SECTION A
SUBSCRIPTS AND SUMMATIONS

A.1. SIMPLE SUMMATIONS

A variable may assume various values, which are denoted by subscripts. For example suppose that we have a sample of 10 individuals with information on their income and on the number of their children. We may use the letters X for the variable 'income' and Y for the variable 'number of children', with a subscript to designate the individual to which the particular value of each variable refers:

$$X_1, X_2, \ldots, X_{10}$$
$$Y_1, Y_2, \ldots, Y_{10}$$

where X_i refers to the income of the ith individual and Y_j refers to the number of children of the jth individual.

In order to simplify formulas which involve a large number of values of variables we use the symbol Σ, which is the capital Greek letter sigma (equivalent of the latin letter S) and represents the summation of various values of a variable. Thus:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \ldots + X_n$$

reads: 'the summation of the values of the variable X from the first value to the nth value of this variable'.

There are three basic rules for the algebraic manipulation of terms including summations of variables.

Rule 1. The summation of the sum (or difference) of two or more variables is equal to the sum (or difference) of their respective summations.

Symbolically we may write this rule as follows

$$\sum_{i=1}^{n} (X_i \pm Y_i) = \sum_{i=1}^{n} X_i \pm \sum_{i=1}^{n} Y_i$$

Proof: $\Sigma(X_i + Y_i) = [(X_1 + X_2 + \ldots + X_n) \pm (Y_1 + Y_2 + \ldots + Y_n)] = \Sigma X_i \pm \Sigma Y_i$

Rule 2. The summation of a constant k, times a variable, x_i, is equal to the constant times the summation of the variable.

Symbolically this rule may be written as

$$\sum_{i=1}^{n} kX_i = k\sum_{i=1}^{n} X_i$$

Proof: $\Sigma kX_i = kX_1 + kX_2 + \ldots + kX_n$
$= k(X_1 + X_2 + \ldots + X_n)$
$= k\,\Sigma X_i$

Rule 3. The summation of a constant, k, from 1 to n times, is equal to the product of k times n.

Symbolically we have

$$\sum_{i=1}^{n} k = nk$$

Proof: We may write $\Sigma k = \Sigma(kX_i)$ where all the X's are equal to 1. Thus we have:

$$\Sigma k = \Sigma kX_i = kX_1 + kX_2 + \ldots + kX_n$$

But $X_1 = X_2 = \ldots = X_n = 1$. Therefore

$$\sum_{i=1}^{n} k = k(1) + k(1) + \ldots + k(1) = (k + k + \ldots + k) = nk$$

A.2. DOUBLE SUBSCRIPTS AND SUMMATIONS

Sometimes we want to add various sums. For example assume we have the incomes of 5 individuals from town A, 5 individuals from town B and 5 individuals from town C. We want the total income of all 15 individuals. We may use the letter X for the variable 'income' with two subscripts, X_{ij}, the first (i) referring to the town ($i = 1, 2, 3$) and the second (j) relating to the individual ($j = 1, 2, 3, 4, 5$).

The sum of incomes of the 5 inhabitants of town A is

$$X_{11} + X_{12} + X_{13} + X_{14} + X_{15} = \sum_{j=1}^{5} X_{1j}$$

The sum of incomes of the 5 inhabitants of town B is

$$X_{21} + X_{22} + X_{23} + X_{24} + X_{25} = \sum_{j=1}^{5} X_{2j}$$

The sum of incomes of the 5 inhabitants of town C is

$$X_{31} + X_{32} + X_{33} + X_{34} + X_{35} = \sum_{j=1}^{5} X_{3j}$$

The total sum of all (15) incomes is

$$\sum_{j=1}^{5} X_{1j} + \sum_{j=1}^{5} X_{2j} + \sum_{j=1}^{5} X_{3j} = \sum_{i=1}^{3} \sum_{j=1}^{5} X_{ij} = \sum_{j=1}^{5} \sum_{i=1}^{3} X_{ij}$$

In general

$$\sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij} = \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ij} = \sum_{i=1}^{n} (X_{i1} + X_{i2} + \ldots + X_{im})$$

$$= (X_{11} + X_{21} + \ldots + X_{n1}) + (X_{12} + X_{22} + \ldots + X_{n2}) + + \ldots + (X_{1m} + X_{2m} + \ldots + X_{nm})$$

The following rules which involve double summation are useful.

Rule 4.

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij} + Y_{ij}) = \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij}$$

Proof:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (X_{ij} + Y_{ij}) = \sum_{j=1}^{m} (X_{1j} + Y_{1j}) \sum_{j=1}^{m} (X_{2j} + Y_{2j}) + \ldots + \sum_{j=1}^{m} (X_{nj} + Y_{nj})$$

$$= \sum_{j=1}^{m} (X_{1j} + X_{2j} + \ldots + X_{nj}) + \sum_{j=1}^{m} (Y_{1j} + Y_{2j} + \ldots + Y_{nj})$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij}$$

Rule 5.

$$\sum_{i=1}^{n} \sum_{j=1}^{m} X_i Y_j = \left(\sum_{i=1}^{n} X_i\right) \left(\sum_{j=1}^{m} Y_j\right)$$

Proof:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} X_i Y_j = \sum_{i=1}^{n} (X_i Y_1 + X_i Y_2 + \ldots + X_i Y_m)$$

$$= Y_1 \sum_{i=1}^{n} X_i + Y_2 \sum_{i=1}^{n} X_i + \ldots + Y_m \sum_{i=1}^{n} X_i$$

$$= \sum_{i=1}^{n} X_i (Y_1 + Y_2 + \ldots + Y_m)$$

$$= \left(\sum_{i=1}^{n} X_i\right) \left(\sum_{j=1}^{m} Y_j\right)$$

SECTION B

FREQUENCY DISTRIBUTIONS AND PROBABILITY DISTRIBUTIONS

B.1. POPULATIONS AND SAMPLES

The *population* of a variable X consists of all the conceptually possible values that the variable may assume. Some of these values may have already been observed; others may not have occurred, but their occurrence is conceivably possible. For example the variable 'income per head' may assume any positive value, from zero to millions of pounds. A very large number of values of income is observed in any one period. Yet there are infinite other values which, although not already assumed, may be observed in some other period.

The number of conceptually possible values of a variable is called the *size of the population*. The size varies according to the phenomenon being investigated. For example a study of incomes may be conducted at a regional level, at a country level, at a world-wide level. In the first instance the population will consist of the incomes of one region; in the second case the population will consist of the incomes of all the residents of the country; in the third case the population will comprise the incomes of the residents of all the countries of the world.

A population may be *finite*, when it consists of a given number of values, or it may be *infinite*, when it includes an infinite number of values of the variable. Most of the populations with which we are concerned in econometrics are infinite.

In most cases we do not know all the values of a population. What we usually have is a certain number of values that any particular variable has assumed and which have been recorded in one way or another. Such data form a sample from the population. *A sample* is a collection of observations on a certain variable. The number of observations included in the sample is called the *size* of the sample.

The main object of the theory of statistics is the development of methods of drawing conclusions about the (unknown) population from the information provided by a sample.

In order to facilitate the study of populations and samples, statisticians have introduced various descriptive measures, that is various characteristic values which describe the important features of the sample or the population. The most important of these characteristics are the mean, the variance and the standard deviation. To distinguish between samples and populations statisticians use the term *parameters* for the basic descriptive measures of the population and the term *statistics* for the characteristic measures of the sample. Furthermore they use Greek letters for the population parameters and Latin letters for the sample statistics. In summary, the basic descriptive measures of populations and samples are

Population Parameters	*Sample Statistics*
1. Population mean: μ	1. Sample mean: $\bar{X}$
2. Population variance: σ_x^2	2. Sample variance: s_x^2
3. Population standard deviation: σ_x	3. Sample standard deviation: s_x

Before giving the formal definitions of these measures, we will develop the concepts of frequency and probability and their distributions (frequency distributions and probability distributions). Frequency distributions refer to samples while probability distributions are associated with populations.

B.2. FREQUENCY DISTRIBUTIONS

Frequency distributions are associated with samples. If we draw values from a population and record them, we may observe that some values appear more frequently than others in the sample. For example if we are concerned with the population 'income per head of the U.K.', values of £1,500 or £1,600 do appear more often than values of £100 or £1,000,000. The number of times that a certain value appears when we draw observations from a population is called the *absolute frequency* of that value. If for example the income £1,000 is observed 10 times in our drawings from the population of incomes, we say that the value £1000 has an absolute frequency of 10. *Relative frequency* of a particular value is the ratio of the absolute frequency divided by the total number of the observations (n) drawn from the population. If, in our example the total number of observations is 400, the relative frequency of the value X = £1,000 is $\frac{10}{400} = 0{\cdot}025$.

Conventionally the absolute frequency of a value of a variable is denoted by the symbol f_i, where the subscript i refers to the ith value of the variable X.

It should be clear that the total number of observations, or drawings, is equal to the sum of the frequencies of the individual values that have been observed:

$$\sum_{i=1}^{k} f_i = n$$

(k = possible values of the variable X).

Thus the relative frequency of the ith value of X may be written as

$$\begin{bmatrix}\text{relative}\\ \text{frequency}\\ \text{of } X_i\end{bmatrix} = f(X_i) = \frac{f_i}{n} = \frac{f_i}{\Sigma f_i}$$

The sum of all the relative frequencies is equal to unity:

$$\Sigma \frac{f_i}{n} = \frac{1}{n} \Sigma fi = \frac{n}{n} = 1$$

given that $\Sigma f_i = n$.

If we record all the sample values of X with their frequencies we obtain a set of pairs of values with their respective frequencies, which is called the *frequency distribution* (or simply the *distribution*) of the variable X. The frequency distribution is an organised presentation of the observed values of a variable in a sample: it shows the number of observations for each value of the variable in the sample (in the case of a discrete variable) or the number of observations in each interval of values of the variable in the sample (in the case of a continuous variable).

Symbolically the frequency distribution is denoted by $f(X)$ and it gives for each value X_i the relative frequency, $f(X_i)$, with which that value occurs in the set of n observations.

Distributions may be presented in a tabular form, on a graph or with a mathematical formula. Graphically frequency distributions of discrete variables are presented by a *frequency polygon* or a *histogram. An important feature of the histogram is that its area represents the sum of the relative frequencies and hence it is equal to 1.*

An example will illustrate these definitions. Assume that we take a sample of the daily incomes of 1,000 individuals. We observe that 20 individuals have an income of £1 per day; 200 individuals have an income of £2; 540 individuals have an income of £3; 220 individuals have an income of £4; and 20 individuals have an income of £5. This information constitutes the frequency distribution of the variable income. The tabular presentation of this distribution is shown in Table 1.

The graph of the above frequency distribution is shown in figures 1 and 2. On the vertical axis of figure 1 we measure the absolute or the relative frequencies and on the horizontal axis we measure the value of the variable. If the data are not grouped, as in our example, the frequency distribution is presented by vertical lines drawn on top of each value of X to a height equal to the frequency (or the relative frequency). If we join the ends of the vertical lines we form a *frequency polygon*. However, it is more useful to draw the graph of the relative frequencies in a slightly different way. On the horizontal axis we measure single unit intervals, that is, we take classes or class-intervals equal to one unit of X. (In our example we have measured income in single units to begin with.) On top of the unit intervals we draw rectangles with height equal to the relative frequency. These rectangles form the histogram of the frequency distribution (figure 2). The base of each rectangle is (by construction) equal to one unit of X. Consequently the area of each rectangle is equal to the relative frequency of the particular value of X. We know that the sum of the relative frequencies of the values of X is equal to unity. Hence the total area of all the rectangles of the histogram is equal to 1.

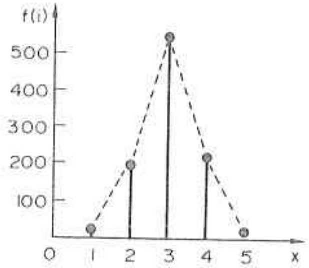


Figure 1(a)
Frequency polygon for the absolute frequencies of X

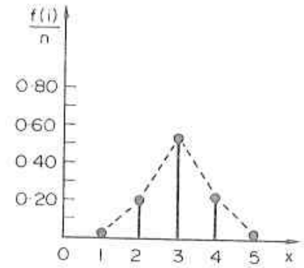


Figure 1(b)
Frequency polygon for the relative frequencies of X

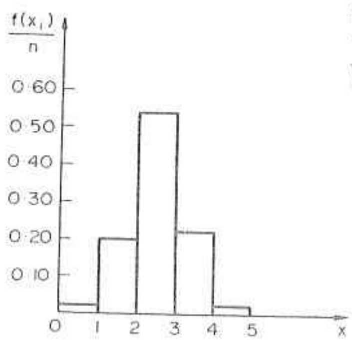


Figure 2
The histogram of the frequency distribution of table 1

Table 1. Frequency distribution of X (daily income)

Values of income (in £) X_i	Absolute frequency f_i	Relative frequency f_i/n
1	20	$\frac{20}{1,000} = 0{\cdot}02$
2	200	$\frac{200}{1,000} = 0{\cdot}20$
3	540	$\frac{540}{1,000} = 0{\cdot}54$
4	220	$\frac{220}{1,000} = 0{\cdot}22$
5	20	$\frac{20}{1,000} = 0{\cdot}02$
Total frequency	$\sum_{i=1}^{n} f_i = 1,000$	$\sum_{i=1}^{n} \frac{f_i}{n} = 1$

Heuristically we may think as follows:

The area of a rectangle (A_i) is equal to the product of its height (H_i) times its base (B_i)

$$A_i = H_i \times B_i$$

But H_i = relative frequency of the ith value of $X = f_i$; and $B_i = 1$ by construction, since we took the class interval equal to one unit of X.

Consequently

$$A_i = (f_i/n)(1) = f_i/n$$

The sum of the areas of all rectangles of the histogram is

$$\sum_{i}^{k} A_i = \sum_{i=1}^{n} \frac{f_i}{n} = 1$$

Thus the area of a histogram represents the sum of the relative frequencies and is equal to one.

B.3. PROBABILITY DISTRIBUTIONS

Probability distributions refer to populations and are analogous to the frequency distributions of samples.

