9 CORRELATION AND REGRESSION

9.1 CONCEPTS OF 'CORRELATION' AND 'REGRESSION'

So far we have studied some characteristics of one variable only, e.g. mean of the distribution of height, standard deviation of weight, skewness of the distribution of income. But, many situations arise in which we may have to study two variables simultaneously, say x and y. For example, the variables may be (i) the amount of rainfall and yield of a certain crop, (ii) the height and weight of a group of children, (iii) income and expenditure of several families, (iv) ages of husband and wife; etc. There are two main problems involved in such studies :

Firstly, the data may reveal some association between x and y, and we may be interested to measure numerically the strength of this association between the variables. Such a measure will determine how well a linear or other equation explains the relationship between the variables. This is the problem of *Correlation.*

Secondly, there may be one variable of particular interest, and the other variable, regarded as an auxiliary variable, may be studied for its possible aid in throwing some light on the former. In such a case, one is then interested in using a mathematical equation for making estimates or predictions regarding the principal variable. This equation is known as a *Regression Equation*, and the problem of making predictions on the basis of the equation is called the problem of *Regression.*

In short, correlation is concerned with the measurement of the 'strength of association' between variable; while regression is concerned with the 'prediction' of the most likely value of one variable when the value of the other variable is known.

In *simple correlation* and *simple regression* (also called *linear correlation* and *linear regression*), we consider the simplest kind of relationship, viz. a linear relationship, as the regression equation. Simple correlation is, therefore, concerned with the strength of linear trpe of relationship between the variables.

9.2 BIVARIATE DATA

The word '*bivariate*' is used to describe situations in which two characters are measured on each individual or item, the characters being represented by two variables. Statistical data relating to the simultaneous measurement of two variables are called *bivariate data.* (Data relating to one variable only are called *univariate data*). The observations on each individual are then paired, one for each variable—(x_1, y_1), (x_2, y_2), ..., (x_n, y_n), suppose.

$\bar{y} = \frac{\Sigma fy}{N} = \frac{24}{5} = 4.8$ $\bar{y} = \frac{42}{10} = 4.2$ $\bar{y} = \frac{102}{14} = 7.3$

The conditional mean values of y when $x = 1, 3, 5$ are respectively 4.8, 4.2, 7.3.

9.4 SCATTER DIAGRAM

Example 9.4 *What is a Scatter Diagram ? Explain how this can be used to indicate the degree and type of association between two variables.*

[D.S.W., Nov., '70; I.C.W.A., Jan. '67, July '71, Dec. '73, '74; C.U., M.Com. '63, '67; B.A. (Econ) '71, '73, '75; B.U., B.A. '77]

Solution When statistical data relating to the simultaneous measurement on two variables are available, each pair of observations can be geometrically represented by a point on the graph paper—the values of one variable being shown along the X-axis and those of the other variable along Y-axis. If there are n pairs of observations, finally the graph paper will contain n points. This diagrammatic representation of bivariate data is known as *Scatter Diagram* (Figs 9.1 and 9.2).

A scatter diagram indicates the nature of association between the two variables, i.e., the type of correlation between them. If the pattern of points (or dots) on the scatter diagram shows a linear path diagonally across the graph paper from the the bottom left-hand corner to the top right, correlation will be positive (Fig. 9.1a). In other words, association between the variables is direct, indicating thereby that high values of one variable are in general, associated with high values of the other variable, and low values are associated with low values.

On the other hand, if the pattern of dots be such as to indicate a straight line path from the upper left-hand corner to the bottom right, correlation is negative, i.e. the association is indirect, high values of one variable being associated with low values of the other (Fig. 9.1b).

When the dots do not indicate any straight line tendency, but a swarm (Fig. 9.1c), or concentration around a curved line, correlation is small (Fig. 9.1d). In fact, if no straight line tendency is noticed, correlation will be zero.

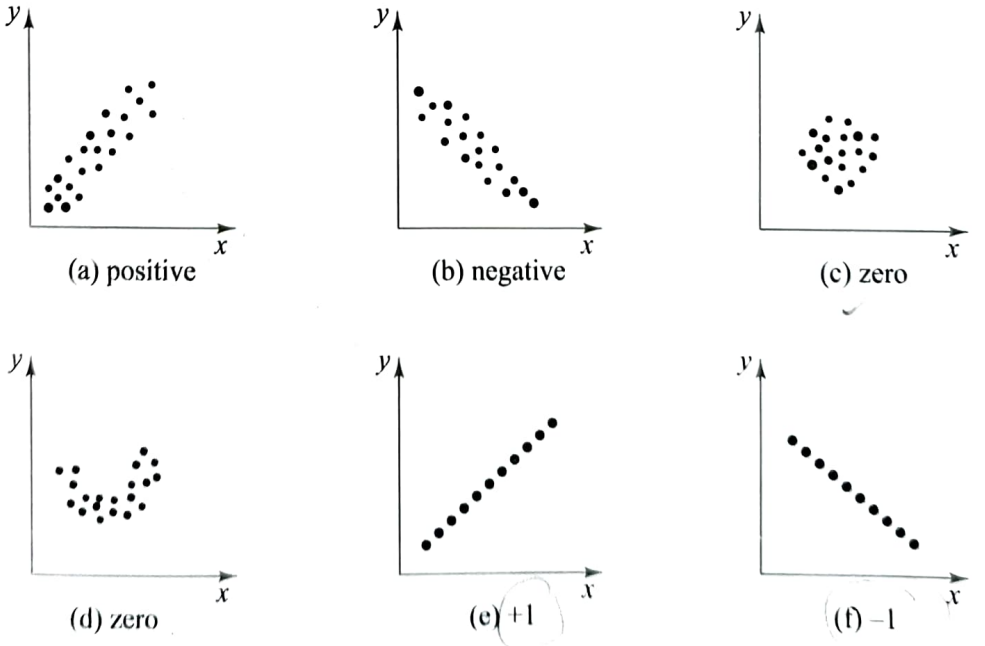


Fig. 9.1 Scatter Diagrams showing Different Types and Degrees of Correlation

When the dots lie exactly on a straight line, correlation is perfect—the correlation coefficient being + 1 or − 1, according as the slope of the straight line is positive or negative (Figs 9.1e, 9.1f).

The scatter diagram also gives an indication of the degree of linear correlation between the variables, i.e. whether correlation is high or low. If the plotted points on the scatter diagram lie approximately on, or near about, a straight line (Figs 9.1e, 9.1f, 9.2), correlation coefficient will be nearly one, numerically. The more scattered the points are around a straight line, the less is the correlation coefficient (Figs 9.1a, 9.1b). A scatter diagram like Fig. 9.1c will indicate almost zero correlation.

