],(b,\infty),[b,\infty),\ a\leq b \in \mathbb{R}, \qquad (48)$$

together with their finite unions, where $[a,a]$ is the singleton $\{a\}$, and (a,a), $(a,a]$ and $[a,a)$ should be interpreted as the empty set $\emptyset$. Then each set in $\Im_0$ can be written as a finite union of *disjoint* sets of the type (48) (Compare (25)), hence $\Im_0$ is an algebra. Define for $-\infty < a < b < \infty$,

$$\begin{aligned}
&\mu((a,a)) = \mu((a,a]) = \mu([a,a)) = \mu(\emptyset) = 0\\
&\mu(\{a\}) = F(a) - \lim_{\delta\downarrow 0}F(a-\delta), \quad \mu((a,b]) = F(b) - F(a)\\
&\mu([a,b)) = \mu((a,b]) - \mu(\{b\}) + \mu(\{a\}), \quad \mu([a,b]) = \mu((a,b]) + \mu(\{a\})\\
&\mu((a,b)) = \mu((a,b]) - \mu(\{b\}), \quad \mu((-\infty,a]) = F(a)\\
&\mu((-\infty,a]) = F(a) - \mu(\{a\}), \quad \mu((b,\infty)) = 1 - F(b)\\
&\mu([b,\infty)) = \mu((b,\infty)) + \mu(\{b\})
\end{aligned} \tag{49}$$

and let for *disjoint* sets $A_1,.......,A_n$ of the type (48), $\mu(\cup_{j=1}^{n}A_j) = \Sigma_{j=1}^{n}\mu(A_j)$. Then the distribution function F defines a probability measure μ on $\mathfrak{I}_0$, and this probability measure μ coincides on $\mathfrak{I}_0$ with the induced probability measure μ_X. It follows now from Theorem 9 that there exists a σ-algebra $\mathfrak{I}$ containing $\mathfrak{I}_0$ for which the same applies. This σ-algebra $\mathfrak{I}$ may be chosen equal to the σ-algebra $\mathcal{B}$ of Borel sets. Q.E.D.

The importance of this result is that there is a one-to-one relationship between the distribution function F of a random variable or vector X and the induced probability measure μ_X. Therefore, the distribution function contains all the information about μ_X.

DEFINITION 12: *A distribution function F on $\mathbb{R}^k$ and its associated probability measure μ on $\{\mathbb{R}^k, \mathcal{B}^k\}$ are called **absolutely continuous with respect to Lebesgue measure** if for every Borel set B in $\mathbb{R}^k$ with zero Lebesgue measure, $\mu(B) = 0$.*

We will need this concept in the next section.

9. Density functions

An important concept is that of a density function. Density functions are usually associated to differentiable distribution functions:

DEFINITION 13: *The distribution of a random variable X is called **absolutely continuous** if there exists a non-negative integrable function f, called the **density function** of X, such that the distribution function F of X can be written as the* (*Lebesgue*) *integral* $F(x) =$

$\int_{-\infty}^{x} f(u)du$. *Similarly, the distribution of a random vector* $X \in \mathbb{R}^k$ *is called absolutely continuous if there exists a non-negative integrable function f on* $\mathbb{R}^k$ *, called the joint density, such that the distribution function F of X can be written as the integral*

$$F(x) = \int_{-\infty}^{x_1} ..... \int_{-\infty}^{x_k} f(u_1,...,u_k)du_1....du_k,$$

where $x = (x_1, ..., x_k)^T$.

Thus, in the case $F(x) = \int_{-\infty}^{x} f(u)du$ the density function $f(x)$ is the derivative of $F(x)$: $f(x) = F'(x)$, and in the multivariate case $F(x_1,...,x_k) = \int_{-\infty}^{x_1} ..... \int_{-\infty}^{x_k} f(u_1,...,u_k)du_1....du_k$ the joint density is $f(x_1,...,x_k) = (\partial/\partial x_1)......(\partial/\partial x_k)F(x_1,...,x_k)$.

The reason for calling the distribution functions in Definition 13 **absolutely continuous** is that in this case the distributions involved are absolutely continuous with respect to Lebesgue measure. See Definition 12. To see this, consider the case $F(x) = \int_{-\infty}^{x} f(u)du$, and verify (*Exercise*) that the corresponding probability measure μ is:

$$\mu(B) = \int_B f(x)dx, \tag{50}$$

where the integral is now the Lebesgue integral over a Borel set *B*. Since the Lebesgue integral over a Borel set with zero Lebesgue measure is zero (*Exercise*), it follows that μ(*B*) = 0 if the Lebesgue measure of *B* is zero.

For example the uniform distribution (47) is absolutely continuous, because we can write (47) as $F(x) = \int_{-\infty}^{x} f(u)du$, with density $f(u) = 1$ for $0 < u < 1$ and zero elsewhere. Note that in this case $F(x)$ is not differentiable in 0 and 1, but that does not matter, as long as the set of points for which the distribution function is not differentiable has zero Lebesgue measure. Moreover, a density of a random variable always integrates to 1, because $1 = \lim_{x\to\infty} F(x) = \int_{-\infty}^{\infty} f(u)du$. Similarly for random vectors $X \in \mathbb{R}^k$: $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ..... \int_{-\infty}^{\infty} f(u_1,...,u_k)du_1....du_k = 1$.