Let us first define the concept of probability. It has been observed that as we increase the number of observations from the population of a random variable the relative frequency of any value X_i tends to stabilize at a certain value, which is called the limiting value of the relative frequency or *probability* of the value X_i of the random variable. In other words the *probability* of a value X_i of a random variable is the limiting value of the relative frequency of that value as the total number of observations on the variable approaches infinity, the value which the relative frequency assumes *at the limit* as the number of observations tends to infinity. This is symbolically written as

$$P(X_i) = \lim_{n \to \infty} \frac{f_i}{\Sigma f_i}$$

Sometimes the symbol $f(X_i)$ is used to denote the probability of the value X_i. Each value of a random variable has some probability associated with itself, i.e. there is some probability of observing any value of a variable. The sum of the probabilities of all values, being the sum of relative frequencies (at the limit), is equal to unity.

Probabilities, interpreted as limiting values of relative frequencies, can be estimated from empirical data. Thus in practice we take *the relative frequency as equal to the probability*. It should be however clear that these two concepts are different. The larger the number of observations the closer will the relative frequency be to the probability. For an illustration let us examine the occurrence of male births in the population. We want to find what is the probability that a birth will be the birth of a male. Every year thousands of births are recorded. If we take the total births over, say, a decade all over the world, the number of observations increases considerably. For any practical purpose such a large number of observations may be considered as adequate for the concept of limiting value of the relative frequency, although there is no actual approach of n to infinity. With such calculations it has been found that male births occur with a relative frequency approximately of $\frac{1}{2}$ (or 0·50, or 50 per cent). Thus we say that the probability of a male birth is 0·50.

The probability of any event (or any value of a variable) can assume any value between 0 and 1. Symbolically:

$$0 \leqslant f(X_i) \leqslant 1$$

If the probability of X assuming the particular value X_i is equal to zero this means that the variable cannot assume the value X_i; in other words a probability of zero for X_i suggests impossibility of occurrence of the value X_i. If the probability of a particular value X^* is equal to unity this means that this value does occur at any time, i.e. the value X^* is the only value that the variable can assume. In this case the 'variable' is really a constant. It may assume only the single value X^*. Thus a probability of one corresponds to certainty; that is, the value X^* is certain to occur. Any probability of a particular value X_i between zero and one shows some uncertainty in the occurrence of this particular value of the variable X.

There are several rules for the calculation of the probabilities of one or more values being observed in any particular instance. These rules are known as *laws of probabilities* and are developed in section F below.

We said that a random variable is a variable whose values are associated with some probability of being observed, and that a random variable can be discrete (when it can assume only finite values) or continuous (when it can assume an

infinite number of values within any given interval). We shall examine the probability distributions of discrete and continuous variables separately.

B.3.1. *Discrete Random Variables and their Probability Distributions*

If a variable is discrete its values are distinct, i.e. they are separated by finite distances. To each value we may assign a given probability. If X is a discrete random variable which may assume the values $X_1, X_2, \ldots, X_n$, with respective probabilities $f(X_1), f(X_2), \ldots, f(X_n)$, then the entire set of pairs of permissible values together with their respective probabilities is called the *probability distribution* of the random variable X.

The probability distribution is mathematically denoted by a function, $f(X_i)$, which gives the probability of any particular value X_i (of the discrete variable X) being observed. If all the permissible values of X have equal probability of being observed, the probability distribution is called the *uniform* or *rectangular* distribution. For example, in casting an unbiased die each of the permissible values (1, 2, 3, 4, 5, 6) has a probability of 1/6 of being observed. Thus the probability distribution of the random variable X which denotes the results of casting a die is a uniform distribution. It is shown in table 2 and its graph appears in figure 3.

Table 2. A uniform probability distribution

Probability distribution of variable X = results of casting a die	
Permissible Values of X	Probability of each value: $f(X_i)$
$X = 1$	1/6
$X = 2$	1/6
$X = 3$	1/6
$X = 4$	1/6
$X = 5$	1/6
$X = 6$	1/6
	$\Sigma f(X_i) = 1$

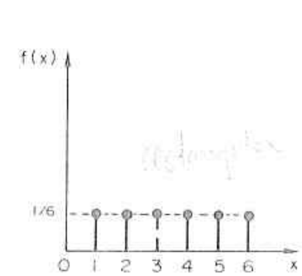


Figure 3. A uniform probability distribution of a discrete variable

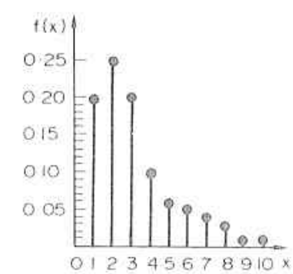


Figure 4. A non-uniform probability distribution of a discrete variable

In most econometric applications the variables involved do not have a uniform probability distribution. Each value of the variable has usually its own probability of being observed. For example suppose that we have the random variable X denoting the number of children per family, and that the permissible values of this variable are the positive integers 1, 2, 3, . . . 10, each with the probability (estimated from a very large number (sample) of families) shown in table 3.

Table 3. A non-uniform probability distribution

Probability distribution of X_i	
Permissible values X_i	Probability of each value $f(X_i)$
0	0·05
1	0·20
2	0·25
3	0·20
4	0·10
5	0·06
6	0·05
7	0·04
8	0·03
9	0·01
10	0·01
	$\Sigma f(X_i) = 1$

The graph of this non-uniform probability distribution is shown in figure 4.

In many cases we are interested in knowing the probability that the discrete variable X will assume a value less than or equal to a given value. These are called *cumulative probabilities* and are usually denoted by $F(X)$. The cumulative probability is the sum of individual probabilities. For example the probability of a couple having 4 or less children is

$$P(0 \leqslant X \leqslant 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$
$$= 0{\cdot}05 + 0{\cdot}20 + 0{\cdot}25 + 0{\cdot}20 + 0{\cdot}10 = 0{\cdot}80$$

In general if $X_1, X_2, \ldots, X_k, \ldots, X_n$ are successive values of X, then the cumulative probability of the first k values is

$$F(X_k) = f(X_1) + f(X_2) + \ldots + f(X_k) = \sum_{i=1}^{k} f(X_i)$$

B.3.2. *Continuous Random Variables and Probability Density Functions*

If a variable is continuous, it may assume an infinite number of values within any given interval. For example, the variable 'income' can assume infinite values between, say, £10 and £100. In the case of a continuous variable the

probability of any particular value must be zero; that is, $P(X = X^*) = 0$. However the probability of X assuming values within an interval (X_1 and X_2), no matter how small this interval might be, is a finite number and can be computed (by integrals). Thus the fact that $P(X = X^*) = 0$ in the case of a continuous variable should not be interpreted as meaning that the value X_i^* is impossible. After all, the variable does assume particular values in any one situation. Rather $f(X_i^*)$ should be interpreted as the average probability of values very close in the neighbourhood of X_i^*. (See A. S. Goldberger, *Econometric Theory*, Wiley, 1964, p. 69.)

The probability distribution of a continuous variable is called *probability density function* or simply *probability function*. However, in this book we use the term 'probability distribution' both for discrete and for continuous variables.

The probability function is mathematically denoted by a continuous function and graphically presented by a curve. Some typical shapes of probability functions (distributions) are shown in figure 5.

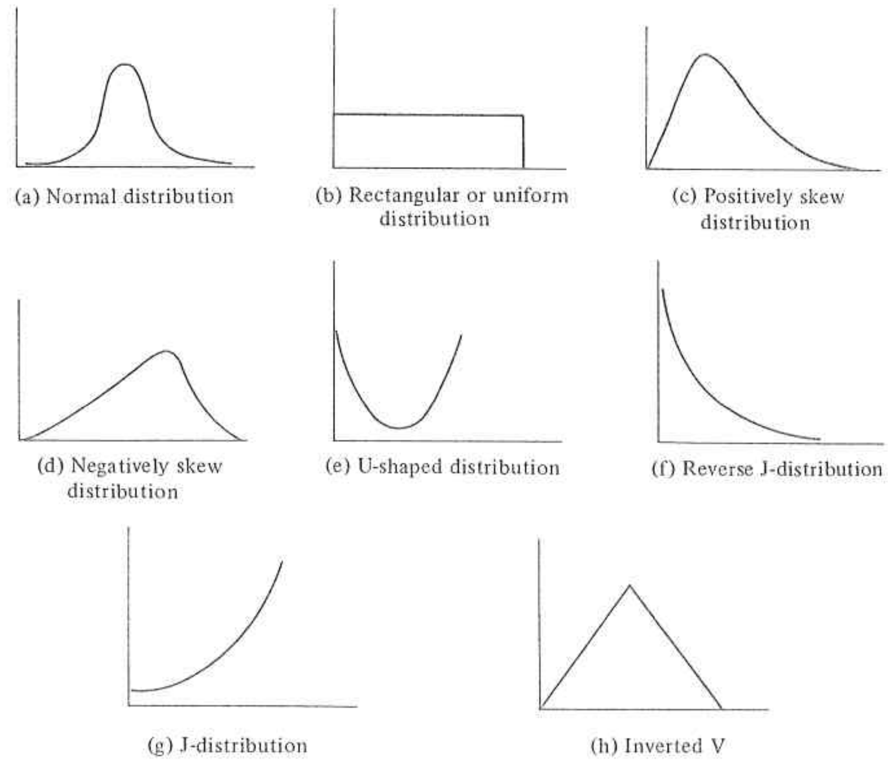


Figure 5

A normal distribution is a symmetrical bell-shaped curve which extends indefinitely in both directions: the curve comes closer and closer to the horizontal axis but never quite reaches it. There is a certain central value of X around which the values of X are clustered symmetrically.

In a rectangular (or uniform) distribution all the values of the variable X are observed with the same probability.

If the majority of values of X are lower than the mean (average) value of X, the distribution is positively skewed; i.e. skewed to the right. If most of the values of X are higher than the mean (average) value of X, the distribution is negatively skewed, i.e. skewed to the left.

An important feature of the curves of probability distributions is that the areas under these curves represent probabilities. *The total area under the curve of a probability distribution,* being the sum of individual probabilities, *is equal to unity.*

For example assume that the probability distribution of a continuous variable X is represented by the curve of figure 6. The probability that X assumes a value

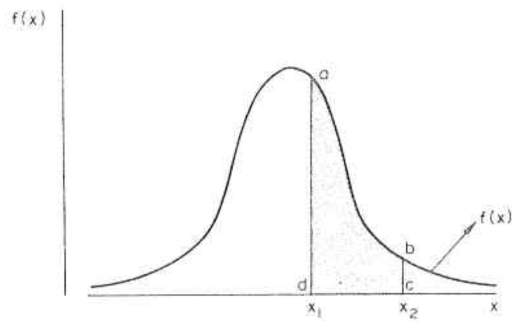


Figure 6

within the interval X_1 to X_2 is given by the shaded area *abcd*. This probability may be computed by integration

$$P(X_1 < X < X_2) = \int_{X_1}^{X_2} f(X)\,dx$$

where $f(X)$ is the equation of the probability function. (The integration in the case of a continuous variable is analogous to the summation of the probabilities of individual values of a discrete variable). Fortunately it will not be necessary to compute probabilities by integrals in econometric applications. Probabilities, when required, will be taken out of particular standard tables, whose use will be explained in subsequent sections.

It should be stressed that the fact that for a continuous variable the probability of any particular value is equal to zero, $P(X = X_i^*) = 0$, does not mean that the value X_i^* is impossible. When we draw a sample from a continuous variable, $P(X_i^*)$ should be interpreted as the average probability of values which are very close in the neighbourhood of X_i (see A. S. Goldberger, *Econometric Theory*, p. 69).

B.3.3. *Joint Probability Distributions*

If we have two random variables X and Y their joint probability, denoted by $f(X_i Y_j)$, is a function which gives the probability that X assumes a given value X_i and Y assumes a given value Y_j, jointly. For example assume that we have the variable X denoting the sex (male or female) of a person, and the variable Y denoting whether the person is a smoker or a non-smoker. Assume that we draw a very large sample of men and women and we find the following probabilities:

P (male and smoker) = 0·5
P (female and smoker) = 0·3
P (male and non-smoker) = 0·1
P (female and non-smoker) = 0·1

This information may be presented in the following tabular form.

Table 4. Joint distribution of X and Y

Values of Y	Values of X: X_1 = Male	Values of X: X_2 = Female	Marginal Probability of Y
Y_1 = smoker	$f(X_1 Y_1) = 0{\cdot}5$	$f(X_2, Y_1) = 0{\cdot}3$	$f(Y_1) = 0{\cdot}8$
Y_2 = non-smoker	$f(X_1, Y_2) = 0{\cdot}1$	$f(X_2, Y_2) = 0{\cdot}1$	$f(Y_2) = 0{\cdot}2$
Marginal Probability of X	$f(X_1) = 0{\cdot}6$	$f(X_2) = 0{\cdot}4$	1·00

Thus the joint probability of $X = X_1$ *and* $Y = Y_1$ is

$$f(X_1, Y_1) = 0{\cdot}5$$

Similarly, the joint probability of $X = X_2$ *and* $Y = Y_1$ is

$$f(X_2, Y_1) = 0{\cdot}3$$

and so on.

Marginal probability of X_i is the probability of X assuming a specific value X_i irrespective of the value of Y, that is, irrespective of what value Y assumes. In our example the marginal probability of X_1 is

$$f(X_1) = f(X_1, Y_1) + f(X_1, Y_2)$$

$$= \sum_{i=1}^{2} (X_1, Y_i)$$

$$= 0\cdot5 + 0\cdot1 = 0\cdot6$$

Similarly the marginal probability of X_2 is

$$f(X_2) = f(X_2, Y_1) + f(X_2, Y_2)$$

$$= \sum_{i=1}^{2} (X_2, Y_i)$$

$$= 0\cdot3 + 0\cdot1 = 0\cdot4$$

In general, the marginal probability of X_i is given by adding up the corresponding probabilities over all the values of the other variable (Y); and similarly the marginal probability of Y_i is found by summing all the corresponding probabilities of the other variable (X). Symbolically

$$\begin{bmatrix} \text{Marginal} \\ \text{Probability } X_i \end{bmatrix} = f(X_i) = \sum_j f(X_i, Y_j)$$

and

$$\begin{bmatrix} \text{Marginal} \\ \text{Probability } Y_j \end{bmatrix} = f(Y_j) = \sum_i (X_i, Y_j)$$

The extension of the above results to the joint distribution of more than two variables (multivariate distributions) is straightforward.

SECTION C

POPULATION PARAMETERS AND SAMPLE STATISTICS

C.1. POPULATION PARAMETERS

Although the construction of probability distributions and their graphic presentation help in the study of populations, they still involve a lot of tedious work. In order to simplify further the study of populations, statisticians have defined various descriptive measures, i.e. basic characteristic values which describe adequately the basic features of the population. These descriptive measures are called *parameters* of the population or *expected values* (or *mathematical expectations*) of the probability distribution of the variables. The term 'expected value' is used in order to denote the fact that the parameters of the population of a continuous variable are never observed, and those of the population of a discrete variable, although in principle observable, may be impossible to observe in practice because the number of values of the population are not all ~~of them~~ known. Hence we say that we *expect* the population parameters to have a certain (expected) value.