Example 9.5 *Construct the scatter diagram of the data given below. Draw a free hand straight line through the group of points and from your diagram discuss the probable amount of correlation:*

	Average Values in Lakhs							
	1946	1947	1948	1949	1950	1951	1952	1953
Raw Cotton Imports	47	64	100	97	126	203	171	115
Cotton Manufacturers' Exports	70	85	100	103	111	139	133	115

[C.U., M.Com. '69, '72]

Solution The variables here are Raw Cotton Imports (x, say) and Cotton Manufactures' Exports (y). The 8 pairs of observations (47, 70), (64, 85), ..., (115, 115) are plotted as points on a graph paper, giving the scatter diagram (Fig. 9.2). A free hand straight line has been drawn. It is found that out of 8 points, 4 lie almost on the straight line and the remaining 4 are also very close to it. This indicates that the correlatioin coefficient between the variables x and y is numerically very high, say about 0.95. Again since the slope of the straight line is positive, the correlation coefficient is also positive. The amount of correlation is thus estimated to be about + 0.95 (The calculated value of the correlation coefficient for the data is + 0.98).

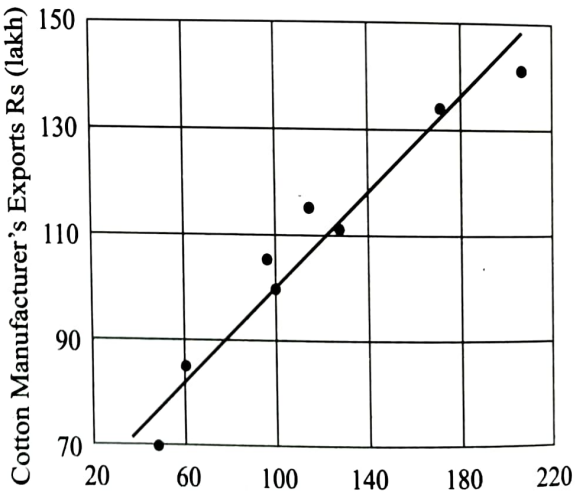


Fig. 9.2 Scatter Diagram

9.5 CORRELATION

The word "correlation" is used to denote the *degree of association between variables.* If two variables x and y are so related that variations in the magnitude of one variable tend to be accompanied by variations in the magnitude of the other variable, they are said to be *correlated.* If y tends to increase as x increases, the variables are said to be *positively correlated.* If y tends to decreases as x increases, the variables are *negatively correlated.* If the values of y are not affected by changes in the values of x, the variables are said to be *uncorrelated* (see Example 9.6).

Correlation may also be *linear* or *non-linear.* If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then correlation is said to be 'linear'; because the scatter diagram would show a linear path. Here, we shall be concerned with *linear correlation* or *simple correlation* only. This is measured by *'Correlation Coefficient'* (Section 9.7).

9.6 COVARIANCE

Given a set of n pairs of observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) relating to two variables x and y, the *Covariance of x and y*, usually represented by *cov* (x, y), is defined as

$$\text{cov}\,(x, y) = \frac{1}{n}\Sigma(x - \bar{x})(y - \bar{y}) \tag{9.6.1}$$

Expanding the expanssion on the right, it can be shown that

$$\text{cov}\,(x, y) = \frac{\Sigma xy}{n} - \left(\frac{\Sigma x}{n}\right)\left(\frac{\Sigma y}{n}\right) \tag{9.6.2}$$

This form is generally used for calculations. Covariance has properties similar to those of variance, i.e. the square of S.D.

(i) If $X = x - c$ and $Y = y - c'$, where c, c' are constants, then

$$\text{cov}(x, y) = \text{cov}\,(X, Y) \tag{9.6.3}$$

(ii) If $u = (x - c)/d$ and $v = (y - c')/d'$, where c, c', d d' are constants, then

$$\text{cov}(x, y) = dd' \,.\, \text{cov}\,(u, v) \tag{9.6.4}$$

However, while variance must be always positive, covariance may be positive, negative or zero.

[**Note:** By definition,

$$\text{Variance of } x = \sigma_x^2 = \frac{1}{n}\Sigma(x - \bar{x})^2 = \frac{1}{n}\Sigma(x - \bar{x})(x - \bar{x})$$

$$\text{Variance of } y = \sigma_y^2 = \frac{1}{n}\Sigma(y - \bar{y})^2 = \frac{1}{n}\Sigma(y - \bar{y})(y - \bar{y})$$

$$\text{Covariance of } x \text{ and } y = \frac{1}{n}\Sigma(x - \bar{x})(y - \bar{y})$$

For the variances, both factors in the expression on the right are same, either both $(x - \bar{x})$, or both $(y - \bar{y})$; both for the covariance one factor is $(x - \bar{x})$ and another $(y - \bar{y})$. Again, the working formulae are

$$\text{Variance of } x = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{\Sigma x.x}{n} - \left(\frac{\Sigma x}{n}\right)\left(\frac{\Sigma x}{n}\right)$$

$$\text{Variance of } y = \frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2 = \frac{\Sigma y.y}{n} - \left(\frac{\Sigma y}{n}\right)\left(\frac{\Sigma y}{n}\right)$$

$$\text{Covariance of } x \text{ and } y = \frac{\Sigma x.y}{n} - \left(\frac{\Sigma x}{n}\right)\left(\frac{\Sigma y}{n}\right)$$

For the variances, the first term on the right contains $x.x$ or $y.y$ and for the covariance it is $x.y$, i.e. one is x and the other y; in the second term similarly instead of both factors same, $\Sigma x/n$ or $\Sigma y/n$, we have one from x and another from y, i.e. $(\Sigma x/n)\,(\Sigma y/n)$.

It is thus seen that covarience is a variance like quantity, but obtained by the combination of two variables. Covariance may, therefore, be looked upon as a 'conjoint variance'.

It is also interesting to note that covariance of a variable and the same variable is the variance itself. Replacing y by x in (9.6.1) or (9.6.2), we see that cov. $(x, x) = \sigma_x^2$; similarly, cov $(y.y) = \sigma_y^2$.

9.7 CORRELATION COEFFICIENT (r)

Let (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) be a given set of n pairs of observations on two variables x and y. The *Correlation Coefficient*, or *Coefficient of Correlation*, between x and y (denoted by the symbol r) is then defined as

$$r = \frac{\text{cov}(x, y)}{\sigma_x\, \sigma_y} \tag{9.7.1}$$

where σ_x and σ_y are the standard derivations of x and y respectively, and cov (x, y) denotes the coveriance of x and y (Section 9.6). This expression is known as *Pearson's product-moment formula,* and is used as a measure of linear correlation between x and y.

The formula for r may be written in various other forms. Putting the explicit expressions for cov(x, y), σ_x and σ_y in (9.7.1), and multiplying both the numerator and the denominator by n, we have

$$r = \frac{\Sigma\,(x - \bar{x})\,(y - \bar{y})}{\sqrt{[\Sigma(x - \bar{x})^2\,.\,\Sigma(y - \bar{y})^2]}} \tag{9.7.2}$$

Now, expanding the expressions

$$r = \frac{\Sigma\, xy - n\bar{x}\bar{y}}{\sqrt{[(\Sigma x^2 - n\bar{x}^2)\,.\,(\Sigma y^2 - n\bar{y}^2)]}} \tag{9.7.3}$$

Multiplying the numerator and the denominator by n again, and since $n\bar{x} = \Sigma x$ and $n\bar{y} = \Sigma y$, we may write

$$r = \frac{n\Sigma xy - (\Sigma x)\,(\Sigma y)}{\sqrt{[\{n\,\Sigma x^2 - (\Sigma x)^2\} \cdot \{n\,\Sigma y^2 - (\Sigma y)^2\}]}} \tag{9.7.4}$$

[**Note:** In all the forms shown above, the denominators contains two factors under the square-root. They may be obtained on replacing y by x, and x by y, in the numerator.]