The concept of a density function can be generalized to other distributions than only absolute continuous distributions, in the sense that $F(x)$ can be written as an integral of a density $f(x)$, although no longer as $F(x) = \int_{-\infty}^{x} f(u)du$. For example, consider the Poisson(λ) distribution (16). The distribution function involved is

$$F(x) = \sum_{k=0}^{[x]} e^{-\lambda}\lambda^k / k! \ \ if\ x \ge 0, \quad F(x) = 0 \ \ if\ x < 0, \tag{51}$$

where $[x]$ denotes the largest integer $\le x$. Define

$$f(x) = \sum_{k=0}^{\infty} \left(e^{-\lambda}\lambda^k / k!\right).I\left(x \in [k,k+1)\right). \tag{52}$$

Then for $x \ge 0$,

$$\int_{-\infty}^{[x]+1} f(u)du = \int_{0}^{[x]+1} \sum_{k=0}^{\infty} \left(e^{-\lambda}\lambda^k / k!\right).I\left(u \in [k,k+1)\right)du$$
$$= \sum_{m=0}^{[x]} \int_{m}^{m+1} \sum_{k=0}^{\infty} \left(e^{-\lambda}\lambda^k / k!\right).I\left(u \in [k,k+1)\right)du = \sum_{k=0}^{[x]} e^{-\lambda}\lambda^k / k! = F(x) \tag{53}$$

Thus also in this case $F(x)$ can be written as in integral, but now as $F(x) = \int_{-\infty}^{[x]+1} f(u)du$ instead of $F(x) = \int_{-\infty}^{x} f(u)du$. Nevertheless, the function (52) is often referred to as the density of the Poisson(λ) distribution.

Note that in this case $F(x)$ is not absolutely continuous with respect to Lebesgue measure. For example, let $B = \{0,1,2,3,........\}$. Clearly, the Lebesgue measure of B is zero, but $\mu(B) = 1$, where μ is the probability measure associated to $F(x)$.

10. Conditional probability, Bayes' rule, and independence

10.1 Conditional probability

Consider statistical experiment with probability space $\{\Omega,\mathscr{F},P\}$, and suppose that it is known that the outcome of this experiment is contained in a set B with $P(B) > 0$. What is the probability of an event A, given that the outcome of the experiment is contained in B? For example, roll a dice. Then $\Omega = \{1,2,3,4,5,6\}$, $\mathscr{F}$ is the σ-algebra of all subsets of Ω, and $P(\{\omega\}) = 1/6$ for $\omega = 1,2,3,4,5,6$. Let B be the event: "the outcome is even": $B = \{2,4,6\}$, and let $A = \{1,2,3\}$. If we know that the outcome is even, then we know that the outcomes $\{1,3\}$ in A will not occur: if the outcome in contained in A, it is contained in $A\cap B = \{2\}$. Knowing that the outcome is either 2,4, or 6, the probability that the outcome is contained in A is therefore $1/3 = P(A\cap B)/P(B)$. This is the conditional probability of A, given B, denoted by $P(A|B)$. If it is revealed that the outcome of a statistical experiment is contained in a particular set B, then the sample space Ω is reduced to B, because we then know that the outcomes in the complement of B will not occur, the σ-algebra $\mathscr{F}$ is reduced to $\mathscr{F}\cap B$, the collection of all intersections of the sets in $\mathscr{F}$ with B: $\mathscr{F}\cap B = \{A\cap B, \mathrm{A}\in\mathscr{F}\}$ (*Exercise*: Is this a σ-algebra?), and the probability measure involved becomes $P(A|\mathrm{B}) = P(A\cap B)/P(B)$, hence the probability space becomes $\{B, \mathscr{F}\cap B, P(\cdot|B)\}$. See Exercise 20 below.

10.2 Bayes' rule

Let A and B be sets in $\mathscr{F}$. Since the sets A and $\tilde{A}$ form a partition of the sample space Ω, we have

$$B = (B \cap A) \cup (B \cap \tilde{A}),$$

hence

$$\mathrm{P}(B) = P(B \cap A) + P(B \cap \tilde{A}) = P(B|A)P(A) + P(B|\tilde{A})P(\tilde{A}).$$

Moreover,

$$\mathrm{P}(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}.$$

Combining these two results now yields Bayes' rule:

$$\mathrm{P}(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\tilde{A})P(\tilde{A})}.$$

Thus, Bayes' rule enables us to compute the conditional probability $P(A|B)$ if $P(A)$ and the conditional probabilities $P(B|A)$ and $P(B|\tilde{A})$ are given.

More generally, if A_j, $j = 1,2,.....n$ ($\leq \infty$) is a partition of the sample space Ω, i.e., the A_j's are disjoint sets in $\mathscr{F}$ such that $\Omega = \cup_{j=1}^{n} A_j$, then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}.$$

Bayes' rule plays an important role in a special branch of statistics [and econometrics], called Bayesian statistics [econometrics].

10.3. Independence

If $P(A|B) = P(A)$, then knowing that the outcome is in B does not give us any information about A. In that case the events A and B are called *independent*. For example, if I tell you that the outcome of the dice experiment is contained in the set $\{1,2,3,4,5,6\} = \Omega$, then you know nothing about the outcome: $P(A|\Omega) = P(A\cap\Omega)/P(\Omega) = P(A)$, hence Ω is independent of any other event A.