The most common parameters of probability distributions are the following.

1. Measures of location (or of central tendency).

There are various measures of location. The most important is *the arithmetic mean* of the population. The mean is the 'central' or 'average' value of the variable whose population we study. It is denoted by the Greek letter μ.

2. Measures of dispersion (or variation).

There are various measures of dispersion, the most important being the variance and the standard deviation of the population of X. The variance shows how close to the mean the various values of X cluster. The standard deviation is the square root of the variance and gives the average distance of the various values of X from the arithmetic mean.

The variance of a population is denoted by the Greek letter σ_x^2. The population standard deviation then is $\sqrt{\sigma_x^2} = \sigma_x$.

3. Measures of the skewness

Such measures show the degree of symmetry of the distribution around the mean value. Two skew distributions are shown in figures 5(c) and 5(d).

4. Measures of the kurtosis

Such measures show the degree of peakedness or flatness of the distribution of X.

Of the above measures the most important are the measures of location and dispersion. Thus when we speak of the parameters of a population we will refer to its mean, variance and standard deviation.

The population mean is called *the expected value* of the population and it is conventionally denoted as $E(X)$ or μ. For a *discrete random variable* the expected value is computed by the sum of the products of the values X_i multiplied by their respective probabilities

$$E(X) = \mu = \sum_{i=1}^{n} X_i f(X_i)$$

where $f(X_i)$ is the probability of the variable X assuming the value X_i.[1,2] The expected value is the weighted arithmetic mean of the (random) variable X, the weights being the individual probabilities.

Example. Assume that the variable X 'number of children per family' assumes values from zero to ten with the probabilities shown in table 3. The expected value of the random variable X, i.e. the average number of children in a family, is computed as follows:

$$E(X) = \Sigma X_i f(X_i)$$

$$E(X) = (0)(0{\cdot}05) + (1)(0{\cdot}20) + (2)(0{\cdot}25) + \ldots + 10(0{\cdot}01) \approx 3$$

The average number of children in a family in our example is 3, approximately.

The variance of a population is the expected value of the squared deviations of the value of X from their expected mean value

$$\text{var}(X) = \sigma_x^2 = E[X_i - E(X)]^2$$

For a discrete random variable the variance is equal to the sum of the squared deviations of the X's from their expected mean value multiplied by the respective

[1] *Proof.*

$$E(X) = \frac{\Sigma X_i f_i}{\Sigma f_i} = \Sigma X_i \frac{f_i}{\Sigma f_i}$$

But

$$\frac{f_i}{\Sigma f_i} = f(X_i) = \text{probability of the value } X_i$$

Therefore $E(X) = \Sigma X_i f(X_i)$.

[2] If the variable is continuous the expected value is given by the same formula by substituting the summation with the integral:

$$E(X) = \int_a^b X f(X)\, dx$$

where $f(X)$ is the probability density function of X.

probabilities of the values of X[1]

$$\text{var}(X) = \sigma_x^2 = \sum_i^n (X_i - \mu)^2 f(X_i)$$

It can be shown that

$$\text{var}(X) = E(X^2) - \mu^2$$

Proof:

$$\begin{aligned}\text{var}(X) = \sigma_x^2 &= E[X - E(X)]^2 \\ &= E(X - \mu)^2 \\ &= E(X^2 + \mu^2 - 2X\mu) \\ &= E(X^2) + E(\mu^2) - 2E(X\mu) \\ &= E(X^2) + \mu^2 - 2\mu E(X) \\ &= E(X^2) + \mu^2 - 2\mu^2 \qquad \text{(given } E(X) = \mu\text{)} \\ &= E(X^2) - \mu^2\end{aligned}$$

The variance shows the way in which the various values of the random variable X are distributed around their expected mean value. The smaller the variance, the closer the cluster of the values of X around the mean and vice versa.

The standard deviation of a population is the square root of the variance. It is denoted by σ_x and is computed from the expression

$$\sigma_x = \sqrt{E[X_i - E(X)]^2} = \sqrt{E(X_i - \mu)^2}$$

Like the variance, the standard deviation is a measure of the dispersion of the values of X around their (population) mean.[2]

[1] If the variable is continuous the variance is given by the same expression by substituting the summation with the integral

$$E(X - E(X))^2 = \int_a^b [X - E(X)]^2 f(X)\, dx$$

where $f(X)$ is the probability density function of X.

[2] The parameters of a population are also called the *moments* of the population. In a distribution of a discrete variable the rth moment about zero is defined as

$$m'_r = \sum_i^n X_i^r f(X_i)$$

and the rth moment about the mean is defined as

$$m_r = \sum_i^n (X_i - \mu)^r f(X_i)$$

The first moment about zero is the mean of the distribution

$$\mu = m'_1 = \Sigma X_i f(X_i)$$

The second moment about the mean is the variance of the distribution

$$\sigma_x^2 = \sum_i^n (X_i - \mu)^2 f(X_i) = \sum_i^n (X_i - \mu)^2 P_i$$

If we know the mean μ and the standard deviation σ_x of a population, we have a pretty good idea of the form of its distribution. From the expected value, μ, we know where the distribution is located along the horizontal axis: we know its 'central' point. From the variance or standard deviation we know whether the values of the variable are clustered closely around the mean or are dispersed widely about it. The shape of any distribution depends on the mean and variance of the population of the variable to which the distribution refers. The following results (which are discussed in detail in Section E below) show the extent of information which we have if we know that a population is normal (that is, has a normal distribution) with mean μ and standard deviation σ_x:

68% of the values of X will lie within the interval $[\mu \pm 1\sigma_x]$

95% of the values of X will lie within the interval $[\mu \pm 2\sigma_x]$

99% of the values of X will lie within the interval $[\mu \pm 3\sigma_x]$

Covariance of two random variables

The *covariance* of two random variables is the expected value of the product of the deviations of the variables from their expected values. Symbolically

$$\text{cov}(XY) = E\{[X_i - E(X)]\,[Y_j - E(Y)]\}$$

which can be shown to reduce to

$$\text{cov}(XY) = E(XY) - E(X)E(Y)$$ (see p. 542)

C.2. SAMPLE STATISTICS

The parameters of a population can be mathematically measured if we know all the values of the population. However in most cases we do not know all the values of a population. In any particular case what we usually have is a sample from the population. Statisticians have introduced various descriptive measures for samples, which describe the basic features of the sample values, and are called *sample statistics.* The basic statistics of a sample, corresponding to the parameters of the population, are: the sample mean, the sample variance and the sample standard deviation.

The *sample mean* is the average value in the sample. It is designated by $\overline{X}$. The simple arithmetic mean is calculated by adding up the observations of the sample and then dividing the total by the number of observations. Thus

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

*

For example suppose that X stands for the incomes of the inhabitants of a country. We choose a sample of 10 individuals and we want to compute the average income.

Suppose that the observations of the sample are as follows:

Individual	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Incomes (£ per head)	95	120	80	150	50	130	70	90	110	105

The average income will be

$$\overline{X} = \frac{\sum_{i=1}^{10} X_i}{n} = \frac{95 + 120 + 80 + \ldots + 105}{10} = \frac{1000}{10} = 100$$

The weighted arithmetic mean is the sum of the products of the X's and their weights, w's, divided by the sum of the weights

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i w_i}{\sum_{i=1}^{n} w_i}$$

Weighted means are widely used in price indexes, or quantity indexes. The weights may be chosen according to the purpose and the nature of the index.

The *sample variance* is a measure of the dispersion of the values of X in the sample around their average value. It is designated by s_x^2. It is computed by the formula.

$$s_x^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n} = \frac{\Sigma X_i^2 - n\overline{X}^2}{n} = \frac{\Sigma X_i^2}{n} - \overline{X}^2$$

For our example the variance of the incomes of ten individuals is

$$s_x^2 = \frac{(-5)^2 + (20)^2 + (-20)^2 + \ldots + (5)^2}{10} = 785$$

Note that in measuring the variance we take the sum of the squares of the deviations of X's from X. This is necessary because the simple sum of the deviations is equal to zero (see p. 529).

The *sample standard deviation* is designated by s_x and is equal to the square root of the variance

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}}$$

In the above example the standard deviation is 28.

The sample covariance of two variables

Given any two sets of n sample observations on two variables

$$X_1, X_2, \ldots, X_n$$

$$Y_1, Y_2, \ldots, Y_n$$

we are often interested in measuring their *covariance*, that is, the way in which they change together (co-vary). The covariance is defined as the sum of the products of deviations of X and Y from their means, $\bar{X}$ and $\bar{Y}$ respectively, divided by the number of observations.

$$\text{cov}_{(XY)} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

If the variables X and Y are not changing together, if they are independent, their covariance is equal to zero, since the sum of the products of their deviations is zero (see p. 542).

From a population we may draw many samples. A *random sample* of a discrete variable is a sample of observations each of which has the same probability of being drawn from the population. For example if we want a random sample of consumers we may go in the street and ask any one whom we happen to meet on his consumption expenditures. Any individual has the same chance of being included in this sample. If the variable is continuous, a sample drawn from its population is random if the values that are observed are independent (see section F below for the definition of independence of any two events, values).

In most cases in econometrics we use data either in the form of time-series observations or in the form of cross-section observations. These data are considered as being randomly drawn from a hypothetically infinite population. In other words in most cases *the data which econometricians use in their work are considered as random samples from a hypothetically infinite population.*[1] For example a time series of prices of a commodity, say cars, recorded on the market may be considered as a random sample of all possible prices of cars, which form the hypothetically infinite population of car prices.

From the observations of a sample we may compute the sample mean, variance and standard deviation. Once the sample statistics are computed we may make inferences about the population parameters. The basic rules of statistical inference are outlined in section E below.

[1] See M. Brennan, *Preface to Econometrics*, p. 293.

SECTION D
THE ALGEBRA OF EXPECTED VALUES AND SAMPLE STATISTICS

D.1. THE ALGEBRA OF TERMS INVOLVING SAMPLE STATISTICS

Rule 1. The mean value of the sum (or difference) of two variables is equal to the sum (or difference) of their means. Symbolically

$$\overline{(X \pm Y)} = \bar{X} \pm \bar{Y}$$

Proof:
$$\overline{(X+Y)} = \frac{\Sigma(X+Y)}{n} = \frac{\Sigma X + \Sigma Y}{n} = \frac{\Sigma X}{n} + \frac{\Sigma Y}{n} = \bar{X} + \bar{Y}$$

Rule 2. The sum of deviations of a variable from its mean is equal to zero. Symbolically

$$\sum_{i=1}^{n} (X_i - \bar{X}) = \sum_{i=1}^{n} x_i = 0 \quad \text{where } x_i = X_i - \bar{X}$$

Proof:
$$\Sigma(X_i - \bar{X}) = \Sigma X_i - n\bar{X} \quad \text{(given that } \bar{X} \text{ is a constant)}$$
$$= \Sigma X_i - n\frac{\Sigma X_i}{n} = 0$$

Rule 3. The mean of a variable multiplied by a constant k is equal to this constant times the mean of the variable. Symbolically

$$\overline{kX} = k\bar{X}$$

Proof:
$$\overline{(kX)} = \frac{\Sigma kX_i}{n} = k\frac{\Sigma X_i}{n} = k\bar{X} \qquad \text{(given that } k \text{ is a constant)}$$

Rule 4. The variance of a variable X multiplied by a constant k is equal to the square of the constant times the variance of the variable. Symbolically

$$\text{var}(kX) = k^2\{\text{var}(X)\}$$

Proof:
$$\text{var}_{(kX)} = \frac{\Sigma(kX - \overline{kX})^2}{n} = \frac{\Sigma(kX - k\bar{X})^2}{n} = k^2\frac{\Sigma(X - \bar{X})^2}{n} = k^2\{\text{var}(X)\}$$

Rule 5. The variance of the sum of two variables is equal to the sum of the variances plus twice their covariance

$$s^2_{(X+Y)} = s^2_X + s^2_Y + 2\,\text{cov}_{(XY)}$$

Proof:
$$s^2{(X+Y)} = \frac{\Sigma[(X+Y) - \overline{(X+Y)}]^2}{n} = \frac{\Sigma[(X-\bar{X}) + (Y-\bar{Y})]^2}{n}$$
$$= \frac{\Sigma\{(X-\bar{X})^2 + (Y-\bar{Y})^2 + 2(X-\bar{X})(Y-\bar{Y})\}}{n}$$
$$= \frac{\Sigma(X-\bar{X})^2}{n} + \frac{\Sigma(Y-\bar{Y})^2}{n} + \frac{2\Sigma(X-\bar{X})(Y-\bar{Y})}{n} = s^2_X + s^2_Y + 2\,\text{cov}_{(XY)}$$

If X and Y are independent, their covariance is zero. Hence, for independent variables

$$s^2_{(X+Y)} = s^2_X + s^2_Y$$

Rule 6. The mean of a constant is the constant itself. Symbolically

$$\bar{k} = k$$

Proof:

$$\bar{k} = \frac{\sum_{i=1}^{n} k}{n} = \frac{nk}{n} = k$$

Rule 7

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

Proof:

$$\Sigma x^2 = \Sigma(X - \bar{X})^2 = \Sigma(X^2 + \bar{X}^2 - 2\bar{X}X)$$
$$= \Sigma X^2 + n\bar{X}^2 - 2\bar{X}\Sigma X$$
$$= \Sigma X^2 + n\bar{X}^2 - 2n\bar{X}^2 = \Sigma X^2 - n\bar{X}^2$$

Rule 8

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

Proof:

$$\Sigma xy = \Sigma(X - \bar{X})(Y - \bar{Y}) = \Sigma(XY - \bar{X}Y - \bar{Y}X + \bar{X}\bar{Y})$$
$$= \Sigma XY - \bar{X}\Sigma Y - \bar{Y}\Sigma X + n\bar{X}\bar{Y}$$
$$= \Sigma XY - n\bar{X}\bar{Y} - n\bar{Y}\bar{X} + n\bar{X}\bar{Y} = \Sigma\bar{X}Y - n\bar{X}\bar{Y}$$

D.2. THE ALGEBRA OF EXPECTED VALUES

There are several rules for the algebraic manipulation of expressions involving expected values, which are used in several chapters. The most important of these rules are listed below.