9.8 PROPERTIES OF CORRELATION COEFFICIENT

(i) The correlation coefficient r is independent of the choice of both origin and scale of observations. This means that if

$$u = \frac{x - c}{d} \quad \text{and} \quad v = \frac{y - c'}{d'}$$

where c, c', d, d' are arbitrary constants (d and d' positive), then

$$r_{xy} = r_{uv} \tag{9.8.1}$$

i.e. *Correlation coefficient between x and y*
= *Correlation coefficient between u and v*

In genreal, if $X = a + bx$, $Y = a' + b'y$, then

$$r_{XY} = \pm r_{xy} \tag{9.8.2}$$

according as b and b' have the same sign, or opposite signs.

(ii) The correlation coefficient r is a pure number and is independent of the units of measurement. This means that if, for example, x represents height in inches and y weight in lbs., then the correlation coefficient between x and y will neigther be in inches nor in lbs. or any other unit, but only a number.

(iii) The correlation coefficient r lies between – 1 and + 1; i.e. r cannot exceed 1 numerically.

$$-1 \leq r \leq +1 \tag{9.8.3}$$

9.9 CALCULATION OF r

Correlation coefficient (r) is unaffected by the choice of origin and scale of one or both the variables (Property (i), Section 9.8). Therefore, it can be calculated from a given set of a n pairs of observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) as follows:

(I) If $X = x - c$ and $Y = y - c'$, (here c, c' are constants), then

$$r_{xy} = r_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \, \sigma_Y} \tag{9.9.1}$$

where $\sigma_X^2 = \dfrac{\Sigma X^2}{n} - \left(\dfrac{\Sigma X}{n}\right)^2$, $\sigma_Y^2 = \dfrac{\Sigma Y^2}{n} - \left(\dfrac{\Sigma Y}{n}\right)^2$

$$\text{cov}(X, Y) = \frac{\Sigma XY}{n} - \left(\frac{\Sigma X}{n}\right)\left(\frac{\Sigma Y}{n}\right) \tag{9.9.2}$$

Thus, we can always reduce the given values of x and y on subtracting convenient numbers c and c', and obtain deviations $X = x - c$, $Y = y - c'$. From these reduced values X and Y, the two standard deviations and the covariance, viz. σ_X, σ_Y and cov (X, Y), are now calculated, and finally the correlation coefficient r_{XY} between them.

This will be exactly equal to the correlation coefficient r_{xy} between the original values of x and y (see Example 9.9).

(II) If $u = \dfrac{x - c}{d}$ and $v = \dfrac{y - c'}{d'}$, (here c, c', d, d' are constants and d, d' are positive), then

$$r_{xy} = r_{uv} = \frac{\text{cov}(u, v)}{\sigma_u \, \sigma_v} \tag{9.9.3}$$

where $\sigma_u^2 = \dfrac{\Sigma u^2}{n} - \left(\dfrac{\Sigma u}{n}\right)^2$, $\sigma_v^2 = \dfrac{\Sigma v^2}{n} - \left(\dfrac{\Sigma v}{n}\right)^2$

$$\text{cov}(u, v) = \frac{\Sigma uv}{n} - \left(\frac{\Sigma u}{n}\right)\left(\frac{\Sigma v}{n}\right) \tag{9.9.4}$$

In some cases, the given values of x and y may be such that it is further possible to reduce the deviations $x - c$ and $y - c'$ on division by constant factors d and d', i.e. $u = (x - c)/d$, $v = (y - c')/d'$. From these values of u and v, the two standard deviations σ_u, σ_v and the covariance cov(u, v) are calculated, and finally the correlation coefficient r_{uv} between u and v is obtained. This will be exactly equal to the correlation coefficient r_{xy} between the original values of x and y (see Example 9.11).

Example 9.6 *State in each case whether you would expect to find a positive correlation, a negative correlation or no correlation:*

(i) The ages of husband and wives,
(ii) Shoe size and intelligence,
(iii) Insurance companies' profit and the number of claims they have to pay,
(iv) Years of education and income,
(v) Amount of rainfall and yield of crop.

Solution (i) Positive, (ii) No correlation, (iii) Negative, (iv) Positive, (v) Positive.

Example 9.7 *Calculate the correlation coefficient and its probable error from the following results:*

$$\sum_{i=1}^{100} x_i = 280, \quad \sum_{i=1}^{100} y_i = 60, \quad \sum_{i=1}^{100} x_i^2 = 2384, \quad \sum_{i=1}^{100} y_i^2 = 117, \quad \sum_{i=1}^{100} x_i y_i = 438$$

[I.C.W.A., July '71]

Solution

Correlation coefficient $r = \dfrac{\text{cov}(x, y)}{\sigma_x \, \sigma_y}$

Here $\sigma_x^2 = \dfrac{\Sigma x^2}{n} - \left(\dfrac{\Sigma x}{n}\right)^2 = \dfrac{2384}{100} - \left(\dfrac{280}{100}\right)^2 = 16$

$\sigma_y^2 = \dfrac{\Sigma y^2}{n} - \left(\dfrac{\Sigma y}{n}\right)^2 = \dfrac{117}{100} - \left(\dfrac{60}{100}\right)^2 = 0.81$

$\text{cov}(x, y) = \dfrac{\Sigma xy}{n} - \left(\dfrac{\Sigma x}{n}\right)\left(\dfrac{\Sigma y}{n}\right) = \dfrac{438}{100} - \dfrac{280}{100} \times \dfrac{60}{100} = 2.7$

Substituting the values,

$$r = \frac{2.7}{\sqrt{16}\,\sqrt{0.81}} = \frac{2.7}{4 \times 0.9} = 0.75$$

The Probable Error (P. E.) of correlation coefficient is

$$\text{P.E.} = 0.6745\frac{1-r^2}{\sqrt{n}}$$

$$= 0.6745 \times \frac{1-(0.75)^2}{\sqrt{100}} = 0.0295.$$

Ans. $r = 0.75$, P.E. $= 0.0295$

Example 9.8 *While calculating the coefficient of correlation between two variables x and y, the following results were obtained : n = 25, Σx = 125, Σy = 100, Σx² = 650, Σy² = 460, Σxy = 508. It was however later discovered at the time of checking that two pairs of observations (x, y) were copied (6, 14) and (8, 6), while the correct values were (8, 12) and (6, 8) respectively. Determine the correct value of the coefficient of correlation.* [I.C.W.A., June '77, '78]

Solution For the two incorrect pairs of observations

x: 6, 8; $\Sigma x = 6 + 8 = 14$; $\Sigma x^2 = 6^2 + 8^2 = 100$;

y: 14, 6; $\Sigma y = 14 + 6 = 20$; $\Sigma y^2 = 14^2 + 6^2 = 232$;

$\Sigma xy = 6 \times 14 + 8 \times 6 = 132$.

For the two correct pairs

x: 8, 6; $\Sigma x = 8 + 6 = 14$; $\Sigma x^2 = 8^2 + 6^2 = 100$;

y: 12, 8; $\Sigma y = 12 + 8 = 20$; $\Sigma y^2 = 12^2 + 8^2 = 208$;

$\Sigma xy = 8 \times 12 + 6 \times 8 = 144$.