Note that $P(A|B) = P(A)$ is equivalent to $P(A\cap B) = P(A)P(B)$. Thus,

DEFINITION 14: *Sets A and B in $\mathscr{F}$ are (pairwise) independent if $P(A\cap B) = P(A)P(B)$.*

If events A and B are independent, and events B and C are independent, are the events A and C independent? The answer is: not necessarily. In order to give a counter example, observe that if A and B are independent, then so are $\tilde{A}$ and B, A and $\tilde{B}$, and $\tilde{A}$ and $\tilde{B}$, because

$$P(\tilde{A}\cap B) = P(B) - P(A\cap B) = P(B) - P(A)P(B) = (1-P(A))P(B) = P(\tilde{A})P(B),$$

and similarly,

$$P(A\cap\tilde{B}) = P(A)P(\tilde{B}) \text{ and } P(\tilde{A}\cap\tilde{B}) = P(\tilde{A})P(\tilde{B}).$$

Now if $C = \tilde{A}$ and $0 < P(A) < 1$, then B and $C = \tilde{A}$ are independent if A and B are independent, but

$$P(A\cap C) = P(A\cap\tilde{A}) = P(\emptyset) = 0,$$

whereas

$$P(A)P(C) = P(A)P(\tilde{A}) = P(A)(1-P(A)) \neq 0.$$

Thus, for more than two events we need a stronger condition for independence than pairwise independence, namely:

DEFINITION 15: *A sequence A_j of sets in $\mathscr{F}$ is independent if for **every** sub-sequence A_{j_i}, $i = 1,2,..,n$, $P(\cap_{i=1}^{n} A_{j_i}) = \prod_{i=1}^{n} P(A_{j_i})$.*

By requiring that the latter holds for all sub-sequences rather than $P(\bigcap_{i=1}^{\infty}A_i) = \prod_{i=1}^{\infty}P(A_i)$, we avoid the problem that a sequence of events would be called independent if one of the events is the empty set.

The independence of a pair or sequence of random variables or vectors can now be defined as follows.

DEFINITION 16: *Let X_j be a sequence of random variables or vectors defined on a common probability space $\{\Omega,\mathscr{F},P\}$. X_1 and X_2 are pairwise independent if for **all** Borel sets B_1, B_2, the sets $A_1 = \{\omega\in\Omega\text{: } X_1(\omega) \in B_1\}$ and $A_2 = \{\omega\in\Omega\text{: } X_2(\omega) \in B_2\}$ are independent. The sequence X_j is independent if for **all** Borel sets B_j the sets $A_j = \{\omega\in\Omega\text{: } X_j(\omega) \in B_j\}$ are independent.*

As we have seen before, the collection $\mathscr{F}_j = \{\{\omega\in\Omega\text{: } X_j(\omega) \in B\},\ B \in \mathfrak{B}\} = \{X_j^{-1}(B),\ B \in \mathfrak{B}\}$ is a sub σ-algebra of $\mathscr{F}$. Therefore, Definition 16 also reads:
The sequence of random variables X_j is independent if for arbitrary $A_j \in \mathscr{F}_j$ the sequence of sets A_j is independent (according to Definition 15).

Independence usually follows from the setup of a statistical experiment. For example, draw randomly *with* replacement n balls from a bowl containing R red balls and $N-R$ white balls, and let $X_j = 1$ if the j-th draw is a red ball, and $X_j = 0$ if the j-th draw is a white ball. Then $X_1,...,X_n$ are independent (and $X_1+...+X_n$ has the Binomial (n,p) distribution, with $p = R/N$). However, if we would draw these balls without replacement, then $X_1,...,X_n$ are not independent.

For a sequence of random variables X_j it suffices to verify the condition in Definition 16 for Borel sets B_j of the type $(-\infty,x_j]$, $x_j \in \mathbb{R}$, only:

THEOREM 14 : *Let $X_1,...,X_n$ be random variables, and denote for $x \in \mathbb{R}$ and $j = 1,....,n$, $A_j(x) = \{\omega\in\Omega\text{: } X_j(\omega) \le x\}$. Then $X_1,...,X_n$ are independent if and only if for arbitrary $(x_1,.....,x_n)^T \in \mathbb{R}^n$ the sets $A_1(x_1),......,A_n(x_n)$ are independent.*

The proof of Theorem 14 is complicated, and therefore omitted.[11]

It follows now from Theorem 15 that:

THEOREM 15: *The random variables $X_1,...,X_n$ are independent if and only if the joint distribution function $F(x)$ of $X = (X_1,...,X_n)^T$ can be written as the product of the distribution functions $F_j(x_j)$ of the X_j 's, i.e., $F(x) = \Pi_{j=1}^n F_j(x_j)$, where $x = (x_1,....,x_n)^T$.*

The latter distribution functions $F_j(x_j)$ are called the *marginal* distribution functions. Moreover, it follows straightforwardly from Theorem 15 that if the joint distribution of $X = (X_1,....,X_n)^T$ is absolutely continuous with joint density function $f(x)$, then $X_1,...,X_n$ are independent if and only if $f(x)$ can be written as the product of the density functions $f_j(x_j)$ of the X_j 's:

$$f(x) = \Pi_{j=1}^n f_j(x_j), \text{ where } x = (x_1,....,x_n)^T.$$

The latter density functions are called the *marginal* density functions.

Exercises:

1. Prove (4).
2. Show that $p_1(r,n)$ in (12) decreases if we increase r.
3. Show that $p_2(r,n)$ in (13) increases with r .
4. Prove (20) by proving that $\ln[(1 - \mu/n)^n] = n\ln(1 - \mu/n) \rightarrow -\mu$ *for* $n \rightarrow \infty$.
5. Let $\mathscr{F}_*$ be the collection of all subsets of $\Omega = (0,1]$ of the type $(a,b]$, where $a < b$ are

[11] Let $\mathscr{F}_j^0 = \{\Omega,\emptyset,X_j^{-1}((-\infty,x]),X_j^{-1}((y,\infty)),\forall\ x,y\in\mathbb{R}$, together with all finite unions and intersections of the latter two types of sets$\}$. Then $\mathscr{F}_j^0$ is an algebra such that for arbitrary $A_j \in \mathscr{F}_j^0$ the sequence of sets A_j is independent. This is not too hard to prove. Now $\mathscr{F}_j = \{X_j^{-1}(B),\ B \in \mathfrak{B}\}\}$ is the smallest σ-algebra containing $\mathscr{F}_j^0$, and is also the smallest monotone class containing $\mathscr{F}_j^0$. It can be shown (but this is the hard part), using the properties of monotone class (see Exercise 13 below), that for arbitrary $A_j \in \mathscr{F}_j$ the sequence of sets A_j is independent as well.

rational numbers in [0,1], together with their *finite* unions and the empty set $\varnothing$. Verify that $\mathscr{F}_*$ is an algebra.