Rule 1. The mathematical expectation of a sum (or a difference) of two random independent variables is the sum (or difference) of their individual expected values.

Let X and Y be two random variables with independent population distributions. We then may write symbolically

$$E(X \pm Y) = E(X) \pm E(Y)$$

Proof:

$$E(X+Y) = \sum_i \sum_j (X_i + Y_j) f(X_i, Y_j) = \sum_i \sum_j X_i f(X_i, Y_j) + \sum_i \sum_j Y_j f(X_i, Y_j)$$
$$= \sum_i X_i \sum_j f(X_i, Y_j) + \sum_j Y_j \Sigma f(X_i, X_y)$$

But $\sum_j f(X_i, Y_j) = f(X_i)$ = marginal probability of X_i

and $\sum_i f(X_i, Y_j) = f(Y_j)$ = marginal probability of Y_j

Therefore $E(X+Y) = \sum_i X_i f(X_i) + \sum_j Y_j f(Y_j) = E(X) + E(Y)$

Rule 2. The expected value of a constant is equal to that constant. Symbolically we may write

$$E(k) = k$$

Proof: $E(k) = \sum_i kf(k) = k\Sigma f(k) = k$

given $\Sigma f(k)$ = sum of probabilities = 1

Rule 3. The expected value of a random variable multiplied by a constant is equal to the constant times the expected value of the variable. Symbolically we may write this rule as follows

$$E(kX_i) = kE(X)$$

Proof: $E(kX_i) = \Sigma(kX_i) f(X_i) = k\,\Sigma X_i f(X_i) = k\,E(X)$

Rule 4. The variance of the sum of two variables is equal to the sum of the individual variances plus twice their covariance:

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(XY)$$

or

$$\sigma^2_{(X+Y)} = \sigma^2_X + \sigma^2_Y + 2\,\text{cov}(XY)$$

Proof:

$$\text{var}(X+Y) = E[(X+Y) - E(X+Y)]^2$$
$$= E\{[X - E(X)] + [Y - E(Y)]\}^2$$
$$= E[X - E(X)]^2 + E[Y - E(Y)]^2 + 2E\{[X - E(X)][Y - E(Y)]\}$$
$$= \text{var}(X) + \text{var}(Y) + 2\,\text{cov}(XY)$$

Rule 5. The expected value of the product of two *independent variables* is equal to the product of the expected values of the two variables:

$$E(XY) = E(X)\,E(Y)$$

Proof: $E(XY) = \sum_i \sum_j X_i Y_j f(X_i, Y_j)$

Since X and Y are independent their joint probability $f(X_i, Y_j)$ is equal to the product of their individual probabilities (see section F):

$$f(X_i, Y_j) = f(X_i) f(Y_j)$$

Therefore

$$E(XY) = \sum_i \sum_j X_i Y_j f(X_i) f(Y_j)$$
$$= [\sum_i X_i f(X_i)][\sum_j Y_j f(Y_j)]$$
$$= E(X)\,E(Y)$$

Rule 6. If X and Y are two independent random variables their covariance is equal to zero:

$$\text{cov}(XY) = 0 \qquad \text{for independent variables}$$

Proof:

$$\begin{aligned}\text{cov}(XY) &= E\{[X - E(X)]\,[Y - E(Y)]\} \\ &= E\{XY - YE(X) - XE(Y) + E(X)\,E(Y)\} \\ &= E\,(XY) - E(X)\,E(Y)\end{aligned}$$

By rule 5, for independent variables

$$E(XY) = E(X)\,E(Y)$$

Therefore

$$\text{cov}(XY) = E(X)\,E(Y) - E(X)\,E(Y) = 0$$

Rule 7. If X is a random variable and a and b are constants, then

$$\text{var}(aX + b) = a^2\,\text{var}(X)$$

Proof:

$$\begin{aligned}\text{var}(aX + b) &= E\{(aX + b) - E(aX + b)\}^2 \\ &= E\{(aX + b) - [aE(X) + b]\}^2 \\ &= E[aX + b - a\mu - b]^2 \\ &= a^2E(X - \mu)^2 = a^2\,\text{var}(X)\end{aligned}$$

Rule 8 The expected value of the sum of products of two variables is equal to the sum of the expected values of the products. Symbolically

$$E\left(\sum_{i=1}^{n} X_iY_i\right) = \sum_{i=1}^{n} [E(X_iY_i)]$$

Proof:

$$\begin{aligned}E\left(\sum_{i=1}^{n} X_iY_i\right) &= E(X_1\,Y_1 + X_2\,Y_2 + X_3\,Y_3 + \ldots + X_nY_n) \\ &= E(X_1\,Y_1) + E(X_2\,Y_2) + \ldots + E(X_nY_n) \\ &= \sum_{i=1}^{n} [E(X_iY_i)]\end{aligned}$$

Rule 9 The expected value of the squared sum of products of two variables is

$$E\left(\sum_{i=1}^{n} X_iY_i\right)^2 = E\left[\sum_{i=1}^{n} X_i^2Y_i^2 + 2\sum_{i\neq j} X_iX_jY_iY_j\right]$$

Proof:

$$\begin{aligned}E\left(\sum_{i=1}^{n} X_iY_i\right)^2 &= E[X_1\,Y_1 + X_2\,Y_2 + X_3X_3 + \ldots + X_nY_n]^2 \\ &= E(X_1^2\,Y_1^2 + X_2^2\,Y_2^2 + \ldots + X_n^2Y_n^2 + \\ &\quad + 2X_1\,X_2\,Y_1\,Y_2 + 2X_1\,X_3\,Y_1\,Y_3 + \ldots] \\ &= E\left[\sum_{i=1}^{n} X_i^2Y_i^2 + 2\sum_{i\neq j}^{n} X_iX_jY_iY_j\right]\end{aligned}$$

SECTION E

ELEMENTS OF STATISTICAL INFERENCE

Statistical inference may be of two kinds, *estimation* and *hypothesis testing*. Estimation is concerned with obtaining numerical values of the parameters from a sample. Hypothesis testing is concerned with passing a judgement on some assumption which we make (on the basis of economic theory or from any other source of information) about the true value of a population parameter. Both types of statistical inference utilise the information of a sample for drawing some conclusions about the parameters of the population, but each type of inference uses this information in different ways. In estimation we use some formulae in which we substitute the observations of the sample in order to obtain a numerical estimate of the population parameters (see below). In hypothesis testing we begin with some assumption about the true value of the population parameter, and then we use the information of the sample to compute a certain *test statistic* with which we will decide whether to accept or reject the assumption (hypothesis) which we made about the true population parameter.

E.1. ESTIMATION

The parameters of the populations of economic variables are unknown, since most of these populations include an infinite number of values. Even when the populations are finite, the number of values they include is very large and, although observable in theory, all of them are not known in practice. Estimation aims at the evaluation of the unknown basic parameters of the populations from the information of a sample. There are various methods of estimation; the most important are the following.

The method of moments, which involves the estimation of a population parameter from the corresponding sample statistic. Let us examine the estimation of the two basic population parameters, the mean μ and the variance σ_x^2, from the sample statistics $\overline{X}$ and s_x^2.

(a) The mean of the sample is an unbiased estimate of the population mean,

$$E(\overline{X}) = \mu$$

Proof

The population mean is $E(X_i) = \mu$

The sample mean is $\overline{X} = \sum_i^n \frac{X_i}{n} = \frac{1}{n}\sum_i^n X_i$

Taking expected values we find

$$E(\bar{X}) = \frac{1}{n}\sum_i^n E(X_i) = \frac{1}{n}\sum_i^n \mu = \frac{1}{n} n\mu = \mu$$

(b) The variance of the sample however is a biased estimate of the population variance, that is, $E(s_x^2) \neq \sigma_x^2$

Proof

The population variance is $E(X - \mu)^2 = \sigma_x^2$

The sample variance is $s_x^2 = \dfrac{\Sigma(X_i - \bar{X})^2}{n}$

Taking expected values we find

$$E(s_x^2) = E\left\{\frac{1}{n}\Sigma(X_i - \bar{X})^2\right\} = E\left\{\frac{1}{n}\Sigma[(X_i - \mu) - (\bar{X} - \mu)]^2\right\}$$

$$= E\left\{\frac{1}{n}\Sigma(X_i - \mu)^2 + \frac{1}{n}\Sigma(\bar{X} - \mu)^2 - 2(\bar{X} - \mu)\frac{1}{n}\Sigma(X_i - \mu)\right\}$$

$$= E\left\{\frac{1}{n}\Sigma(X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(\bar{X} - \mu)^2\right\}$$

$$= \frac{1}{n}\Sigma E(X_i - \mu)^2 - E(\bar{X} - \mu)^2$$

But $E(X_i - \mu) = \sigma_x^2$, and $E(\bar{X} - \mu)^2 = \sigma_{\bar{x}}^2 = \dfrac{\sigma_x^2}{n}$. (For a proof of this result see below, p. 553.) Therefore

$$E(s_x^2) = \frac{1}{n}\Sigma\sigma_x^2 - \frac{\sigma_x^2}{n} = \frac{n\sigma_x^2}{n} - \frac{\sigma_x^2}{n}$$

$$E(s_x^2) = \sigma_x^2 \quad \frac{n-1}{n}$$

Clearly $E(s_x^2) \neq \sigma_x^2$. However, we may obtain an unbiased estimate of the population variance from the sample variance as follows. We found

$$E(s_x^2) = \frac{n-1}{n}\sigma_x^2$$

Rearranging we get

$$\frac{n}{n-1}\, E(s_x^2) = \sigma_x^2$$

$$\frac{n}{n-1}\, E\left[\frac{\Sigma(X - \bar{X})^2}{n}\right] = \sigma_x^2$$

$$\frac{n}{n-1} \cdot \frac{1}{n}\Sigma E(X - \bar{X})^2 = \sigma_x^2$$

and

$$E\left[\frac{\Sigma(X - \bar{X})^2}{n-1}\right] = \sigma_x^2$$

Thus an unbiased estimator of the population variance from the sample observations may be obtained by the formula

$$\hat{\sigma}_x^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$$

Note the difference between the sample variance as a descriptive measure of the dispersion of the sample observations from their mean

$$s_x^2 = \frac{\Sigma(X_i - \bar{X})^2}{n}$$

and the unbiased estimate of the population variance obtained from the sample observations

$$\hat{\sigma}_x^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$$

If the sample is small the subtraction of the unity from the denominator makes a lot of difference to the estimate $\hat{\sigma}_x^2$. However, if n is large the sample variance is a satisfactory approximation to the population variance (the bias involved in s_x^2 is unimportant for large samples, usually for $n > 30$).

The other two important methods of estimation are the *least squares method* and the *maximum likelihood method* which are explained in detail in Chapters 4 and 18. In this Appendix we will examine systematically the procedure of hypothesis testing. For this it is necessary to develop the basic concepts and theorems of sampling theory.

E.2. ELEMENTS OF SAMPLING THEORY

Sampling theory is concerned with establishing relationships between the distributions of populations and distributions of sample statistics (for example the relationship between the population distribution and the distribution of the sample mean, or the relationship between the population distribution and the distribution of the sample variance).

The basis of sampling theory is the normal distribution, which we will examine first. We will next introduce four basic distributions: (a) the Standard Normal distribution, (b) the χ^2 distribution, (c) the t distribution, (d) the F distribution. These are distributions of corresponding statistics, the z, t, χ^2 and F statistics. These statistics are formulae (expressions) which transform the units of the original population parameters into units of the sampling distribution of these statistics. The probabilities for these distributions have been computed by various writers and have been tabulated. Thus by using the appropriate transformation procedure one can find the probabilities of the original distribution of X or of the sampling distributions indirectly instead of estimating these probabilities directly from the original or sampling distributions which usually involve highly complex expressions.

THE NORMAL DISTRIBUTION

A normal distribution is a bell-shaped curve which extends indefinitely in both directions. It is symmetrical round the mean of the variable, whose values are measured on the horizontal axis (figure 7).

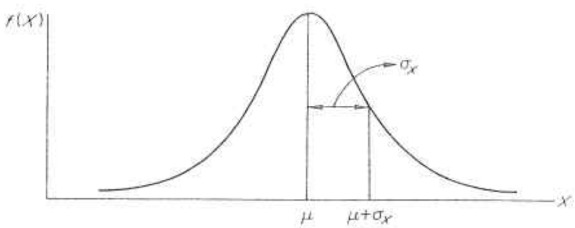


Figure 7

The vertical axis depicts the value of the probability density function $f(X_i)$.

The equation of the normal curve is

$$f(X_i) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \cdot \exp\left\{-\frac{1}{2}\left(\frac{X_i - \mu}{\sigma_x}\right)^2\right\}$$

where $f(X_i)$ = the probability of X assuming the value X_i
μ = the mean of the variable X
σ_x = the standard deviation of X
π = 3·14 = the ratio of the circumference of the circle to its diameter
e = 2·71828, the base of the natural logarithms

From the above equation it is obvious that a normal curve is completely determined if we know its mean μ and standard deviation σ_x, since all the other coefficients appearing in the equation are known constants.

Any normal curve will be bell-shaped and symmetrical round its mean, but its actual form (height and width) will differ according to the value of its mean u and its standard deviation σ_x. In figure 8 both curves are normal, they both have the same mean $\mu = 5$, but the inner curve has a smaller standard deviation than the outer curve ($\sigma_{x_1} < \sigma_{x_2}$). In figure 9 two normal curves are depicted with different means ($\mu_1 < \mu_2$) and different standard deviations ($\sigma_{x_1} < \sigma_{x_2}$). If we know the mean μ and the standard deviation σ_x, we can draw the normal curve, by assigning various values to X and computing the respective value of $f(X)$.

The areas under the normal curve are probabilities of X assuming various values. The total area under the curve is equal to unity, because it is the sum of all the probabilities of X assuming all its possible values (from $-\infty$ to $+\infty$ in the case of the normal curve).