When the incorrect pairs of observations are replaced by the correct pairs, the revised results for the 25 pairs of observations are:

$$\Sigma x = 125 - 14 + 14 = 125; \qquad \Sigma y = 100 - 20 + 20 = 100;$$

$$\Sigma x^2 = 650 - 100 + 100 = 650; \qquad \Sigma y^2 = 460 - 232 + 280 = 436;$$

$$\Sigma xy = 508 - 132 + 144 = 520$$

Using these results,

$$\sigma_x^2 = 650/25 - (125/25)^2 = 26 - 25 = 1; \qquad \sigma_x = 1$$

$$\sigma_y^2 = 436/25 - (100/25)^2 = 17.44 - 16 = 1.44; \qquad \sigma_y = 1.2$$

$$\text{cov}(x, y) = 520/25 - (125/25)(100/25) = 20.8 - 20 = 0.8$$

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{0.8}{1 \times 1.2} = \frac{2}{3}$$

Ans. 2/3.

Example 9.9 *Find the coefficient of correlation from the following data:*

X	65	63	67	64	68	62	70	66
Y	68	66	68	65	69	66	68	65

[C.U., B.A. (Econ) '69]

Solution Since the correlation coefficient is unaffected by change of origin (and also scale), let us change the origins of X and Y to 65 and 67 respectively, i.e. write $x = X - 65$, $y = Y - 67$.

Table 9.9 Calculations for Correlation Coefficient

X	Y	$x = X - 65$	$y = Y - 67$	x^2	y^2	xy
(1)	(2)	(3)	(4)	(5)	(6)	(7)
65	68	0	1	0	1	0
63	66	−2	−1	4	1	2
67	68	2	1	4	1	2
64	65	−1	−2	1	4	2
68	69	3	2	9	4	6
62	66	−3	−1	9	1	3
70	68	5	1	25	1	5
66	65	1	−2	1	4	−2
525	535	5	−1	53	17	18

$$\sigma_x^2 = \frac{53}{8} - \left(\frac{5}{8}\right)^2 = \frac{399}{64}$$

$$\sigma_y^2 = \frac{17}{8} - \left(\frac{-1}{8}\right)^2 = \frac{135}{64}$$

$$\text{cov}(x, y) = \frac{18}{8} - \left(\frac{5}{8}\right)\left(\frac{-1}{8}\right) = \frac{149}{64}$$

$$r_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{149}{64}}{\sqrt{\frac{399}{64}}\sqrt{\frac{135}{64}}} = \frac{149}{\sqrt{(399 \times 135)}} = 0.64$$

Therefore, the required correlation coefficienty is 0.64.

[**Note:** (i) The origins should be changed to values near the means. Here, the means of X and Y are 525/8 = 65.625 and 535/8 = 66.875 respectively.

(*ii*) The following check may be introduced for the accuracy of figures in cols. (3) and (4), on which the calculations in the subsequent columns depend. (3) must be $525 - 8 \times 65 = 525 - 520 = 5$. Similarly the total of col. (2) may be used to check the total of col. (4), viz. $535 - 8 \times 67 = 535 - 536 = -1$.]

Example 9.10 *Marks of 10 students in Mathematics and Statistics are given below:*

Mathematics (X)	32	38	48	43	40	22	41	69	35	64
Statistics (Y)	30	31	38	43	33	11	27	76	40	59

Calculate (i) product-moment correlation coefficient, and (ii) its standard error.

[I.C.W.A., June'74]

Solution Since the correlation coefficient is unaffected by changes of origin, we write $x = X - 43$, $y = Y - 38$, and construct a table like Table 9.9. It will be seen that

$$n = 10, \Sigma x = 2, \Sigma y = 8, \Sigma x^2 = 1806, \Sigma y^2 = 2902, \Sigma xy = 2140.$$

Using these results, $\sigma_x^2 = 180.56$, $\sigma_y^2 = 289.56$, $\text{cov}(x, y) = 213.84$.

$$r = \frac{213.84}{\sqrt{(180.56 \times 289.56)}} = 0.94$$

Standard Error (*S.E.*) of correlation coefficient is

$$S.E. = \frac{1 - r^2}{\sqrt{n}} = \frac{1 - (0.94)^2}{\sqrt{10}} = 0.04$$

Ans. (i) 0.94, (ii) .04

Example 9.11 *Calculate the coefficient of correlation from the following data :*

x	2.52	2.49	2.49	2.45	2.43	2.42	2.41	2.40
y	730	710	770	890	970	1020	970	1040

[C.U., B.A. (Econ) '73 : M.Com. '76]

Solution Since the coefficient of correlation is unaffected by changes of origin and scale, let us write

$$u = \frac{x - 2.45}{.01}, \quad v = \frac{y - 900}{10}$$

Table 9.10 Calculations for Correlation Coefficient

x	y	$u = \frac{x - 2.45}{.01}$	$v = \frac{y - 900}{10}$	u^2	v^2	uv
2.52	730	7	– 17	49	289	– 119
2.49	710	4	– 19	16	361	– 76
2.49	770	4	– 13	16	169	– 52
2.45	890	0	– 1	0	1	0
2.43	970	– 2	7	4	49	– 14
2.42	1020	– 3	12	9	144	– 36
2.41	970	– 4	7	16	49	– 28
2.40	1040	– 5	14	25	196	– 70
Total 19.61	7100	1	– 10	135	1258	– 395

[**Note:** (i) For the calculation of correlation coefficient, a change of origin is always possible, and must be made; a change of scale, wherever possible, should also be made to simplify the calculations. In the present case, the given values of y are divisible by 10, indicating that changes of both origin and scale and possible. Also, since the values of x are given in two decimals, division by .01 (i.e. multiplication by 100) will change them to whole numbers, and consequently worries about fixing the decimal point can be avoided.

(ii) Another important point to note here is how to effect changes of origin and scale both in one step. When data are given in two decimals the origin should be changed to a conveninet two-decimal number near the mean, and the scale shall be changed to 01. Here, $u = (x - 2.45)/.01 = 100_x - 245$. This shows that for obtaining the values of u, each x is to be multiplied by 100 (which can be achieved simply by ignoring the decimal point) and then 245 subtracted. Thus, 2.52 changes to 252, and 252 – 245 = 7; 2.49 changes to 2.49 and 249 – 245 = 4; etc. (For data given in one-decimal figures, the origin should be changed to a one-decimai number and scale to 0.1).

Again, $v = (y - 900)/10 = y/10 - 90$. This shows that for obtaining the values of v, each y is to be divided by 10 (i.e. ignore the last zero) and then subtract 90. For example, 730 changes to 73 and $73 - 90 = -17$; 710 changes to 71 and $71 - 90 = -19$; etc.]