6. Prove Theorem 2.

7. Prove Theorem 5.

8. Let $\Omega = (0,1]$, and let $\mathfrak{C}$ be the collection of all intervals of the type $(a,b]$ with $0 \le a < b \le 1$. We have shown that the smallest algebra containing the collection $\mathfrak{C}$ consists of the sets in $\mathfrak{C}$ together with the empty set $\varnothing$, and all finite unions of disjoint sets in $\mathfrak{C}$. Determine along the same lines $\sigma(\mathfrak{C})$, the smallest σ-algebra containing this collection $\mathfrak{C}$.

9. Show that $\sigma(\{[a,b]: \forall a \le b,\ a,b \in \mathbb{R}\}) = \mathscr{B}$.

10. Prove part (*g*) of Theorem 8.

11. Prove that $\mathscr{F}_0$ defined by (25) is an algebra.

12. Prove Theorem 11. *Hint*: Use Definition 12 and Theorem 8. Determine first which parts of Theorem 8 apply.

13. A collection $\mathscr{F}$ of subsets of a set Ω is called a *monotone class* if the following two conditions hold:

$$A_n \in \mathscr{F},\ A_n \subset A_{n+1},\ n = 1,2,3,\text{.....}\ \text{imply}\ \cup_{n=1}^{\infty} A_n \in \mathscr{F},$$

$$A_n \in \mathscr{F},\ A_n \supset A_{n+1},\ n = 1,2,3,\text{.....}\ \text{imply}\ \cap_{n=1}^{\infty} A_n \in \mathscr{F}.$$

Show that an algebra is a σ-algebra if and only if it is a monotone class.

14. A collection $\mathscr{F}_\lambda$ of subsets of a set Ω is called a $\lambda-$system if $A \in \mathscr{F}_\lambda$ implies $\tilde{A} \in \mathscr{F}_\lambda$, and for **disjoint** sets $A_j \in \mathscr{F}_\lambda$, $\cup_{j=1}^{\infty} A_j \in \mathscr{F}_\lambda$. A collection $\mathscr{F}_\pi$ of subsets of a set Ω is called a $\pi-$system if $A,B \in \mathscr{F}_\pi$ implies that $A \cap B \in \mathscr{F}_\pi$. Prove that if a $\lambda-$system is also a $\pi-$system, then it is a σ-algebra .

15. Let $\mathscr{F}$ be the smallest $\sigma-$algebra of subsets of $\mathbb{R}$ containing the (countable) collection of half-open intervals $(-\infty, q]$ with *rational* endpoints *q*. Prove that $\mathscr{F}$ contains all the Borel subsets of $\mathbb{R}$: $\mathscr{B} = \mathscr{F}$.

16. Consider the following subset of $\mathbb{R}^2$: $L = \{(x,y) \in \mathbb{R}^2 \text{: } y = x,\ 0 \le x \le 1\}$. Explain why *L* is a Borel set.

17. Consider the following subset of $\mathbb{R}^2$: $C = \{(x,y) \in \mathbb{R}^2 \text{: } x^2 + y^2 \le 1\}$. Explain why *C*

is a Borel set.

18. Let $F(x) = \int_{-\infty}^{x} f(u)du$ be an absolutely continuous distribution function. Prove that corresponding probability measure μ is given by the Lebesgue integral (50).

19. Prove that the Lebesgue integral over a Borel set with zero Lebesgue measure is zero.

20. Let $\{\Omega,\mathscr{F},P\}$ be a probability space, and let $B \in \mathscr{F}$ with $P(B) > 0$. Verify that $\{B,\mathscr{F}\cap B,P(\cdot|B)\}$ is a probability space.

21. Are disjoint sets in $\mathscr{F}$ independent?

22. (Application of Bayes' rule): Suppose that 1 out of 10,000 people suffer from a certain disease, say HIV+. Moreover, suppose that there exists a medical test for this disease which is 90% reliable: If you don't have the disease, the test will confirm that with probability 0.9, and the same if you do have the disease. If a randomly selected person is subjected to this test, and the test indicates that this person has the disease, what is the probability that this person actually has this disease? In other words, if you were this person, would you be scared or not?

23. Let A and B in $\mathscr{F}$ be pairwise independent. Prove that $\tilde{A}$ and B are independent (and therefore A and $\tilde{B}$ are independent and $\tilde{A}$ and $\tilde{B}$ are independent).

24. Draw randomly **without** replacement n balls from a bowl containing R red balls and $N-R$ white balls, and let $X_j = 1$ if the j-th draw is a red ball, and $X_j = 0$ if the j-th draw is a white ball. Show that $X_1,...,X_n$ are **not** independent.

APPENDIX A: Sets and set operations

In the main text I will assume that the reader is familiar with the basic set operations, notations, and results listed here.