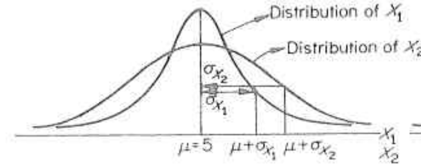


Figure 8

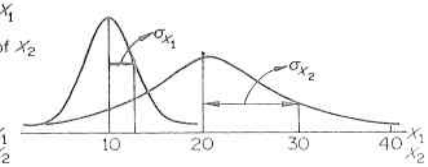


Figure 9

THE STANDARD NORMAL DISTRIBUTION (OR GAUSS DISTRIBUTION)

It should be obvious that the probability of X_i will be different according to the form of the curve. In figure 10 we see that the probability of X taking a value between 10 and 15 ($P\{10 < X < 15\}$) is bigger for the distribution A, as compared to the distribution B. Consequently in order to find the probability

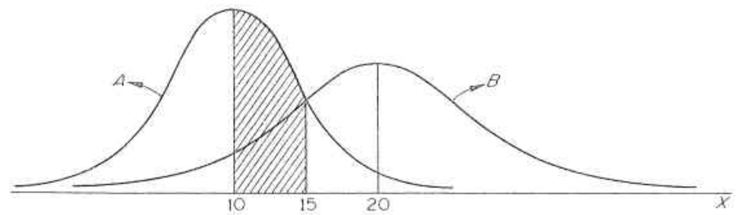


Figure 10

of X assuming a value between X_a and X_b we should compute a separate table for each normal curve, giving the relevant probabilities according to the values of the mean μ, and standard deviation, σ_x, of the variable. Fortunately this is not necessary because we can transform any normal curve into a standard form, which is called *Standard Normal Distribution*, or Gauss Distribution. The Standard Normal Distribution is the probability distribution of a variable Z which has a normal distribution with zero mean and unit variance

$$Z \sim N(0, 1)$$

The probabilities of the various values of Z have been tabulated by Gauss and are shown in Table 1 of Appendix IV.

The standardisation procedure may be outlined as follows. If a variable X has a normal distribution with mean μ and variance σ_x^2, then the statistic

$$Z_i = \frac{X_i - \mu}{\sigma_x} \sim N(0, 1)$$

where X_i = the values of the variable X, whose distribution we want to convert into units of the Standard Normal Curve.
μ = the mean of the distribution of X.
σ_x = the standard deviation of X.

The transformation of any normal curve into the Standard Normal Curve is a very simple operation, through which we actually change the scales of measurement of the variable X: we transform the units of measurement of X into standard Z units. In figure 11, on the horizontal axis of which we measure the variable X in x-units, we see that when $X = \mu$, then $Z = 0$, because

$$Z_i = \frac{X_i - \mu}{\sigma_x} = \frac{\mu - \mu}{\sigma_x} = 0$$

Similarly, if $X = \mu + \sigma_x$, then

$$Z_i = \frac{X_i - \mu}{\sigma_x} = \frac{(\mu + \sigma_x) - \mu}{\sigma_x} = 1$$

In the same way we see that when $X = \mu + 2\sigma_x$ on the X-scale, the corresponding value on the Z-scale is 2,

$$Z_i = \frac{X_i - \mu}{\sigma_x} = \frac{(\mu + 2\sigma_x) - \mu}{\sigma_x} = 2$$

and so on.

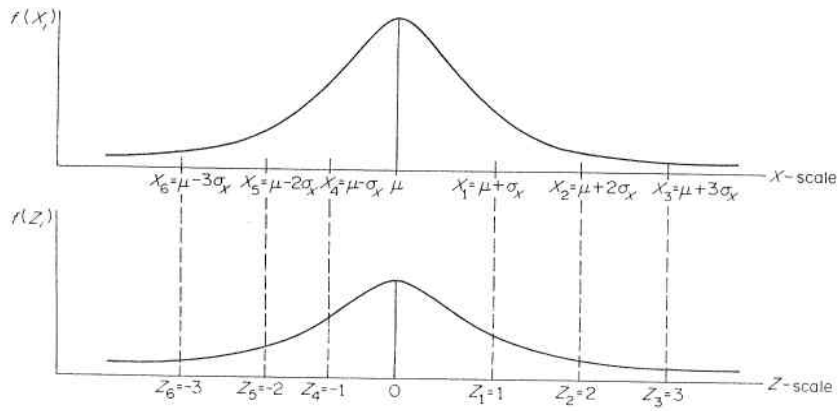


Figure 11

We will show that the probability of $X \sim N(\mu, \sigma_x^2)$ assuming any value between X_1 and X_2 is equal to the probability of Z assuming any value between Z_1 and Z_2 where

$$Z_1 = \frac{X_1 - \mu}{\sigma_x} \quad \text{and} \quad Z_2 = \frac{X_2 - \mu}{\sigma_x}$$

Proof. The general formula for the transformation of X units into Z units is

$$Z_i = \frac{X_i - \mu}{\sigma_x}$$

Now from $Z = \dfrac{X - \mu}{\sigma_x}$ $\longrightarrow$ we obtain $X = \sigma_x Z + \mu$

similarly $Z_1 = \dfrac{X_1 - \mu}{\sigma_x}$ $\longrightarrow$ yields $X_1 = \sigma_x Z_1 + \mu$

and $Z_2 = \dfrac{X_2 - \mu}{\sigma_x}$ $\longrightarrow$ gives $X_2 = \sigma_x Z_2 + \mu$

We want to prove that

$$P(X_1 < X < X_2) = P(Z_1 < Z < Z_2)$$

Substituting X, X_1 and X_2 in this expression we find

$$P(X_1 < X < X_2) = P\{(\sigma_x Z_1 + \mu) < (\sigma_x Z + \mu) < (\sigma_x Z_2 + \mu)\}$$

Subtracting from all three terms μ we obtain

$$P(X_1 < X < X_2) = P(\sigma_x Z_1 < \sigma_x Z < \sigma_x Z_2)$$

Finally dividing all terms by σ_x we find

$$P(X_1 < X < X_2) = P(Z_1 < Z < Z_2)$$

where

$$Z_i = \frac{X_i - \mu}{\sigma_x}$$

Thus the probability that X lies between X_1 and X_2 is equal to the probability that the standard normal (Z) variable lies between

$$Z_1 = \frac{X_1 - \mu}{\sigma_x} \quad \text{and} \quad Z_2 = \frac{X_2 - \mu}{\sigma_x}$$

With the above transformation we avoid the tedious task of constructing separate tables of normal curve areas for each pair of values of μ and σ_x. We need only the table giving the areas (probabilities) under the Standard Normal Curve. The Normal Curve Table is reproduced on page 579. If in a given problem we want to determine an area under a normal curve whose mean and standard deviation are μ and σ_x, we have only to change the X's to Z's and then use the Standard Normal Table.

Examples for the use of the Standard Normal Table

Example 1. The Standard Normal Curve Table shows the area (probability) to the right of any particular value of Z. For example if $Z = 1{\cdot}96$ the area to the right of this value is 0·025 (figure 12). This is interpreted as follows: the probability of Z_i assuming any value greater than 1·96 is 0·025 (or 2·5 per cent). Symbolically

$$P\{Z_i > 1{\cdot}96\} = 0{\cdot}025$$

Example 2. The Standard Normal Curve Table does not contain areas (probabilities), corresponding to negative values of Z. Since the normal curve is

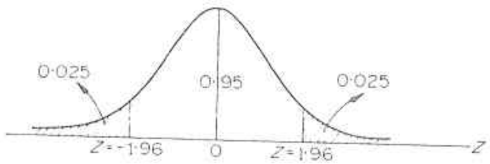


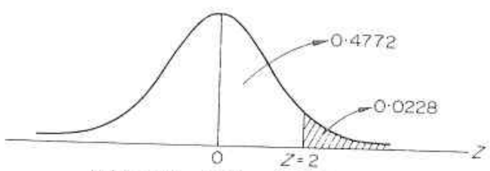
Figure 12

symmetrical, the area for negative values is the same as the area for positive values. For example the area to the left of $Z = -1{\cdot}96$ is $0{\cdot}025$.

Example 3. Suppose we want the area (probability) between $Z = 0$ and any positive value Z_i. We subtract the corresponding (to this Z_i) area from 0·50, so that the area between $Z = 0$ and any positive value Z_i is 0·50 minus the tabular value corresponding to Z_i. For instance the area $0 < Z_i < 2$ is found as follows.

(a) The area to the right of the zero mean is 0·50.
(b) The area to the right of $Z = 2$ is 0·0228
(c) Therefore the area between $Z = 0$ and $Z = 2$ is

$$0{\cdot}5000 - 0{\cdot}0228 = 0{\cdot}4772.$$



$P\{0 < Z_i < 2\} = 0{\cdot}4772$

Figure 13

This reads: the probability of Z assuming a value bigger than 0 but smaller than 2 is 47·7 per cent.

Example 4. Assume $X \sim N(\mu, \sigma_x^2) \sim N(10, 4)$. We want to find the probability that X will assume a value lying between $X_1 = 8$ and $X_2 = 12$.

We work as follows

$$Z_1 = \frac{X_1 - \mu}{\sigma_x} = \frac{8 - 10}{2} = -1$$

$$Z_2 = \frac{X_2 - \mu}{\sigma_x} = \frac{12 - 10}{2} = +1$$

We know that $P(X_1 < X < X_2) = P(Z_1 < Z < Z_2)$. From the Standard Normal Table (p. 579) we find the probabilities

$$P(Z > 1) = 0{\cdot}159 \quad \text{and} \quad P(Z < -1) = 0{\cdot}159$$

Therefore $P(-1 < Z < 1) = 1 - (0{\cdot}159 + 0{\cdot}159) = 0{\cdot}682$
Therefore $P(8 < X < 10) = P(-1 < Z < 1) = 0{\cdot}682$

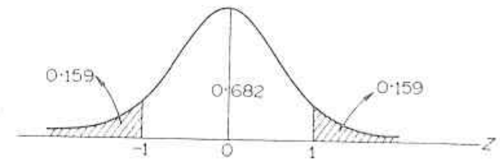


Figure 14

SOME BASIC RESULTS OF THE STANDARD NORMAL CURVE

From the Standard Normal Curve Table we see that:

the area between $Z = 0$ and $Z = 1$ is 0·3413;
the area between $Z = 0$ and $Z = 1{\cdot}96$ is 0·475;
the area between $Z = 0$ and $Z = 3$ is 0·498.

These results are shown in figure 15.

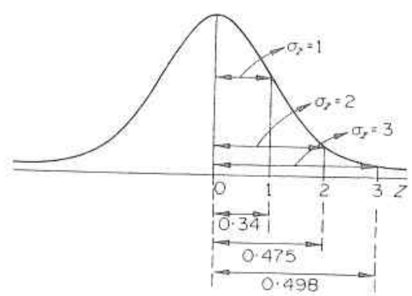


Figure 15

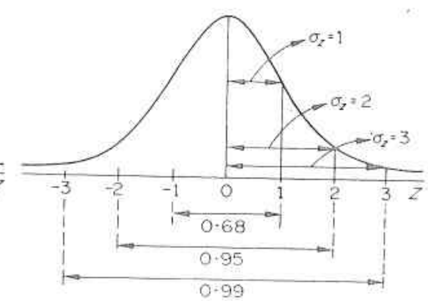


Figure 16

Since the Z distribution is symmetrical with unit standard deviation we can state the following very important results (shown in figure 16).

(a) The range 0 ± 1 contains 68 per cent of the values of Z (or, the probability of Z taking a value between -1 and $+1$ is 0·68.)
(b) The range $0 \pm 1{\cdot}96$ contains 95 per cent of the values of Z (or, the probability of Z taking a value between $-1{\cdot}96$ and $+1{\cdot}96$ is 0·95).
(c) The range 0 ± 3 contains 99 per cent of the values of Z (or, the probability of Z assuming a value between -3 and $+3$ is 0·99).

The above results apply to any normal distribution. This is easily understood if we look back at the transformation formula $Z_i = (X_i - \mu)/\sigma_x$ and the explanation of the change of scales of units.

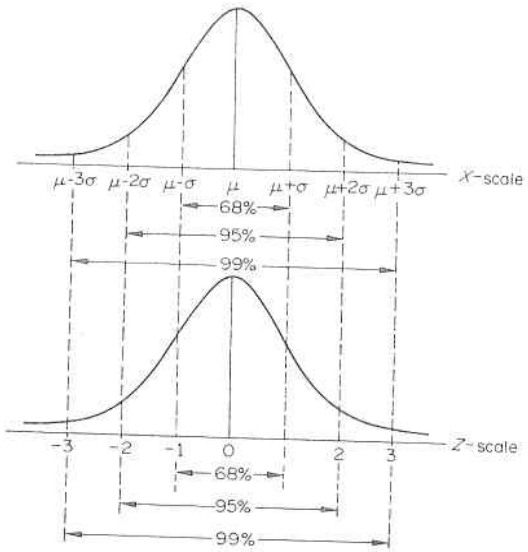


Figure 17

From figure 17 it is clear that:

(i) the range $\mu \pm \sigma_x$ contains 68 per cent of the values of X (or, there is 68 per cent probability that X will take a value between $\mu \pm \sigma_x$);

(ii) the range $\mu \pm 2\sigma_x$ contains 95 per cent of the values of X (or there is 95 per cent probability that X will assume a value between $\mu \pm 2\sigma_x$);

(iii) the range $\mu \pm 3\sigma_x$ contains 99 per cent of the values of X (or there is 99 per cent probability that X will assume a value between $\mu \pm 3\sigma_x$).

CONDITIONS FOR THE APPLICATION OF THE Z STATISTIC

The Z transformation statistic is applicable only in the following cases. *Firstly*, if the population variance σ_x^2 is known, irrespective of the size of the sample. *Secondly*, if the population variance is unknown but the sample is large ($n > 30$); because in this case the sample estimate of the unknown population variance s_x^2 is a good approximation to σ_x^2 (see page 545).

If none of these conditions is fulfilled we cannot use the Z statistic and the standard normal distribution for conducting tests of significance. However, if the parent population is normal (and the sample small, $n < 30$) we can apply another transformation procedure, based on Student's t distribution, which is examined in a subsequent section.

BASIC THEOREMS OF RANDOM SAMPLING: SAMPLING DISTRIBUTIONS

The basic tool of hypothesis testing is the sampling distribution, that is the probability distribution of the sample mean. The concept of the sampling distribution of the sample mean may be explained as follows.

We assume that we draw (hypothetically) an infinite number of samples

each of size n from the population of values of the variable X, which has a normal distribution with mean μ and variance σ_x^2.

This step is known as *hypothetical repeated sampling*.

Next we assume that for each (hypothetical) sample we compute the sample mean $\bar{X}_i$.

Thus we have an infinite number of sample means, one for each sample, which form the (hypothetical) population of sample means. The distribution of the sample mean is called the *sampling distribution*, and its parameters are designated by

$\bar{\bar{X}}$ = mean of the sampling distribution
$\sigma_{\bar{x}}^2$ = variance of the sampling distribution
$\sigma_{\bar{x}}$ = standard deviation of the sampling distribution.

Statisticians have established the following two important theorems which relate the true population parameters to the parameters of the sampling distribution.