Now using formula (9.9.3),

$$\sigma_u^2 = \Sigma u^2/n - (\Sigma u/n)^2 = 135/8 - (1/8)^2 = 1079/64$$

$$\sigma_v^2 = \Sigma v^2/n - (\Sigma v/n)^2 = 1258/8 - (-10/8)^2 = 9964/64$$

$$\text{cov}(u, v) = \Sigma uv/n - (\Sigma u/n)(\Sigma v/n) = -395/8 - (1/8)(-10/8) = -3150/64$$

$$r = \frac{\frac{-3150}{64}}{\sqrt{\frac{1079}{64}}\sqrt{\frac{9964}{64}}} = \frac{-3150}{\sqrt{1079 \times 9964}} = -0.96$$

Example 9.12 *The following table gives the index numbers of industrial production in a country and the number of registered unemployed persons in the same country during the eight consecutive years. Calculate the coefficient of correlation and comment on the result.*

Year	1954	1955	1956	1957	1958	1959	1960	1961
Index of industrial production	100	102	103	105	106	104	103	98
No. of registered unemployed (in thousands)	10.5	11.4	13.0	11.5	12.0	12.5	15.6	20.8

[I.C.W.A., July '64; C.U., B.A. (Econ)' 65]

Solution Let x denote the index of industrial production the y the number of registered unemployed (in thousands). Since the correlation coefficient remains unaffected by changes of origin and scale, we write $u = x - 100$ and $v = (y - 13.0)/0.1$

$$\sigma_u^2 = \Sigma u^2/n - (\Sigma u/n)^2 = 103/8 - (21/8)^2 = 6.0$$

$$\sigma_v^2 = \Sigma v^2/n - (\Sigma v/n)^2 = 7991/8 - (33/8)^2 = 982$$

$$\text{cov}(u, v) = \Sigma uv/n - (\Sigma u/n)(\Sigma v/n) = -265/8 - (21/8)(33/8) = -44$$

Table 9.11 Calculations for Correlation Coefficient

x (1)	y (2)	$u = x - 100$ (3)	$v = (y - 13.0)/0.1$ (4)	u^2 (5)	v^2 (6)	uv (7)
100	10.5	0	−25	0	625	0
102	11.4	2	−16	4	256	−32
103	13.0	3	0	9	0	0
105	11.5	5	−15	25	225	−75
106	12.0	6	−10	36	100	−60
104	12.5	4	−5	16	25	−20
103	15.6	3	26	9	676	78
98	20.8	−2	78	4	6084	−156
Total 821	107.3	21	33	103	7991	−265

$$\therefore \quad r_{xy} = r_{uv} = \frac{\text{cov}(u, v)}{\sigma_u \, \sigma_v} = \frac{-44}{\sqrt{(6.0 \times 982)}} = -0.57$$

Comments—The scatter diagram (not show here) does not indicate any straight line path and hence $r = -0.57$ does not really serve as a useful measure of association between the variables. In fact, if the data for 1961 are excluded, we find $r = +0.21$, a result which differs much from the former. The observed correlation coefficient $r = -0.57$ may be called *non-sense correlation* (Page 286).

Example 9.13 *Prove that the correlation coefficient does not depend on the origin or scale of the observations.*

[C.U., B.A. (Econ) '66, M. Com. '76; B.U., B.A. (Econ)'68; I.C.W.A., July '70, Jan. '72, June '75]

Solution Let x and y be the variables, and (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) denote n pairs of observations. Let us change the origins of x and y to a and c, and the units of measurement (i.e. scales) to b and d respectively, and write

$$u_i = \frac{x_i - a}{b}, \quad v_i = \frac{y_i - c}{d} \tag{i}$$

where a, b, c, d are arbitrary constants (b and d positive). We have to prove that the correlation coefficient between x and y is the same as that between u and v; i.e. $r_{xy} = r_{uv}$.

From (i), $\quad x_i = a + bu_i, \quad y_i = c + dv_i$

Therefore, $\quad \bar{x} = \Sigma x_i/n = \Sigma(a + bu_i)/n = (na + b\,\Sigma u_i)/n = a + b\,\bar{u}$

Similarly, $\quad \bar{y} = c + d\,\bar{v}$

Hence, $\quad \sigma_x^2 = \Sigma(x_i - \bar{x})^2/n = \Sigma(a + bu_i - a - b\,\bar{u})^2/n$

$$= b^2\,\Sigma(u_i - \bar{u})^2/n = b^2 \,.\, \sigma_u^2$$

$\therefore \quad \sigma_x = b \,.\, \sigma_u$ (since b is positive)

Similarly, $\quad \sigma_y = d.\ \sigma_v$ (since d is positive) (ii)

Again, $\quad \text{cov}(x, y) = \Sigma(x_i - \bar{x})\,(y_i - \bar{y})/n$

$$= \Sigma\{(a + bu_i - a - b\,\bar{u})\,(c + dv_i - c - d\,\bar{v})\}/n$$

$$= bd\,\Sigma(u_i - \bar{u})\,(v_i - \bar{v})/n$$

$$= bd.\ \text{cov}(u, v) \tag{iii}$$

Substituting the values from (ii) and (iii),

$$r_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \, \sigma_y} = \frac{bd.\ \text{cov}(u, v)}{(b.\ \sigma_u)\ (d.\ \sigma_v)} = \frac{\text{cov}(u, v)}{\sigma_u \, \sigma_v} = r_{uv}$$

i.e. Correlation coefficient between x and y
= Correlation coefficient between u and v.

Example 9.14 *If $X_i = (x_i - a)/b$, $Y_i = (y_i - c)/d$, $(i = 1, 2, ..., n)$ where a, b, c, d are arbitrary constants, prove that $r_{xy} = r_{XY}$, if b and d are of the same sign, and $r_{xy} = -r_{XY}$ if they are of opposite signs.*

[C.U., B.A. (Econ) '68, ' 70; M. Com.' 74; B.U., B.A., (Econ)'72]

Solution As in Example 9.13, $\sigma_x^2 = b^2.\sigma_X^2$, $\sigma_y^2 = d^2 \,.\, \sigma_Y^2$, and $\text{cov}(x, y) = bd.\ \text{cov}(X, Y)$. These results are true whatever be the signs of b and d. But since the standard deviation is always positive, we must have $\sigma_x = |b| \,.\, \sigma_X$, $\sigma_y = |d| \,.\, \sigma_y$ (see 6.6.2, page 185)

$$r_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \, \sigma_y} = \frac{bd.\ \text{cov}(X, Y)}{|b| \,.\, \sigma_X \; |d| \,.\, \sigma_Y}$$

$$= \frac{bd}{|b| \,.\, |d|} \cdot \frac{\text{cov}(X, Y)}{\sigma_X \, \sigma_Y} = \frac{bd}{|b| \,.\, |d|} \; r_{XY}$$

If b and d are of the same sign (i.e. either both positive or both negative), the product bd will be positive, and exactly equal to $|b| \,.\, |d|$, which is always positive, Therefore,

$$\frac{bd}{|b| \,.\, |d|} = +1, \text{ so that } \quad r_{xy} = r_{XY}.$$

If b and d are of opposite signs (i.e. one of them positive and the other negative), the product bd will be negative, but numerically equal to $|b| \,.\, |d|$. Therefore,

$$\frac{bd}{|b| \,.\, |d|} = -1, \text{ so that } r_{xy} = -r_{XY}. \quad (\textit{Proved})$$

Example 9.15 *If the correlation coefficient between x and y is 0.5, what would be the correlation coefficient between $5x$ and $-3y$?*

Solution Let us write $X = 5x$ and $Y = -3y$. Then by formula (9.8.2), we have (here $a = 0$, $b = 5$, $a' = 0$, $b' = -3$),

$$r_{XY} = -r_{xy} = -0.5$$

since the coefficients of x and y (viz. 5 and –3) have opposite signs.