1. The union $A \cup B$ of two sets A and B is the set of elements that belong either to A or B or both. The finite union $\cup_{j=1}^{n} A_j$ of sets $A_1,...,A_n$ is the set with the property that for each[12] $x \in \cup_{j=1}^{n} A_j$ there exists an index i, $1 \le i \le n$, for which $x \in A_i$, and vice versa: If $x \in A_i$ for some index i, $1 \le i \le n$, then $x \in \cup_{j=1}^{n} A_j$. Similarly, the countable union $\cup_{j=1}^{\infty} A_j$ of an infinite sequence of sets A_j, $j = 1,2,3,....$, is a set with the property that for each $x \in \cup_{j=1}^{\infty} A_j$ there exists a finite index $i \ge 1$ for which $x \in A_i$, and vice versa: If $x \in A_i$ for some finite index $i \ge 1$ then $x \in \cup_{j=1}^{\infty} A_j$.

2. The intersection $A \cap B$ of two sets A an B is the set of elements belong to both A and B. The finite intersection $\cap_{j=1}^{n} A_j$ of sets $A_1,...,A_n$ is the set with the property that if $x \in \cap_{j=1}^{n} A_j$ then for all $i = 1,...,n$, $x \in A_i$, and vice versa: If $x \in A_i$ for all $i = 1, ..., n$, then $x \in \cap_{j=1}^{n} A_j$. Similarly, the countable intersection $\cap_{j=1}^{\infty} A_j$ of an infinite sequence of sets A_j, $j = 1,2,...$, is a set with the property that if $x \in \cap_{j=1}^{\infty} A_j$ then for all indices $i \ge 1$, $x \in A_i$, and vice versa: If $x \in A_i$ for all indices $i \ge 1$ then $x \in \cap_{j=1}^{\infty} A_j$.

3. A is a subset of a set B, $A \subset B$, if all the elements of A are contained in B. If $A \subset B$ and $B \subset A$ then $A = B$.

4. The difference $A \backslash B$ (also denoted by A-B) of sets A and B is the set of elements of A that are not contained in B. The symmetric difference of two sets A and B is denoted and defined by $A \Delta B = (A/B) \cup (B/A)$.

5. If $A \subset B$ then the set $\tilde{A} = B/A$ (also denoted by $\sim A$) is called the complement of A with respect to B. If A_j for $j = 1,2,3,....$ are subsets of B then $\sim \cup_j A_j = \cap_j \tilde{A}_j$ and $\sim \cap_j A_j = \cup_j \tilde{A}_j$,[13] for finite as well as countable infinite unions and intersections.

[12] The symbol $\in$ means "is element of".

[13] These results are called DeMorgan's laws.

6. Sets A and B are disjoint if they do not have elements in common: $A \cap B = \emptyset$, where $\emptyset$ denotes the empty set[14], i.e., a set without elements. In general, a finite or countable infinite sequence of sets is disjoint if their finite or countable intersection is the empty set $\emptyset$.

7. For every sequence of sets A_j , j = 1,2,3,....., there exists a sequence B_j , j = 1,2,3,....., of disjoint sets such that for each j, $B_j \subset A_j$, and $\cup_j A_j = \cup_j B_j$.[15]

The order in which you take unions does not matter, and the same applies to intersections. However, if you take unions and intersections sequentially it matters what is done first. For example, $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$, which is in general different from $A \cup (B \cap C)$, except if $A \subset C$. Similarly, $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$, which is in general different from $A \cap (B \cup C)$, except if $A \subset B$.

APPENDIX B: Supremum and infimum

The supremum of a sequence of real numbers, or a real function, is akin to the notion of a maximum value. In the latter case the maximum value is taken at some element of the sequence, or in the function case some value of the argument. Take for example the sequence $a_n = (-1)^n/n$ for n = 1,2,......., i.e., a_1 = -1, a_2 = 1/2, a_3 = -1/3, a_4 = 1/4, ... Then clearly the maximum value is ½, which is taken by a_2. The latter is what distinguishes a maximum from a supremum. For example, the sequence $a_n = 1 - 1/n$ for n = 1,2,....... is bounded by 1: $a_n < 1$ for all indices $n \geq 1$, and the upper bound 1 is the lowest possible upper bound, but there does not exist a finite index n for which $a_n = 1$. More formally, the (finite) supremum of a sequence a_n (n = 1,2,3,.......) is a number b, denoted by $\sup_{n \geq 1} a_n$, such that $a_n \leq b$ for all indices $n \geq 1$, and for every arbitrary small positive number ε there exists a finite index n such that $a_n > b - \varepsilon$. Clearly, this definition fits a maximum as well: a maximum is a supremum, but a supremum is not always a maximum.

If the sequence a_n is unbounded from above, in the sense that for every arbitrary large real number M there exists an index $n \geq 1$ for which $a_n > M$, then we say that the supremum is

[14] Note that $A \cup \emptyset = A$ and $A \cap \emptyset = \emptyset$. Thus the empty set $\emptyset$ is a subset of any set, including $\emptyset$ itself.

[15] Let $B_1 = A_1$ and $B_n = A_n \backslash \left(\cup_{j=1}^{n-1} A_j \right)$ for n = 2,3,4,.....,

infinite: $\sup_{n\geq 1}a_n = \infty$.

The notion of a supremum also applies to functions. For example the function $f(x) = \exp(-x^2)$ takes its maximum 1 at $x = 0$, but the function $f(x) = 1-\exp(-x^2)$ does not have a maximum; it has supremum 1 because $f(x) \leq 1$ for all x but there does not exists a finite x for which $f(x) = 1$. As another example, let $f(x) = x$ on the interval $[a,b]$. Then b is the maximum of $f(x)$ on $[a,b]$ but b is only the supremum $f(x)$ on $[a,b)$ because b is not contained in $[a,b)$. More generally, the finite supremum of a real function $f(x)$ on a set A, denoted by $\sup_{x\in A}f(x)$, is a real number b such that $f(x) \leq b$ for all x in A, and for every arbitrary small positive number ε there exists an x in A such that $f(x) > b-\varepsilon$. If $f(x) = b$ for some x in A then the supremum coincides with the maximum. Moreover, the supremum involved is infinite, $\sup_{x\in A}f(x) = \infty$, if for every arbitrary large real number M there exists an x in A for which $f(x) > M$.