Theorem 1

If we have a population of a variable X that has a normal distribution with mean μ and variance σ_x^2, and if repeated random samples, all of size n, are taken from this population and for each sample we compute the sample mean $\bar{X}_i$, the theoretical distribution of the sample means (the sampling distribution of the mean) will be normal with the same mean of the population, μ, and variance equal to σ_n^2/n.

Symbolically, if $\quad X_i \sim N(\mu, \sigma_x^2)$

then $\quad \bar{X}_i \sim N(\mu, \sigma_{\bar{x}}^2 = \sigma_x^2/n)$

We will not attempt to establish the normality of the sampling distribution since this requires the use of complicated expressions. Normality will be assumed here. It is important to stress that normality of the sampling distribution is crucial, because it is required for the use of the Standard Normal Curve Table to situations involving sample statistics (see below).

We will derive the mean and the variance of the sampling distribution.

(1) The mean of the sampling distribution is equal to the population mean

$$\bar{\bar{X}} = E(\bar{X}_i) = \mu$$

Proof

Each of the n observations of the sample may be considered as n distinct variables, each possessing the same distribution as X, that is

$$\begin{aligned} X_1 &\sim N(\mu, \sigma_x^2) \\ X_2 &\sim N(\mu, \sigma_x^2) \\ &\vdots \\ X_n &\sim N(\mu, \sigma_x^2) \end{aligned}$$

In other words we may think of X_1 as representing all the possible values of X that can be observed when drawing the first observation of the sample, X_2 as representing all possible values of X when drawing the second observation of the sample, and so on.

Now we want to prove that the mean of the distribution of the sample means is equal to the population mean. By the definition of the mean

$$E(\bar{X}_i) = E\left[\frac{1}{n}\Sigma X_i\right] = E\left[\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right]$$

$$= \frac{1}{n}\{E(X_1) + E(X_2) + \ldots + E(X_n)\}$$

$$= \frac{1}{n}(\mu + \mu + \mu + \ldots + \mu) = \frac{1}{n}\sum_i^n \mu = \frac{1}{n}(n\mu) = \mu$$

(2) The variance of the sampling distribution is equal to the variance of the population divided by the size of the samples, n

$$\sigma_{\bar{X}}^2 = \sigma_X^2/n$$

Proof

By definition

$$\text{var}(\bar{X}) = \sigma_{\bar{X}}^2 = \text{var}\left(\frac{1}{n}\Sigma X_i\right) = \frac{1}{n^2}\Sigma \text{ var } X_i$$

$$= \frac{1}{n^2}\{\text{var}(X_1) + \text{var}(X_2) + \ldots + \text{var}(X_n)\}$$

$$= \frac{1}{n^2}\sum_i^n \sigma_X^2 = \frac{1}{n^2} n\sigma_X^2 = \sigma_X^2/n$$

Clearly the standard deviation of the sampling distribution decreases as n increases. This means that when n is large, a sample mean will be a more reliable estimate of μ. When n becomes large we have more information and the sample mean can be expected to be closer to the mean of the population.

We said that the normality of the sampling distribution is crucial for statistical inference. We also said that the normality of the sampling distribution is secured if the parent population (from which the samples are drawn) has a normal distribution. However, for most populations we cannot be sure about the exact form of their distribution and therefore the above theorem may not be applicable. Yet we can derive similar results for the sampling distribution of $\bar{X}_i$ even when the population, from which the samples are drawn, is not normal. This we can do by making use of a second theorem, which is known as Central Limit Theorem.

Theorem 2. Central Limit Theorem

If the size of the sample is large ($n \to \infty$) the theoretical sampling distribution of $\bar{X}_i$ will be close to a normal curve regardless of the shape of the distribution of the basic (parent) population. Symbolically

if $$X \nsim N(\mu, \sigma_x^2)$$

then $$\bar{X}_i \to N(\mu, \sigma_x^2/n) \qquad \text{for } n \to \infty$$

The Central Limit Theorem is more powerful than may appear at first sight, because although the approximation is derived for $n \to \infty$ a good approximation is generally obtained for quite small values of $n (n > 30)$.

The standardisation procedure with the Z statistic applies for the distribution of sample means. That is, we can calculate the probabilities of getting various values of $\bar{X}_i$ by using the Z transformation formula. For example suppose we have a sample of size 36 from a population, with $\mu = 48$ and $\sigma_x = 12$ and we want to find the probability of $\bar{X}$ lying between 49 and 50 (figure 18). We compute the corresponding Z values for $\bar{X}_1 = 49$ and $\bar{X}_2 = 50$.

$$Z_1 = \frac{49 - 48}{12/\sqrt{36}} = 0{\cdot}50$$

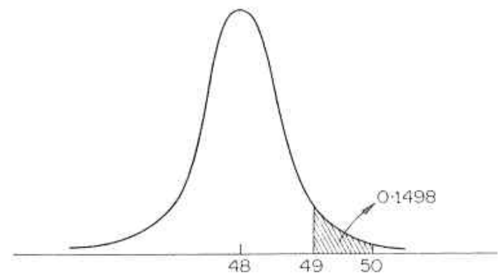


Figure 18

From the Standard Normal Curve Table (p. 659) we find

$$P(Z > 0{\cdot}50) = 0{\cdot}3085$$

Similarly,

$$Z_2 = \frac{50 - 48}{12/\sqrt{36}} = 1{\cdot}00$$

and from the Standard Normal Curve Table we see that

$$P(Z < 1) = 0{\cdot}1587$$

Thus $P\{49 < \bar{X} < 50\} = P\{0{\cdot}50 < Z < 1{\cdot}00\} = 0{\cdot}1498$

Summary of sampling distribution theorems

Theorem I. If $X_i \sim N(\mu, \sigma_x^2)$, then $Z = (X_i - \mu)/\sigma_x \sim N(0, 1)$

If $\bar{X}_i \sim N(\mu, \sigma_{\bar{x}}^2 = \sigma_x^2/n)$, then $Z = \dfrac{\bar{X}_i - \mu}{\sigma_x/\sqrt{n}} \sim N(0, 1)$

Theorem II. Central Limit Theorem. If $X_i \nsim N(\mu, \sigma_x^2)$ then

$$\overline{X}_i \to N(\mu, \sigma_x^2/n) \quad \text{and} \quad Z_i \to \frac{\overline{X}_i - \mu}{\sigma_x/\sqrt{n}} \to N(0, 1) \qquad \text{as } n \to \infty$$

(In practice the approximation of the sampling distribution to a normal shape is good for $n > 30$.)

In actual applied problems we do not have many samples; we usually work with one sample and hence we have one sample mean only. However, we consider that this sample mean has been drawn from a hypothetical repeated sampling process, which gives a sampling distribution with the above characteristics (mean μ, and standard deviation $\sigma_{\overline{x}} = \sigma_x/\sqrt{n}$).

We said that the Z statistic and the use of the Standard Normal Curve is applied only when the population variance σ_x^2 is known, or when the population variance is not known but we have its estimate $\hat{\sigma}_x^2$ from a big *sample* ($n > 30$), because when the sample is large the sample estimate $\hat{\sigma}_x^2$ is a good estimate of σ_x^2. However, when the sample is small ($n < 30$) the sampling distribution (of $\overline{X}$'s) will not be normal. In this case we can apply another transformation, based on the Student's t statistic, $t = (X_i - \mu)/\{\hat{\sigma}/\sqrt{n}\}$ (with n − k *degrees of freedom*) provided that the parent population is normal. For the formal exposition of the t transformation we must first examine the meaning of another theoretical distribution, namely the χ^2 distribution.

THE CHI-SQUARE (χ^2) DISTRIBUTION

If we have a set of normal and independent variables, $X_1, X_2, \ldots X_\nu$, and we normalise them by taking their respective standard normal values

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1} \sim N(0, 1) \qquad Z_2 = \frac{X_2 - \mu_2}{\sigma_2} \sim N(0, 1) \ldots Z_\nu = \frac{X_\nu - \mu_\nu}{\sigma_\nu}$$

the sum of the squares of the normalised variables has a Chi-square (χ^2) distribution

$$\chi^2 = \Sigma Z_i^2 = \Sigma \left(\frac{X_i - \mu_i}{\sigma_{x_i}}\right)^2$$

with ν degrees of freedom. The number of degrees of freedom is equal to the number of independent variables. The chi-square distribution is skewed to the right, starts from the origin and extends to $+\infty$ to the right tail (figure 19). As n increases the χ^2 distribution becomes more and more symmetric.

The mean and variance of a χ^2 distribution with ν degrees of freedom are

$$E(\chi^2) = n$$

$$\text{var}(\chi^2) = 2n$$

The chi-square distribution is tabulated for up to 30 degrees of freedom. The

χ^2 table is cumulative: it gives the probability of χ^2 assuming a value higher than a certain figure, given the number of degrees of freedom and the level of significance. The χ^2 table is reproduced on page 661. The table includes only values of χ^2 above which we find 5, 2·5, 1 and 0·5 per cent of the area under the curve. Symbolically we write $\chi^2_{0\cdot025}$ for the value of χ^2 to the right of which

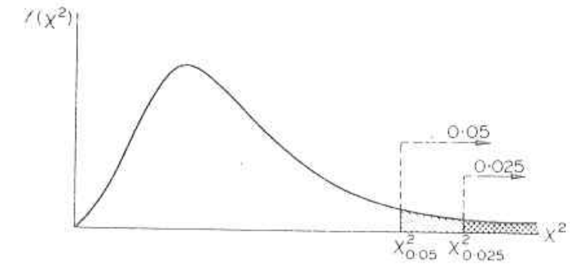


Figure 19

lies 2·5 per cent of the total area under the curve. $\chi^2_{0\cdot05}$ for the value of χ^2 to the right of which lies 5 per cent of the total area under the curve, and so on. Thus if the degrees of freedom are 10 we see from table 3 of Appendix III that $\chi^2_{0\cdot05} = 18\cdot3$. This means that the probability is 5 per cent that χ^2 will assume a value higher than 18·3 given that we have 10 degrees of freedom,

$$P\{18\cdot3 < \chi^2 < \infty\} = 0\cdot05 \qquad (\text{for } \nu = 10)$$

Apart from its use in forming the t statistic, the Chi-square distribution has many other applications. (See T. Yamane, *Statistics*, pp. 613–41.)

THE STUDENT'S t DISTRIBUTION

If a variable Z_i has a standard normal distribution with zero mean and unit standard deviation, $Z_i \sim N(0, 1)$, and another variable V^2 has an independent χ^2 distribution with ν degrees of freedom, then the quantity $t = Z\sqrt{\nu}/V$ has a Student's t distribution with ν degrees of freedom.

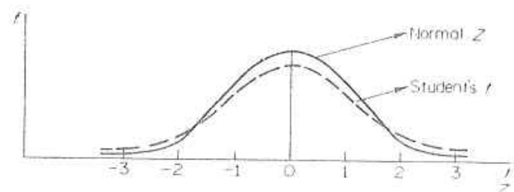


Figure 20

The characteristics of the t distribution may be summarised as follows.

(1) The t distribution is a bell-shaped distribution symmetric about zero (with zero mean). In general it has a variance greater than 1, but the variance approaches 1 as the sample size increases.

(2) The range of values of t is $-\infty < t < +\infty$.

(3) The t distribution is flatter than the normal distribution. This means that the area at the tails is larger for the t distribution than for the standard normal distribution (figure 20).

(4) As the sample size (n) becomes larger, the t distribution approaches the standard normal distribution. In fact if $n > 30$ one makes a very small error if one decides to use Z instead of t.

(5) The t distribution depends on the degrees of freedom, that is we need to know the degrees of freedom ν to obtain probabilities from the t table. The t table is reproduced on page 660. It differs from the Standard Normal Curve Table in that it includes the degrees of freedom. The table lists the values of t to the right of which we find, e.g., 10, 5, 2·5, 1 and 0·5 per cent of the area under the curve. Symbolically, we shall write $t_{0.025}$ for the value of t to the right of which the area under the curve is 2·5 per cent, $t_{0.01}$ for the value of t to the right of which the area under the curve is 1 per cent of the total, and so forth. For example assume we are given the degrees of freedom $\nu = 15$. We can find from the t table the values of t that correspond to various probabilities. (a) For $\nu = 15$, $t_{0.025} = 2{\cdot}131$, and since the t distribution is symmetrical the value $-t_{0.025} = -2{\cdot}131$, (b) For $\nu = 15$, $t_{0.05} = 1{\cdot}753$ and $-t_{0.05} = -1{\cdot}753$. These results may be stated as follows

$$P\{-2{\cdot}131 < t < 2{\cdot}131\} = 0{\cdot}95 \qquad \text{with 15 degrees of freedom}$$

$$P\{-1{\cdot}753 < t < 1{\cdot}753\} = 0{\cdot}90 \qquad \text{with 15 degrees of freedom}$$

The same results are shown graphically in figure 21.

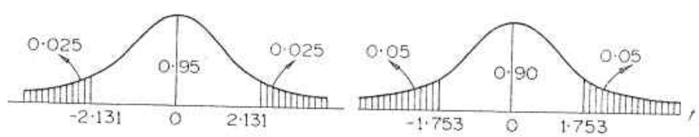


Figure 21

The transformation of the units of a normal variable X_i and its sampling distribution $\bar{X}_i$ into t units is implemented by applying the formulae

$$t = \frac{X_i - \mu}{\hat{\sigma}_x} \quad \text{and} \quad t = \frac{\bar{X}_i - \mu}{\hat{\sigma}_x/\sqrt{n}}$$

where $\hat{\sigma}_x$ is the unbiased estimate of the unknown population variance from a small sample ($n < 30$)

$$\hat{\sigma}_x^2 = \frac{\Sigma(X_i - \bar{X})^2}{n - 1}$$

The t transformation is appropriate if the parent population is normal with unknown variance and the sample with which we are working is small ($n < 30$).

This is so because, in forming the t statistic, $t = (Z\sqrt{\nu})/V$, the true variance σ_x^2 of the population is eliminated and we are left with a formula which includes the unbiased estimate of the population variance $\hat{\sigma}_x^2$.

Let us examine how σ_x^2 is eliminated from the t statistic.

(1) From the theory developed on p. 547 we know that

if
$$X_i \sim N(\mu, \sigma_x^2)$$

then
$$Z_i = \frac{X_i - \mu}{\sigma_x} \sim N(0, 1)$$

(2) It can be shown that if X is normally distributed with variance σ_x^2, and s_x^2 is the sample variance, $\dfrac{\Sigma(X_i - \bar{X})^2}{n}$, based on a random sample of size n, then the ratio $\dfrac{ns_x^2}{\sigma_x^2}$ has a χ^2 distribution with $n - 1$ degrees of freedom. (See P. G. Hoel, *Introduction to Mathematical Statistics*, Wiley, 1954, pp. 218–19.)