Example 9.16 *Prove that the correlation coefficient r lies between -1 and $+1$.*
[C.U., B.A.(Econ) '67, '74; M.Com. '74; B.U., B.A. (Econ) '65, '68, '72, '73]

Solution Let x and y be the variables, and (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) denote n pairs of observations, with means $\bar{x}$, $\bar{y}$ and standard deviations σ_x, σ_y respectively. If we write

$$u_i = \frac{x_i - \bar{x}}{\sigma_x} \quad \text{and} \quad v_i = \frac{y_i - \bar{y}}{\sigma_y}$$

then $$\Sigma u_i^2 = \sum \left(\frac{x_i - \bar{x}}{\sigma_x} \right)^2 = \frac{\Sigma (x_i - \bar{x})^2}{\sigma_x^2} = \frac{n \, \sigma_x^2}{\sigma_x^2} = n$$

Similarly, $$\Sigma v_i^2 = n$$

Again, $$\Sigma u_i v_i = \sum \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right) = \frac{\Sigma (x_i - \bar{x}) \, (y_i - \bar{y})}{\sigma_x \, \sigma_y}$$

$$= \frac{n \,.\, \text{cov}(x, y)}{\sigma_x \sigma_y} = nr$$

where r denotes the correlation coefficient between x and y.

Now, $(u_i + v_i)^2$ can never be negative, because it is a perfect square. Hence, the sum of all such squares for $i = 1, 2, \ldots, n$, cannot be negative; i.e.

$$\Sigma (u_i + v_i)^2 \geq 0$$

or, $$\Sigma u_i^2 + \Sigma v_i^2 + 2\Sigma u_i v_i \geq 0$$

or, $$n + n + 2nr \geq 0$$

or, $$2n + 2nr \geq 0$$

or, $$2n(1 + r) \geq 0$$

or, $$1 + r \geq 0$$

or, $$r \geq -1; \quad \text{i.e.} -1 \leq r$$

Similarly, since $(u_i - v_i)^2$ cannot be negative,

$$\Sigma(u_i - v_i)^2 \geq 0$$

or, $$\Sigma u_i^2 + \Sigma v_i^2 - 2\Sigma u_i v_i \geq 0$$

or, $$n + n - 2nr > 0$$

or, $$2n(1 - r) \geq 0$$

or, $$1 - r \geq 0$$

or, $$1 \geq r; \quad \text{i.e. } r \leq 1$$

Combining the results $-1 \leq r$ and $r \leq 1$, we have $-1 \leq r \leq +1$. This proves that the correlation coefficient lies between -1 and $+1$.

9.10 INTERPRETATION AND USE OF *r*

Example 9.17 *Discuss the various uses and limitations of coefficient of correlation, indicating the difficulties involved in the interpretation of the coefficient of correlation.*

[C.U., B.A.(Econ) '66; M.Com '68]

Solution *Uses—*

(i) The coefficient of correlation (r) is a measure of the degree of association between two variables. For comparing two series of observations, it is sometimes necessary to determine whether they are associated or not, and to establish relations of cause and effect. The coefficient of correlation furnishes a method of determining numerically the existence of such causal connection between them, e.g. whether weights of school boys are connected with their parents salaries. When the pattern of dots in the scatter diagram is linear, the correlation coefficient (r) can be considered as a useful measure of such relationship. A positive value of r indicates that high values of one variable are in general associated with high values of the other, and low values with low values. When r is negative, high values of one variable are in general associated with low values of the other.

(*ii*) Again, the proportion of variation explained by regression is equal to the square of correlation coefficient (Section 9.14), i.e.

$$r^2 = \textit{Proportion of variation explained by regression}$$

The value of r^2, therefore, enables us to state the relative amount of variation in the dependent variable which can be explained by the regression equation.

(*iii*) The coefficient of correlation also helps in estimating the value of the dependent variable, when the value of the independent variable is known. Thus, for a given x, the estimated value of y is obtained from the regression equation of y on x (page 291)

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

When $r = 0$, neither y nor x can be estimated by a linear function of the other variable. Again, when $r = \pm 1$, the two regression lines coincide, and the value of one variable can be accurately predicted by the linear regression equation.

(iv) In educational and psychological measurements, coefficients of correlation is used in problems of *reliability and validity* of tests.

Limitations—

(i) In linear correlation, it is assumed that there is a straight line relationship between the variables. A small value of r therefore indicates only a poor *linear type* of relationship between the variables. This however, does not rule out the possibility that the association is very cloose, but the relationship is non-linear (Example 9.21). Before using r as a measure of the degree of association between the variables, it is, therefore, advisable to draw a scatter diagram and see whether the pattern of points is linear or not.

(ii) Again, a high value of r also does not imply that there is a direct cause-and-effect relationship between the variables. The high value of r may be generated solely due to the influence of a third variable affecting both. In this cause, the effect of the third variable should be eliminated from the first two and then the 'partial' correlation coefficient between them found out (Section 9.17).

(iii) Sometimes, it may happen that two series of observations show a high correlation coefficient even though there is no logical basis for any relationship between them. For example, a renowned statistician observed a high correlation between the stork population is Oslo, Norway, over a period of several years and the number of babies born there each year. Yet it is hard to develop a theory as to why this should be. Such correlation is said to be *Spurious correlation* or *Non-sense correlation*. One should apply common-sense is deciding whether the association indicated by the value of r is real or spurious (Example 9.12).

(iv) If the data are not reasonably homogeneous the coefficient of correlation may give a misleading picture of the extent of association. For example, if the scatter diagram shows the points in separate clusters or groups, the correlation coefficient based on all the groups taken together may be very high; yet if seperate values of r are computed for each group, they may be close to zero. If some reasonable basis can be found for separating the data into groups, it is desirable to compute values of r for each group.

9.11 VARIANCE OF THE SUM (DIFFERENCE) OF TWO SERIES

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of n pairs of observations on the variables x and y, and a new series $(x + y)$ is formed by combining the corresponding values of the two series, giving $x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n$. If we denote the variances of x-series, y-series and $(x + y)$-series by $\text{var}(x)$, $\text{var}(y)$ and $\text{var}(x + y)$ respectively, and $\text{cov}(x, y)$ denotes the covariance between x and y, then

$$\text{var}(x + y) = \text{var}(x) + \text{var}(y) + 2 \cdot \text{cov}(x, y) \qquad (9.11.1)$$

Similarly, if another series $(x - y)$ is formed from the given data, then

$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2.\ \text{cov}(x, y) \qquad (9.11.2)$$

In the usual notations, if we write σ_x, σ_y, σ_{x+y}, σ_{x-y} to denote the standard deviations of x, y, $x+y$, $x-y$ respectively, the above formulae may be written as

$$\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 + 2r\sigma_x\sigma_y \tag{9.11.3}$$

$$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y \tag{9.11.4}$$

When x and y are uncorrelated i.e. $r = 0$, then cov $(x, y) = 0$, and the last terms on the right of the above formulae vanish. Hence

$$\text{var}(x + y) = \text{var}(x - y) = \text{var}(x) + \text{var}(y)$$

i.e.
$$\sigma_{x+y}^2 = \sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 \tag{9.11.5}$$

This implies that the variance of the sum (or difference) of uncorrelated variables is equal to the *sum* of their variances.