The minimum versus infimum cases are similar: $\inf_{n\geq 1}a_n = -\sup_{n\geq 1}(-a_n)$ and $\inf_{x\in A}f(x) = -\sup_{x\in A}(-f(x))$.

The concepts of supremum and infimum apply to any collection $\{c_\alpha, \alpha \in A\}$ of real numbers, where the index set A may be uncountable, as we may interpret c_α as a real function on the index set A, say $c_\alpha = f(\alpha)$.

APPENDIX C: Common structure of the proofs of Theorems 6 and 10

The proofs of Theorems 6 and 10 employ a similar argument, namely the following:

THEOREM C.1. *Let* $\mathfrak{C}$ *be a collection of subsets of a set* Ω, *and let* $\sigma(\mathfrak{C})$ *be the smallest* σ*-algebra containing* $\mathfrak{C}$. *Moreover, let* ρ *be a Boolean function on* $\sigma(\mathfrak{C})$, *i.e.,* ρ *is a set function which takes either the value "True" or "False". Furthermore, let* $\rho(A) = True$ *for all sets A in* $\mathfrak{C}$. *If the collection* $\mathscr{D}$ *of sets A in* $\sigma(\mathfrak{C})$ *for which* $\rho(A) = True$ *is a* σ*-algebra itself, then* $\rho(A) = True$ *for all sets A in* $\sigma(\mathfrak{C})$.

Proof: Since $\mathscr{D}$ is a collection of sets in $\sigma(\mathfrak{C})$ we have $\mathscr{D} \subset \sigma(\mathfrak{C})$. Moreover, by assumption, $\mathfrak{C} \subset \mathscr{D}$, and $\mathscr{D}$ is a σ-algebra . But $\sigma(\mathfrak{C})$ is the smallest σ-algebra containing $\mathfrak{C}$, hence $\sigma(\mathfrak{C}) \subset \mathscr{D}$. Thus, $\mathscr{D} = \sigma(\mathfrak{C})$, and consequently, $\rho(A) = $ True for all sets A in $\sigma(\mathfrak{C})$.

Q.E.D.

This type of proof will also be used later on.

Of course, the hard part is to prove that $\mathscr{D}$ is σ-algebra. In particular, the collection $\mathscr{D}$ is not automatically a σ-algebra. Take for example the case where $\Omega = [0,1]$, $\mathfrak{C}$ is the collection of all intervals $[a,b]$ with $0 \le a < b \le 1$, and $\rho(A) = \text{True}$ if the smallest interval $[a,b]$ containing A has positive length: b-$a > 0$, and $\rho(A) = \text{False}$ otherwise. In this case $\sigma(\mathfrak{C})$ consists of all the Borel subsets of [0,1], but $\mathscr{D}$ does not contain singletons whereas $\sigma(\mathfrak{C})$ does, so $\mathscr{D}$ is smaller than $\sigma(\mathfrak{C})$, and therefore not a σ-algebra.

APPENDIX D: Extension of an outer measure to a probability measure

In order to use the outer measure as a probability measure for more general sets that those in $\mathscr{F}_0$, we have to extend the algebra $\mathscr{F}_0$ to a σ-algebra $\mathscr{F}$ of events for which the outer measure is a probability measure. In this appendix it will be shown how $\mathscr{F}$ can be constructed.

LEMMA D.1: *For any sequence B_n of disjoint sets in Ω, $P^*(\cup_{n=1}^{\infty}B_n) \le \Sigma_{n=1}^{\infty}P^*(B_n)$.*

Proof: Given an arbitrary $\varepsilon > 0$ it follows from (26) that there exists a countable sequence of sets $A_{n,j}$ in $\mathscr{F}_0$ such that $B_n \subset \cup_{j=1}^{\infty}A_{n,j}$ and $P^*(B_n) > \Sigma_{j=1}^{\infty}P(A_{n,j}) - \varepsilon 2^{-n}$, hence

$$\Sigma_{n=1}^{\infty}P^*(B_n) > \Sigma_{n=1}^{\infty}\Sigma_{j=1}^{\infty}P(A_{n,j}) - \varepsilon\Sigma_{n=1}^{\infty}2^{-n} = \Sigma_{n=1}^{\infty}\Sigma_{j=1}^{\infty}P(A_{n,j}) - \varepsilon. \tag{54}$$

Moreover, $\cup_{n=1}^{\infty}B_n \subset \cup_{n=1}^{\infty}\cup_{j=1}^{\infty}A_{n,j}$, where the latter is a countable union of sets in $\mathscr{F}_0$, hence it follows from (26) that

$$P^*(\cup_{n=1}^{\infty}B_n) \le \Sigma_{n=1}^{\infty}\Sigma_{j=1}^{\infty}P(A_{n,j}). \tag{55}$$

Combining (54) and (55) it follows that for arbitrary $\varepsilon > 0$,

$$\Sigma_{n=1}^{\infty}P^*(B_n) > P^*(\cup_{n=1}^{\infty}B_n) - \varepsilon. \tag{56}$$

Letting $\varepsilon \downarrow 0$, the lemma follows now from (56) . Q.E.D.