Let

$$V^2 = \frac{ns_x^2}{\sigma_x^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_x}\right)^2 \sim \chi^2 \qquad \text{with } \nu = n - 1 \text{ degrees of freedom}$$

(3) It can be shown that Z_i and V^2 have independent distributions. (See P. G. Hoel, op. cit.) Hence we can substitute for Z_i and V in the expression of the t statistic and obtain

$$t = \frac{Z\sqrt{\nu}}{V} = \frac{Z}{\sqrt{V^2/\nu}}$$

$$= \frac{(X_i - \mu)/\sigma_x}{\sqrt{\dfrac{1}{n-1}\Sigma\left(\dfrac{X_i - \bar{X}}{\sigma_x}\right)^2}} = \frac{X_i - \mu}{\sqrt{\dfrac{\Sigma(X_i - \bar{X})^2}{n - 1}}}$$

But the expression in the denominator is the unbiased estimate of the population variance $\hat{\sigma}_x^2$.

Hence
$$t = \frac{X_i - \mu}{\hat{\sigma}_x}$$

Thus the unknown population variance σ_x^2 has been eliminated and we are left with the sample estimate $\hat{\sigma}_x^2$ which can be computed from the observations of the sample with the formula

$$\hat{\sigma}_x^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} \qquad \text{(see page 545)}$$

In summary

(1) If
$$X \sim N(\mu, \sigma_x^2)$$
$$\bar{X} \sim N(\mu, \sigma_x^2/n)$$

(2) If σ_x^2 is known, or if it is unknown but $n > 30$, then the appropriate transformation is

$$Z_i = \frac{X_i - \mu}{\sigma_x} \qquad \text{for } X_i\text{'s}$$

and

$$Z_i = \frac{\overline{X}_i - \mu}{\sigma_x/\sqrt{n}} \qquad \text{for } \overline{X}_i\text{'s}$$

(3) If σ_x^2 is unknown and the sample is small ($n < 30$) the appropriate transformation is

$$t = \frac{X_i - \mu}{\hat{\sigma}_x} \qquad \text{for } X_i\text{'s}$$

and

$$t = \frac{\overline{X}_i - \mu}{\hat{\sigma}_x/\sqrt{n}} \qquad \text{for } \overline{X}_i\text{'s}$$

In this case the parent population must be normal.

THE F DISTRIBUTION

If two variables have *independent* chi-square distributions, χ_1^2 and χ_2^2, with ν_1 and ν_2 degrees of freedom respectively, the statistic

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$$

has an F distribution with ν_1 and ν_2 degrees of freedom.

The F statistic usually involves the ratio of two independent estimates of variances, and the F distribution is used to test the equality of these estimates. For this reason the F statistic is often called the *variance ratio*

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \text{variance ratio}$$

with ν_1 and ν_2 degrees of freedom.

The F distribution has a skewed shape (figure 22). The value of F is always positive. The range of values of F is $0 \leqslant F \leqslant +\infty$. The values of the F distribution

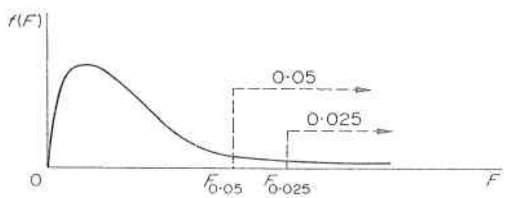


Figure 22

with various degrees of freedom (and various levels of significance) have been tabulated by Dr Snedecor and are reproduced on pp. 663–4. The degrees of freedom ν_1 and ν_2, depend on the way in which we obtain the estimates of the two variances appearing in the numerator and the denominator of the F ratio. The F table gives the probabilities of the right-hand tail. Given that the F

distribution is not symmetrical, the values of the left-hand tail cannot be directly deduced from the regular F table. To avoid complicated calculations statisticians have adopted the following convention. For a two-tailed test the F ratio is always evaluated with the larger estimate of the variance in the numerator and the smaller estimate in the denominator. (See Bugg, Henderson *et al.*, *Statistical Methods in the Social Sciences*, North-Holland, Amsterdam, 1968, p. 282.) With this convention the F ratio is always greater than unity. If the two variance estimates are close to each other their ratio will approach the value of one. The greater the difference between the two variances the greater the value of the F ratio. Thus, in general, high values of F suggest that the difference between the two estimates is significant. However, when conducting a *two-tail test* we must halve the value of our level of significance in looking at the regular F table. For example if we choose the 5 per cent level of significance for a two-tail test we take the value $F_{0.025}$ (with the appropriate degrees of freedom) as our critical value of F.

There is a formal relationship between the t and the F statistics as applied in regression analysis, which is explained in Chapter 8.

E.3. HYPOTHESIS TESTING

The procedure for testing a hypothesis concerning the value of population parameters includes the following steps.

(1) Formulate the null and alternative hypotheses.
(2) Choose the level of significance of the test.
(3) Choose the location of the critical region.
(4) Choose the appropriate test statistic (for example Z, t, χ^2, F) and find from the relevant tables the critical value(s) of the chosen statistic, that is the value(s) that defines the boundary of the critical region.
(5) Compute, from the sample observations, the observed value (or sample value, or empirical value) of the chosen statistic, using the relevant formula (for example $Z^* = (X - \mu)/\sigma_x$, or $t^* = (X - \mu)/\hat{\sigma}_x$).
(6) Compare the sample value of the chosen statistic with the theoretical (tabular) value(s) that define the critical region. If the observed value of the statistic falls in the critical region we reject the null hypothesis. Otherwise we accept the null hypothesis.

We will examine the above stages of hypothesis testing in some detail.

Step 1. Formulate the Null Hypothesis (H_0) *and the Alternative Hypothesis* (H_1)

We said that the aim of statistical inference is to draw conclusions about the population parameters from the sample statistics. In econometrics, using a set of observations, we obtain estimates of the parameters of economic relationships. We next wish to test their statistical reliability, that is to apply some rule which will enable us to decide whether to accept our estimate or to reject it. To make such a decision the best way is to compare the estimate with the true value of the population parameter. However, the population parameter is unknown. Under these circumstances how are we going to make the decision whether to accept or reject the sample estimate given that we do not have the appropriate yardstick (that is the true population parameter) for making the comparison required? To bypass this difficulty we make some assumption about the value of the true population parameter and use our sample estimate in order to decide whether our assumption is acceptable or not. *A hypothesis is an assumption which we make about a population parameter.* The hypothesis which we wish to test (on the basis of the evidence of our sample estimate) is called the *null hypothesis*, because it implies that there is no difference between the true parameter and the hypothesised value, or the difference between the true value and the hypothesised value is nil. Symbolically $H_0: \mu = \mu_0$.

In most applications it is difficult to hypothesise any special value for the true population parameter. What is worse, we may find that a very large number of hypothetical values are compatible with our sample estimate and in this case we run into the problem of choosing among these possible hypotheses. To avoid these problems it has become customary in econometrics to make the hypothesis that the true population parameter is equal to zero. That is the null hypothesis typically takes the form

$$H_0: \mu = 0$$

or, in the case of parameters of economic relationships,

$$H_0: b_i = 0$$

The *alternative hypothesis* (H_1) is an alternative assumption about the population parameter, a counter proposition to the null hypothesis. It is conventionally denoted by H_1 and may take one of the following forms

(a) $H_1: \mu \neq \mu_0$ (or $H_1: b_i \neq 0$)

(b) $H_1: \mu > \mu_0$ (or $H_1: b_i > 0$)

(c) $H_1: \mu < \mu_0$ (or $H_1: b_i < 0$)

The form in which we express the alternative hypothesis is important in defining the location of the rejection region or critical region of the test. For the purpose of conducting the test of a certain hypothesis concerning the population parameter we divide the whole set of the values of the population into two regions. The *acceptance region* includes the values of X which have a high probability of being observed, and the *rejection region* or *critical region* includes the values of the population which are highly unlikely, that is they have a low probability of being observed. Conventionally in econometrics we consider that highly unlikely values of a variable are those values whose total probability is less than 5 per cent or 1 per cent, that is those values which define an area of the probability distribution equal to 0·05 or 0·01. These probabilities are called *level of significance of the test.*

Step. 2. Choose the Level of Significance of the Test

In making a decision one can never be 100 per cent sure that one will make the right decision. In making any decision we are liable to commit one of the following types of error.

Type I Error: We reject the null hypothesis, when it is actually true.

Type II Error: We accept the null hypothesis, when it is actually false.

It is obvious that we want this probability (of 'being wrong') to be small. Thus we assign a low value to this probability, and we call it the *level of significance* of our test. Choosing a certain level of significance involves our specifying the probability of committing a Type I error. Usually we determine (specify) a value of the level of significance equal to 5 per cent (and more rarely 1 per cent). When we choose the 5 per cent level of significance, we tolerate to make the wrong decision (of rejecting the null hypothesis when it is actually true) five times in a hundred.

Step 3. Choose the Location of the Critical Region

We said that the critical region includes the values of the variable which have a low probability of being observed, that is a (total) probability of 5 per cent or 1 per cent. In other words the critical region includes the values which correspond to the level of significance. The critical region may be chosen either at the right end (right tail) of the distribution of the variable, or at the left tail, or half at each end of the distribution (see figure 23). In the first and second cases, we say that we conduct a *one-tail test*, in the third case we say that we conduct a *two-tail test.*

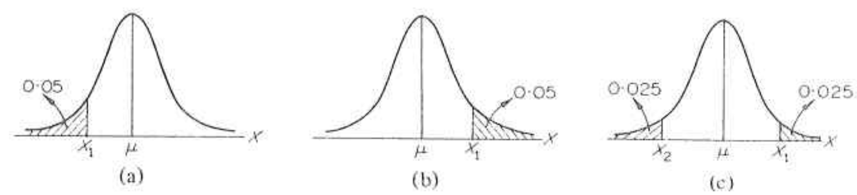


Figure 23. (a) Left-tail critical region at 5 per cent level of significance; (b) Right-tail critical regions at 5 per cent level of significance; (c) Two-tail critical region at 5 per cent level of significance.

The decision of whether to choose a one-tail or a two-tail critical region depends on the form in which the alternative hypothesis is expressed. If the alternative hypothesis is of the form

$$H_1: \mu \neq \mu_0 \qquad (\text{or } H_1: b_i \neq b_i^*)$$

we choose a two-tail critical region. If the alternative hypothesis is of the form

$$H_1: \mu > \mu_0 \qquad (\text{or } H_1: b_i > b_i^*)$$

we choose a right-tail critical region. Finally if the alternative hypothesis is of the form

$$H_1: \mu < \mu_0 \qquad (\text{or } H_1: b_i < b_i')$$

we choose the left tail of the distribution as the critical region of our test.[1] The following rule of thumb may summarise the above statements regarding the choice of the location of the critical region. Choose the location of the critical region on the basis of the direction at which the inequality sign points:

if $>$, choose the right tail as the critical region;
if $<$, choose the left tail as the critical region;
if $\neq$, choose a two-tail critical region.

In econometrics it has become customary to choose a two-tail critical region, although a one-tail test would be preferable on *a priori* economic considerations (see pp. 85 and 90).

Step 4. Choose the Appropriate Test Statistic

In econometrics the most common test statistics are the Z, t, and F statistics. The choice among them depends on the type of test which we want to conduct, on the size of the sample and on the information which we have about the population variance.

(a) If we know the variance of the parent population we may apply the Z transformation statistic, irrespective of the normality of the population and irrespective of the sample size.

(b) If the variance of the parent population is unknown, but the size of the sample is large ($n > 30$), we may still apply the Z statistic since the estimate of the population variance from a large sample is a satisfactory estimate of the true σ^2.

(c) If the variance of the parent population is unknown and our sample is small ($n < 30$) we may apply the t statistic provided that the parent population is normal. For the application of the t statistic normality is crucial.

(d) In econometric research the population variance (σ_u^2) is one of the unknowns of the estimated model. Furthermore, the samples usually available are small ($n < 30$). Thus for testing the reliability of the estimates, $\hat{b}$'s, we may apply the t statistic, or the F statistic. It has been established that

$$t^2 = F$$

(see Chapter 8).

(e) The F statistic is used for conducting various tests of significance in econometric applications. The most important of these tests have been explained in Chapter 8.

It should be clear that the choice of the test statistic aims at the transformation of the units of the variable into units of the chosen statistic, through the corresponding transformation formula. For example

$$t = \frac{X_i - \mu}{\hat{\sigma}_x} \quad \text{or} \quad Z = \frac{X_i - \mu}{\sigma_x}$$

[1] The above results concerning the choice of the location of the critical region are based on the examination of the power functions of tests. See T. Yamane, *Statistics*, pp. 196–226.

The transformation enables us to find the probability of the variable assuming any value within a certain range indirectly by using the results (established in section E.2)

$$P\{X_1 < X < X_2\} = P\{Z_1 < Z < Z_2\}$$

where

$$Z_i = \frac{X_i - \mu}{\sigma_x}$$

or

$$P\{X_1 < X < X_2\} = P\{t_1 < t < t_2\}$$

where

$$t_i = \frac{X_i - \mu}{\hat{\sigma}_x}$$

Having chosen the level of significance, the location of the critical region, and the test statistic, we may use the relevant table of the probability of this statisic and find its critical values, that is the values which define the boundaries of the critical region.

Example. Assume that we have estimated the following consumption function of a certain country for the period 1950–70

$$\hat{C} = 165{\cdot}3 + 0{\cdot}74\,Y$$
$$(43{\cdot}2) \quad (0{\cdot}20)$$

where the numbers in brackets are the standard errors of the $\hat{b}$'s. We wish to test the statistical significance of the estimated marginal propensity to consume $\hat{b}_1 = 0{\cdot}74$.

The null and alternative hypotheses are

$$H_0: b_1 = 0$$
$$H_1: b_1 \neq 0.$$

We choose the 5 per cent level of significance for our test, that is we specify that 5 times out of hundred we may make the wrong decision (of rejecting H_0 when it is actually true).

Since the alternative hypothesis is expressed in the customary form of $b_i \neq 0$, we choose a two-tail critical region. Each tail will correspond to half the chosen level of significance, that is, the area of each tail is 0·025 (or 2·5 per cent).