In general, if $z = ax + by$, where a and b are constants, then

$$\text{var}(ax + by) = a^2.\text{var}(x) + b^2.\text{var}(y) + 2ab.\text{cov}(x, y)$$

i.e.
$$\sigma_x^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\, r\, \sigma_x\sigma_y \tag{9.11.6}$$

Note that formulae (9.11.3) and (9.11.4) may be obtained from this by putting $a = 1, b = 1$, and $a = 1, b = -1$ respectively.

Again, solving (9.11.3) and (9.11.4) for r, we see that

$$r = \frac{\sigma_{x+y}^2 - \sigma_x^2 - \sigma_y^2}{2\sigma_x\sigma_y}, \quad r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y} \tag{9.11.7}$$

Thus, the standard deviations of x, y and $x + y$ (or $x - y$) may be used to find the correlation coefficient between x and y.

Example 9.18 *n pairs of values of two variables x and y are given. The variances of x, y and (x – y) are* σ_x^2, σ_y^2 *and* σ_{x-y}^2 *respectively. Show that the correlation coefficient* r_{xy} *between x and y is given by*

$$\frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

[I.C.W.A., July '66, '72; C.U., B.A..(Econ) '72]

Solution Let (x_i, y_i), $i = 1, 2, \ldots, n$, denote n pairs of observations with means $\bar{x}$, $\bar{y}$ and standard deviations σ_x, σ_y for the x-series and y-series respectively. If we write $u_i = x_i - y_i$, then $\bar{u} = \bar{x} - \bar{y}$.

$$u_i - \bar{u} = (x_i - y_i) - (\bar{x} - \bar{y}) = (x_i - \bar{x}) - (y_i - \bar{y})$$

$$\therefore \quad \sigma_u^2 = \Sigma(u_i - \bar{u})^2/n$$
$$= \Sigma\{(x_i - \bar{x}) - (y_i - \bar{y})\}^2/n$$
$$= \Sigma(x_i - \bar{x})^2/n - 2\Sigma(x_i - \bar{x})(y_i - \bar{y})/n + \Sigma(y_i - \bar{y})^2/n$$
$$= \sigma_x^2 - 2\,.\,\text{cov}(x, y) + \sigma_y^2$$
$$= \sigma_x^2 - 2r_{xy}\,\sigma_x\,\sigma_y + \sigma_y^2,$$

because by definition $r_{xy} = \dfrac{\text{cov}(x, y)}{\sigma_x\,\sigma_y}$ and hence $\text{cov}(x, y) = r_{xy}\,\sigma_x\,\sigma_y$

i.e. $\sigma_{x-y}^2 = \sigma_x^2 - 2r_{xy}\,\sigma_x\,\sigma_y + \sigma_y^2$

or, $2r_{xy}\,\sigma_x\,\sigma_y = \sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2$

or, $r_{xy} = (\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2)/(2\sigma_x\sigma_y)$ (Shown)

Example 9. 19 *The differences $(y_i - x_i)$, $i = 1, 2, ..., 15$, are given for a set of 15 pairs of observations below :*

15, 15, 13, 14, 11, 13, 12, 12, 15, 15, 15, 19, 19, 19, 14.

If $\bar{x} = 49.34$, $\bar{y} = 64.07$, $\sigma_x = 3.53$, $\sigma_y = 4.30$, find the coefficient of correlation between x and y.

[C.U., B.Sc. '64; B.A. (Econ) '67]

Solution We apply the result of the previous Example. If $u_i = y_i - x_i$, then $\bar{u} = \bar{y} - \bar{x}$. In the present case, the mean of the given differences is $(15 + 15 + 13 + 14 + ... + 19 + 14)/15 = 221/15 = 14.73$, which agrees with $\bar{y} - \bar{x} = 64.07 - 49.34 = 14.73$. For the calculation of σ_{y-x}^2, we use deviations from 15 (Table 9.12).

Table 9.12 Calculations for S.D.

Deviation from 15

$y - x$	(z)	z^2
15	0	0
15	0	0
13	− 2	4
14	− 1	1
11	− 4	16
13	− 2	4
12	− 3	9
12	− 3	9
15	0	0
15	0	0
15	0	0
19	4	16
19	4	16
19	4	16
14	− 1	1
Total	− 4	92

$$\sigma_{y-x}^2 = \Sigma z^2/n - (\Sigma z/n)^2$$
$$= 92/15 - (-4/15)^2 = 6.06$$

$\therefore \quad \sigma_{x-y}^2 = \sigma_{y-x}^2 = 6.06$

Now, putting the values

$$r_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

$$= \frac{(3.53)^2 + (4.30)^2 - 6.06}{2 \times 3.53 \times 4.30}$$

$$= \frac{24.8909}{30.3580} = 0.82.$$

Ans. 0.82.

Example 9.20 *If three uncorrelated variables x_1, x_2 and x_3 have the same standard deviation, find the coefficient of correlation between $x_1 + x_2$ and $x_2 + x_3$.*

[B.U., B.A. (Econ)' 71]

Solution Let us write $u = x_1 + x_2$, $v = x_2 + x_3$. It is required to find r_{uv}. Let σ_1, σ_2, σ_3, denote the standard deviations of the variables x_1, x_2, x_3 respectively; and $\text{cov}(x_1, x_2)$, $\text{cov}(x_2, x_3)$, $\text{cov}(x_1, x_3)$ denote the covariances between pairs of variables. Since x_1, x_2, x_3 are uncorrelated.

$$\text{cov}(x_1, x_2) = 0, \quad \text{cov}(x_2, x_3) = 0, \quad \text{cov}(x_1, x_3) = 0$$

Also since the variables are assumed to have the same s.d.,

$$\sigma_1 = \sigma_2 = \sigma_3 = \sigma \text{ (suppose)}$$

Using the formula (9.11.5) for uncorrelated variables

$$\sigma_u^2 = \sigma_1^2 + \sigma_2^2 = 2\sigma^2; \quad \sigma_v^2 = \sigma_2^2 + \sigma_3^2 = 2\sigma^2$$

If $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ denote the means, then $\bar{u} = \bar{x}_1 + \bar{x}_2$ and $\bar{v} = \bar{x}_2 + \bar{x}_3$.

$$\begin{aligned}\text{cov}(u, v) &= \Sigma(u - \bar{u})(v - \bar{v})/n \\ &= \Sigma[\{(x_1 - \bar{x}) + (x_2 - \bar{x}_2)\}\{(x_2 - \bar{x}_2) + (x_3 - \bar{x}_3)\}]/n \\ &= \Sigma(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)/n + \Sigma(x_1 - \bar{x}_1)(x_3 - \bar{x}_3)/n + \Sigma(x_2 - \bar{x}_2)^2/n \\ &\quad + \Sigma(x_2 - \bar{x}_2)(x_3 - \bar{x}_3)/n \\ &= \text{cov}(x_1, x_2) + \text{cov}(x_1, x_3) + \sigma_2^2 + \text{cov}(x_2, x_3) \\ &= 0 + 0 + \sigma^2 + 0 = \sigma^2\end{aligned}$$