Thus, in order for the outer measure to be a probability measure, we have to impose conditions on the collection $\mathscr{F}$ of subsets of Ω such that for any sequence B_j of disjoint sets in $\mathscr{F}$, $P^*(\cup_{j=1}^{\infty}B_j) \geq \Sigma_{j=1}^{\infty}P^*(B_j)$. The latter is satisfied if we choose $\mathscr{F}$ as follows:

LEMMA D.2: *Let $\mathscr{F}$ be a collection of subsets sets B of Ω such that for* ***any*** *subset A of Ω:*

$$P^*(A) = P^*(A \cap B) + P^*(A \cap \tilde{B}). \tag{57}$$

Then for all countable sequences of ***disjoint*** *sets $A_j \in \mathscr{F}$, $P^*(\cup_{j=1}^{\infty}A_j) = \Sigma_{j=1}^{\infty}P^*(A_j)$.*

Proof: Let $A = \cup_{j=1}^{\infty}A_j$, $B = A_1$. Then $A \cap B = A \cap A_1 = A_1$ and $A \cap \tilde{B} = \cup_{j=2}^{\infty}A_j$ are disjoint, hence

$$P^*(\cup_{j=1}^{\infty}A_j) = P^*(A) = P^*(A \cap B) + P^*(A \cap \tilde{B}) = P^*(A_1) + P^*(\cup_{j=2}^{\infty}A_j). \tag{58}$$

Repeating (58) for $P^*(\cup_{j=k}^{\infty}A_j)$ with $B = A_k$, k=2,...,n, it follows by induction that

$$P^*(\cup_{j=1}^{\infty}A_j) = \Sigma_{j=1}^{n}P^*(A_j) + P^*(\cup_{j=n+1}^{\infty}A_j) \geq \Sigma_{j=1}^{n}P^*(A_j) \text{ for all } n \geq 1,$$

hence $P^*(\cup_{j=1}^{\infty}A_j) \geq \Sigma_{j=1}^{\infty}P^*(A_j)$. Q.E.D.

Note that condition (57) automatically holds if $B \in \mathscr{F}_0$: Choose an arbitrary set A and an arbitrary small number $\varepsilon > 0$. Then there exists an covering $A \subset \cup_{j=1}^{\infty}A_j$, where $A_j \in \mathscr{F}_0$, such that $\Sigma_{j=1}^{\infty}P(A_j) \leq P^*(A) + \varepsilon$. Moreover, since $A \cap B \subset \cup_{j=1}^{\infty}A_j \cap B$, where $A_j \cap B \in \mathscr{F}_0$, and $A \cap \tilde{B} \subset \cup_{j=1}^{\infty}A_j \cap \tilde{B}$, where $A_j \cap \tilde{B} \in \mathscr{F}_0$, we have $P^*(A \cap B) \leq \Sigma_{j=1}^{\infty}P(A_j \cap B)$ and $P^*(A \cap \tilde{B}) \leq \Sigma_{j=1}^{\infty}P(A_j \cap \tilde{B})$, hence $P^*(A \cap B) + P^*(A \cap \tilde{B}) \leq P^*(A) + \varepsilon$. Since ε is arbitrary, it follows now that $P^*(A) \geq P^*(A \cap B) + P^*(A \cap \tilde{B})$.

We show now that

LEMMA D.3: *The collection $\mathscr{F}$ in Lemma D.2 is a $\sigma-$algebra of subsets of Ω, containing the algebra $\mathscr{F}_0$.*

Proof: First, it follows trivially from (57) that $B \in \mathscr{F}$ implies $\tilde{B} \in \mathscr{F}$. Now let $B_j \in \mathscr{F}$. It remains to show that $\cup_{j=1}^{\infty} B_j \in \mathscr{F}$, which we will do in two steps. First, we show that $\mathscr{F}$ is an algebra, and then we use Theorem 4 to show that $\mathscr{F}$ is also a $\sigma-$algebra.

(a) *Proof that $\mathscr{F}$ is an algebra*: We have to show that $B_1, B_2 \in \mathscr{F}$ implies that $B_1 \cup B_2 \in \mathscr{F}$. We have

$$P^*(A \cap \tilde{B}_1) = P^*(A \cap \tilde{B}_1 \cap B_2) + P^*(A \cap \tilde{B}_1 \cap \tilde{B}_2),$$

and since

$$A \cap (B_1 \cup B_2) = (A \cap B_1) \cup (A \cap B_2 \cap \tilde{B}_1)$$

we have

$$P^*(A \cap (B_1 \cup B_2)) \leq P^*(A \cap B_1) + P^*(A \cap B_2 \cap \tilde{B}_1).$$

Thus:

$$\begin{aligned} P^*(A \cap (B_1 \cup B_2)) + P^*(A \cap \tilde{B}_1 \cap \tilde{B}_2) &\leq P^*(A \cap B_1) + P^*(A \cap B_2 \cap \tilde{B}_1) + P^*(A \cap \tilde{B}_2 \cap \tilde{B}_1) \\ &= P^*(A \cap B_1) + P^*(A \cap \tilde{B}_1) = P^*(A). \end{aligned} \tag{59}$$

Since $\sim(B_1 \cup B_2) = \tilde{B}_1 \cap \tilde{B}_2$ and $P^*(A) \leq P^*(A \cap (B_1 \cup B_2)) + P^*(A \cap (\sim(B_1 \cup B_2))$, it follows now from (59) that $P^*(A) = P^*(A \cap (B_1 \cup B_2)) + P^*(A \cap (\sim(B_1 \cup B_2))$. Thus, $B_1, B_2 \in \mathscr{F}$ implies that $B_1 \cup B_2 \in \mathscr{F}$, hence $\mathscr{F}$ is an algebra (containing the algebra $\mathscr{F}_0$).