Given that the population variance (σ_u^2) is unknown and our sample is small ($n = 21 < 30$) we will apply the t statistic. From the t table we find the critical values of t, that is, the values which define the boundaries of the critical region. In our example the critical values of t (tabular values) will be those that cut off 2·5 per cent of the area of the t distribution. From the t table we find:

$$-t_{0{\cdot}025} = -2{\cdot}093 \quad \text{and} \quad +t_{0{\cdot}025} = +2{\cdot}093$$

(with $n - 2 = 19$ degrees of freedom). The critical region of our example is shown in figure 24.

Step 5. Compute the Sample Value of the Chosen Test Statistic

Using the sample information we compute the *sample value (or observed value)* of the chosen test statistic, which we denote with an asterisk (for example

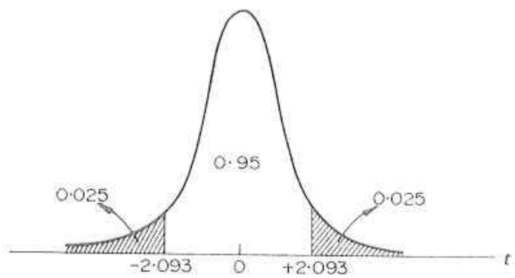


Figure 24

t^*, F^*) to distinguish it from the *critical value* or *theoretical value* or *tabular value* of the statistic which is defined in step 4.

In our example the observed (sample) value of the chosen t statistic is

$$t^* = \frac{\hat{b}_1 - 0}{s_{(\hat{b}_1)}} = \frac{0{\cdot}74}{0{\cdot}20} = 3{\cdot}7$$

Step 6. Compare the Sample Value of the Chosen Statistic with its Critical Value

The final step of our test is to compare the observed value of the chosen statistic (as estimated in step 5) with the critical (tabular) value of this statistic, as defined in step 4.

If the observed value (t^*, F^*) falls in the critical region (for example $t^* < -t_{0{\cdot}025}$ or $t^* > +t_{0{\cdot}025}$) we reject the null hypothesis and we infer that $b_i \neq 0$, that is we accept the alternative hypothesis. This of course does not imply that our estimate $\hat{b}_i$ is the correct value of the population parameter b_i. Rejection of the null hypothesis suggests merely that the true b_i has a value different from zero (see below section E.4).

If the observed value (t^*, F^*) falls outside the critical region (for example $-t_{0{\cdot}025} < t^* < t_{0{\cdot}025}$) we accept the null hypothesis, that is, we infer that the true population parameter b_i is zero.

In our example $t^* = 3{\cdot}7$. Clearly $t^* > t_{0{\cdot}025}$ and hence we reject the null hypothesis: the true population b_i has a value different from zero.

E.4. CONFIDENCE INTERVALS – INTERVAL ESTIMATION

We have concluded in the preceding paragraph that when we reject the null hypothesis we actually accept that our estimate is significantly different from zero, that is, the estimate of the parameter is obtained from a sample drawn from a population, whose true parameter is different from zero. Rejection of the null hypothesis, however, does not mean that our sample estimate is the correct estimate of the true parameter; it simply implies that the true parameter has some value different from zero. The null hypothesis test does not by itself determine how close to the true parameter our estimate is. In order to establish, as a further result, how close to the true parameter our estimate lies, we should

construct *confidence intervals* of the true parameter, that is find the range of values, round the estimate, within which (with a given probability) the true value of the population parameter will fall. This is known as *interval estimation.*

In this section we will examine how we construct confidence intervals of the true parameters, using our sample estimate and its standard deviation. Suppose that from a random sample we obtain $\overline{X} = 4$. We wish to establish how close to the population mean our estimate is. The population mean is unknown, but we assume that we know its standard deviation σ_x, or that we have a large sample from which we can get a satisfactory estimate of the standard deviation.

We know from the Central Limit Theorem that

$$Z_i = \frac{\overline{X}_i - \mu}{\sigma_{\overline{x}}} \to N(0, 1) \qquad \text{as } n \to \infty$$

where

$$\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$$

Furthermore, from the Z table we find that the range $0 \pm 1{\cdot}96$ contains 95 per cent of the values of Z, or the probability of Z taking a value between $-1{\cdot}96$ and $1{\cdot}96$ is 95 per cent. This may be written

$$P\{-1{\cdot}96 < Z < 1{\cdot}96\} = 0{\cdot}95$$

Substituting $Z = (\overline{X}_i - \mu)/\sigma_{\overline{x}}$ we obtain

$$P\left\{-1{\cdot}96 < \frac{\overline{X} - \mu}{\sigma_{\overline{x}}} < 1{\cdot}96\right\} = 0{\cdot}95$$

Subtracting $\overline{X}$ from all terms of the inequality we find

$$P\{\overline{X} - 1{\cdot}96(\sigma_{\overline{x}}) < \mu < X + 1{\cdot}96(\sigma_{\overline{x}})\} = 0{\cdot}95$$

This result reads: the probability that the true parameter μ will be in the interval $\overline{X} \pm 1{\cdot}96(\sigma_{\overline{x}})$ is 0·95.

Thus when we construct the interval $\overline{X} \pm 1{\cdot}96(\sigma_{\overline{x}})$ we can be sure with a probability of 95 per cent that this interval will include the true value of the population parameter μ. This is the reason why the interval

$$\overline{X} \pm 1{\cdot}96(\sigma_{\overline{x}})$$

is called the 95 per cent confidence interval of the true population parameter.

Example. A university lecturer wishes to estimate the average marks gained by the students of his university which form a finite population. A random sample of size $n = 36$ is selected and the sample mean is found to be $\overline{X} = 56$. The standard deviation of the marks of all the students (population) is known to be $\sigma_x = 4$. Therefore the sample standard deviation will be

$$\sigma_{\overline{x}} = \sigma_x/\sqrt{n} = 4/6 \approx 0{\cdot}67$$

The 95 per cent confidence interval will be

$$56 - (1{\cdot}96)(0{\cdot}67) < \mu < 56 + (1{\cdot}96)(0{\cdot}67)$$

Thus the true (population) average marks scored by the students of the university in question will most probably (with a probability of 95 per cent) lie between 54·69 and 57·31.

E.5. SOME NOTES ON THE MEANING OF 'DEGREES OF FREEDOM'

The concept of "degrees of freedom" is very important in performing tests of the reliability of estimates obtained from a sample. We will attempt to explain it with various illustrations.

Let us assume that we want to select various values in order to form a set of such values. The set may be a sample of values on a variable, such as

$$X_1, X_2, \ldots, X_n$$

or it may be a set of deviations of the n values of a variable from their mean:

$$(X_1 - \overline{X}), (X_2 - \overline{X}), \ldots, (X_{n-1} - \overline{X}), (X_n - \overline{X})$$

In order to form a particular set of values we select its various item-components. In some cases we are free to select all the n items of the set, while in others our freedom is limited by the knowledge of some of the values of the set. For example in choosing the n values of a random variable for a sample, we are free to choose all these values. Knowlege of any of the values does not tell us anything about the other values of the sample. Any value of the variable X is chosen withour reference to the values of the other members of the sample. The n items consist of free choices: we have n degrees of freedom in our choice.

Assume next that we want to form the set of n deviations of the n values of X from their mean $\overline{X}$. We know the n values of X and therefore we know their mean $\overline{X}$ which is computed by the formula

$$\frac{\Sigma X_i}{n} = \overline{X}.$$

We said that the sum of the deviations from the mean is always equal to zero: $\Sigma(X_i - \overline{X}) = 0$. Hence if we know $n - 1$ of the deviations, the nth is automatically determined. For example if we have 4 observations and we know that the three deviations are 5, 3, 2, then the 4th deviation from the mean must be equal to -4. In any set of n deviations from a sample mean, all but one can be chosen arbitrarily: the nth must assume such a value as to bring the sum to zero. Thus in this case we can choose only $n - 1$ items of our set: we have $n - 1$ degrees of freedom. We have lost one degree of freedom in computing $\overline{X}$ from the sample.

In general, we may say that degrees of freedom is the number of elements (of a set) that can be chosen freely, or the number of variables that can vary freely. (See D. B. Suits, *Statistics*, Rand-McNally, 1963, p. 62–3.)

SECTION F

BASIC LAWS OF PROBABILITY

In many cases we wish to calculate the probability of two (or more) events occurring at the same time. Any two events may be (a) dependent, (b) independent, (c) mutually exclusive and (d) mutually not exclusive.

Independent events. Two events are said to be independent if the occurrence of one event is not connected in any way with the occurrence of the other. The occurrence of event A does not depend on whether B has occurred or not. For example:

(a) Casting a die twice: the result of the first throwing is independent of the result of the second throwing.
(b) The sex of the children of a family: the sex of the second child is independent of the sex of the first child.
(c) The results of a horse-race and a boxing match are two independent events.
(d) Tossing a coin twice: The result of the second toss is independent of the result of the first toss.

Mutually exclusive events. Two events are mutually exclusive if the occurrence of A precludes the occurrence of B. In other words the two events cannot occur together. For example:

(a) The results of a fooball match between Wales and Scotland. There can be either a win or a loss by the Welsh team or an equal result.
(b) The results of tossing a coin once. Heads and tails are the two possible results, which are obviously mutually exclusive.
(c) The results of drawing a card from a pack. When we draw a card from a pack we may have 52 mutually exclusive events: the 'result' of the card precludes the appearance of any other card.
(d) The result of an examination. One either passes or fails. Thus the 'pass' and 'fail' are two mutually exclusive events.

Dependent Events. Two events are said to be dependent if the occurrence of the one is connected in some way with the occurrence of the other. For example:

(a) The results of drawing two cards from a pack, one at a time, without replacement. The result of the second card depends on the first, because when the first card is drawn there are only 51 cards left from which we have to draw the second card.

(b) Drawing two red balls from a bag which contains 10 red and 10 white balls. If the first ball drawn is red and is not replaced, the probability of the second ball drawn is affected, since now there are only 19 balls in the bag, and of them only 9 are red.

Not mutually exclusive events. Two events are said to be not mutually exclusive when they may occur at the same time: The occurrence of the one does not preclude the occurrence of the other. For example:

(a) Drawing a card from a pack. What is the probability of the card being either a spade or a king? These two results are not mutually exclusive since a king can be the king of spades.

(b) Electing the mayor of a town. What is the probability of the mayor being either a female or a catholic? These two events are not mutually exclusive, since a female can be a catholic.

There are four basic laws of probability.

1. *Addition rule* (Mutually exclusive events)

If two events A and B are mutually exclusive, the probability of either A or B occurring is the sum of their respective probabilities. Symbolically we may write this law as follows:

$$P(A \text{ or } B) = P(A) + P(B)$$

For example assume that we throw a die and we want to find the probability of either a two (event A) or a six (event B) appearing. These events are mutually exclusive because if the die shows up two, the occurrence of a six is precluded. Therefore the probability of either A or B is the sum of the individual probabilities:

$$P(A) = \tfrac{1}{6}$$

$$P(B) = \tfrac{1}{6}$$

$$P(A \text{ or } B) = P(A) + P(B) = \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{2}{6}$$

The same rule can be extended to the occurrence of three or more mutually exclusive events.

2. *Multiplication rule* (Independent events)

If two events are independent the probability of both A and B occurring simultaneously is the product of their individual probabilities. Symbolically we have

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

For example let A be the winning of a cricket match by an English team playing against an Australian team, and B the winning by a black horse in a horse-race. What is the probability of both these results occurring?

The two events are obviously independent. Assume that the English and Australian teams are equally good, so that the probability of England winning

is $\tfrac{1}{2}$ (or 50 per cent). Furthermore assume that in the horse-race there are 5 horses running, all of equally good shape, so that the probability of the black horse winning is $\tfrac{1}{5}$. Thus

$$P(A) = \tfrac{1}{2}$$

$$P(B) = \tfrac{1}{5}$$

and $P(A \text{ and } B) = P(A) \,.\, P(B) = \tfrac{1}{2} \times \tfrac{1}{5} = \tfrac{1}{10}$

The same rule may be extended to the occurrence of three or more independent events.

3. *Multiplication rule* (Dependent events)

If two events are dependent on each other (jointly dependent) the probability of both occurring (joint probability) is the probability of one event multiplied by the probability of the other, given that the first event has occurred. Symbolically we have

$$P(A \text{ and } B) = P(A) \,.\, P(B/A)$$

The term $P(B/A)$ designates the *conditional probability* of B, that is, the probability of B occurring given that A has already occurred.

For example suppose that we draw two cards from a pack of 52 cards. What is the probability of both cards being clubs? For the first card to be a club the probability is

$$P(A) = \tfrac{13}{52}$$

given that there are 13 clubs in a pack. Suppose the first card drawn is actually a club and that we do not return the card to the pack. Under these circumstances the probability of the second card being a club depends on the first event in two ways: (a) since the first card is a club, there remain only 12 cards in the pack; (b) since the first card is not returned to the pack there are only 51 cards left from which to draw the second card. Thus

$$P(B/A) = \frac{12}{51}$$

Therefore the joint probability of A and B occurring simultaneously is

$$P(A \text{ and } B) = P(A) \,.\, P(B/A) = \frac{13}{52} \cdot \frac{12}{51}$$

The above rule may be extended to the joint occurrence of more than two dependent events.

4. *Addition rule* (Events not mutually exclusive)

If two events are not mutally exclusive, the occurrence of either A or B means the occurrence of either A or B or both A and B. The probability of either A or B is given by the formula:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

For example suppose a composer of pop-songs decides to choose a singer from a group of singers consisting of males and females, some of them foreigners. Their names are put in a hat and the composer draws one of them at random. He wishes the singer to be either a female or a foreigner (of either sex) so as to make the performance more exciting. What is the probability that his wish will be fulfilled and that he will hit on the name of a female (event A) or a foreigner (event B)? These two events are not mutually exclusive since a female singer may also be a foreigner. Assume that the total group includes ten singers, five male and five female, two male foreigners and two female foreigners.

Event A is the selection of a female singer

Therefore $$P(A) = \frac{5}{10}$$

Event B is the selection of a foreign singer irrespective of sex

Therefore $$P(B) = \frac{4}{10}$$

However, A and B may occur simultaneously, hence

$$P(A \text{ and } B) = P(A) \,.\, P(B) = \frac{5}{10} \cdot \frac{4}{10} = \frac{20}{100} = \frac{2}{10}$$

(A and B being independent)

Therefore $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

$$= \frac{5}{10} + \frac{4}{10} - \frac{2}{10} = \frac{7}{10}$$

The rule can be extended to more than two not mutually exclusive events.

APPENDIX II

Determinants and the Solution of Systems of Equations