(b) *Proof that $\mathscr{F}$ is a $\sigma-$algebra*: Since we have established that $\mathscr{F}$ is an algebra, it follows from Theorem 4 that in proving that $\mathscr{F}$ is also a $\sigma-$algebra it suffices to verify that $\cup_{j=1}^{\infty} B_j \in \mathscr{F}$ for disjoint sets $B_j \in \mathscr{F}$: For such sets we have: $A \cap (\cup_{j=1}^{n} B_j) \cap B_n = A \cap B_n$, and $A \cap (\cup_{j=1}^{n} B_j) \cap \tilde{B}_n = A \cap (\cup_{j=1}^{n-1} B_j)$, hence

$$P^*(A \cap (\cup_{j=1}^{n} B_j)) = P^*(A \cap (\cup_{j=1}^{n} B_j) \cap B_n) + P^*(A \cap (\cup_{j=1}^{n} B_j) \cap \tilde{B}_n) = P^*(A \cap B_n) + P^*(A \cap (\cup_{j=1}^{n-1} B_j)).$$

Consequently,

$$P^*(A \cap (\cup_{j=1}^{n} B_j)) = \Sigma_{j=1}^{n} P^*(A \cap B_j). \tag{60}$$

Next, let $B = \cup_{j=1}^{\infty} B_j$. Then $\tilde{B} = \cap_{j=1}^{\infty} \tilde{B}_j \subset \cap_{j=1}^{n} \tilde{B}_j = \sim(\cup_{j=1}^{n} B_j)$, hence

$$P^*(A \cap \tilde{B}) \leq P^*(A \cap (\sim[\cup_{j=1}^{n} B_j])). \tag{61}$$

It follows now from (60) and (61) that for all $n \geq 1$,

$$P^*(A) = P^*(A\cap(\cup_{j=1}^n B_j)) + P^*(A\cap(\sim[\cup_{j=1}^n B_j])) \geq \Sigma_{j=1}^n P^*(A\cap B_j) + P^*(A\cap\tilde{B}),$$

hence

$$P^*(A) \geq \Sigma_{j=1}^\infty P^*(A\cap B_j) + P^*(A\cap\tilde{B}) \geq P^*(A\cap B) + P^*(A\cap\tilde{B}), \tag{62}$$

where the last inequality is due to

$$P^*(A\cap B) = P^*(\cup_{j=1}^\infty (A\cap B_j)) \leq \Sigma_{j=1}^\infty P^*(A\cap B_j).$$

Since we always have $P^*(A) \leq P^*(A\cap B) + P^*(A\cap\tilde{B})$ (compare Lemma D.1), it follows from (62) that for countable unions $B = \cup_{j=1}^\infty B_j$ of disjoint sets $B_j \in \mathscr{F}$,

$$P^*(A) = P^*(A\cap B) + P^*(A\cap\tilde{B}), \tag{63}$$

hence $B \in \mathscr{F}$. Consequently, $\mathscr{F}$ is a $\sigma-$algebra, and the outer measure P^* is a probability measure on $\{\Omega, \mathscr{F}\}$. Q.E.D.

LEMMA D.4: *The $\sigma-$algebra $\mathscr{F}$ in Lemma D.3 can be chosen such that P^* is unique: any probability measure P_* on $\{\Omega,\mathscr{F}\}$ which coincide with P on $\mathscr{F}_0$ is equal to the outer measure P^*.*

Partial proof: Let $\{\mathscr{F}_\theta,\ \theta \in \Theta\}$ be the collection of all $\sigma-$algebras such that for each $\theta \in \Theta$:

(a) $\mathscr{F}_0 \subset \mathscr{F}_\theta$;

(b) there exists a probability measure P_θ on $\{\Omega,\mathscr{F}_\theta\}$ which coincide with *P* on $\{\Omega,\mathscr{F}_0\}$.

According to Lemmas D.2 and D.3 such a collection $\{\mathscr{F}_\theta,\ \theta \in \Theta\}$ exists: For at least one index θ we have that $P_\theta = P^*$ and $\mathscr{F}_\theta = \mathscr{F}$. Let $\mathscr{F}_* = \cap_{\theta\in\Theta}\mathscr{F}_\theta$, which is the smallest $\sigma-$algebra satisfying the properties (a) and (b). Then for each $\theta \in \Theta$, P_θ is a probability measure on $\{\Omega,\mathscr{F}_*\}$. We show now that $P_\theta(A) = P^*(A)$ for all $A \in \mathscr{F}_*$, as follows. Let $\mathfrak{C}$ be the collection of all sets $A \in \mathscr{F}_*$ for which $P^*(A)$ is unique: any probability measure P_* on $\{\Omega,\mathscr{F}_*\}$ which coincide with *P* on $\mathscr{F}_0$ coincides with P^* on $\mathfrak{C}$. Observe first that $\mathscr{F}_0 \subset \mathfrak{C} \subset \mathscr{F}_*$. (*Exercise*: Why?). If $\mathfrak{C}$ is a $\sigma-$algebra, then it must be one of the members of the collection $\{\mathscr{F}_\theta,\ \theta \in \Theta\}$ (*Exercise*: Why?), hence $\mathscr{F}_* \subset \mathfrak{C}$, and consequently $\mathscr{F}_* = \mathfrak{C}$. Thus it suffices to prove that $\mathfrak{C}$ is a $\sigma-$algebra. However, the proof of the latter is too difficult and too long [see Billingsley 1986, Theorems 3.2-3.3], and therefore omitted. Q.E.D.

Combining Lemmas D.2-D.4, Theorem 9 follows.

References:

Billingsley, P. (1986): *Probability and Measure*. New York: John Wiley

Royden, H.L. (1968): *Real Analysis*. London: Macmillan