Therefore, $$r_{uv} = \frac{\text{cov}(u, v)}{\sigma_u \sigma_v} = \frac{\sigma^2}{2\sigma^2} = 0.5.$$

Example 9.21 *If two variates are independent, their correlation coefficient is zero. Is the converse true? Explain by means of an example.*

[I.C.W.A., June '64, '77]

Solution If two variates are independent, their correlation coefficient is zero. But the converse is *not* true. A zero correlation coefficient does not necessarily signify that the variables are independent. It only implies that there is no linear relationship between the variables. However, the possible existence of a non-linear relationship cannot be ruled out altogether. For example, let us consider the following data:

x	-3	-2	-1	0	1	2	3
y	9	4	1	0	1	4	9

Here $\Sigma x = 0$, $\Sigma y = 28$, $\Sigma xy = 0$, $n = 7$. Therefore, using (9.6.2.),

$$\text{cov}(x, y) = \frac{\Sigma xy}{n} - \left(\frac{\Sigma x}{n}\right)\left(\frac{\Sigma y}{n}\right) = \frac{0}{7} - \left(\frac{0}{7}\right)\left(\frac{28}{7}\right) = 0.$$

Hence, $$r_{xy} = \text{cov}(x, y)/(\sigma_x \sigma_y) = 0/(\sigma_x \sigma_y) = 0$$

i.e. the correlation coefficient between x and y is zero. But it may be noticed that x and y are bound by the relation $y = x^2$. So, x and y are not independent. Thus the correlation coefficient may be zero, even when the variables are not independent.

9.12 REGRESSION

The word "regression" is used to denote *estimation* or *prediction* of the average value of one variable for a specified value of the other variable. The estimation is done by means of suitable equations, derived on the basis of available bivariate data. Such an equation is known as a *Regression equation* an its geometrical representation is called a *Regression curve*.

In *linear regression* (or *simple regression*) the relationship between the variables is assumed to be linear. The estimate of y (say, y') is obtained from an equation of the form

$$y' - \bar{y} = b_{yx}(x - \bar{x}) \qquad (i)$$

and the estimate of x (say, x') from another equation (usually different from the former) of the form

$$x' - \bar{x} = b_{xy}(y - \bar{y}) \qquad (ii)$$

Equation (i) is known as *Regression equation of y on x*, and equation (ii) as *Regression equation of x on y*. The coefficient b_{yx} appearing in the regression equation of y on x is known as the *Regression coefficient of y on x*. Similarly, b_{xy} is called the *Regression coefficient of x on y*. The geometrical representation of linear regression equations (i) and (ii) are known as *Regression lines*. These lines are "best fitting" straight lines obtained by the Method of Least Squares (page 249).

Example 9.22 *Derive the regression equations of y on x and x on y.*

[C.U., B.A. (Econ) '71, '73, '78]

Or,

Obtain the equations of the two lines of regression for a bivariate distribution.

[I.C.W.A., July '70]

Solution (1) *Regression Equation of y on x*

The regression equation of y on x is the equation of the best-fitting straight line in the form $y = a + bx$, obtained by the method of least squares (Section 8.5, page 251).

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of n pairs of observations, and let us fit a straight line of the form

$$y = a + bx \qquad (i)$$

to these data (here, x is considered to be the independent variable and y and dependent variable, i.e. a value of y is obtained when a value of x is given). Applying the method of least squares, the constants a and b are obtained by solving the normoal equations (8.5.2), viz.

$$\Sigma y = an + b\,\Sigma x \qquad (ii)$$

$$\Sigma xy = a\,\Sigma x + b\,\Sigma x^2 \qquad (iii)$$

Dividing both sides of (ii) by n, we get $\bar{y} = a + b\bar{x}$ so that $a = \bar{y} - b\bar{x}$.

Substituting this in equation (*i*),

$$y - \bar{y} = b(x - \bar{x}) \qquad (iv)$$

Again, multiplying (ii) by Σx and (iii) by n, we have

$$(\Sigma x)(\Sigma y) = na(\Sigma x) + b(\Sigma x)^2$$

$$n(\Sigma xy) = na(\Sigma x) + nb(\Sigma x^2)$$

Subtracting the first from the second,

$$n\Sigma xy - (\Sigma x)(\Sigma y) = b\{n\,\Sigma x^2 - (\Sigma x)^2\}$$

$$\therefore \quad b = \frac{n\,\Sigma xy - (\Sigma x)\,(\Sigma y)}{n\,\Sigma x^2 - (\Sigma x)^2}$$

Dividing both numerator and denominator by n^2,

$$b = \frac{\Sigma xy/n - (\Sigma x/n)\,(\Sigma y/n)}{n\,\Sigma x^2/n - (\Sigma x/n)^2} = \frac{\text{cov}(x,\ y)}{\sigma_x^{\ 2}} \qquad \text{(v)}$$

Writing b with the usual subscripts, we have from (iv),

$$y - \bar{y} = b_{yx}\,(x - \bar{x}) \qquad \text{(vi)}$$

where $b_{yx} = \text{cov}\,(x, y)/\sigma_x^{\ 2}$. This is the required regression equation of y on x. Since $r = \text{cov}(x, y)/(\sigma_x\sigma_y)$, we see that $\text{cov}(x, y) = r\sigma_x\,\sigma_y$.

Substituting this

$$b_{yx} = \frac{\text{cov}(x,\ y)}{\sigma_x^{\ 2}} = r\,\frac{\sigma_y}{\sigma_x} \qquad \text{(vii)}$$

(2) *Regression Equation of x on y*

If we fit an equation of the form $x = a' + b'\,y$ (assuming x to be the dependent variable and y the independent variable), we obtain the regression equation of x on y. Applying the method of least squares, the normal equations for determined a' and b' are

$$\Sigma x = a'\,n + b'\,\Sigma y; \quad \Sigma xy = a'\,\Sigma y + b'\,\Sigma y^2 \qquad \text{(viii)}$$

Proceeding the same way as before, the regression equation of x on y is found to be

$$x - \bar{x} = b_{xy}\,(y - \bar{y}) \qquad \text{(ix)}$$

where $$b_{xy} = \frac{\text{cov}(x,\ y)}{\sigma_y^{\ 2}} = r\,\frac{\sigma_x}{\sigma_y} \qquad \text{(x)}$$

9.13 PROPERTIES OF LINEAR REGRESSION

The regression equation of y on x is

$$y - \bar{y} = b_{yx}\,(x - \bar{x}) \qquad (9.13.1)$$

wehre $b_{yx} = \dfrac{\text{cov}(x,\ y)}{\sigma_x^{\ 2}} = r\,\dfrac{\sigma_y}{\sigma_x}$. This equation is used to estimate y, when the value of x is known.

The regression equation of x on y is

$$x - \bar{x} = b_{xy}\,(y - \bar{y}) \qquad (9.13.2)$$

where $b_{xy} = \dfrac{\text{cov}(x,\ y)}{\sigma_y^{\ 2}} = r\,\dfrac{\sigma_x}{\sigma_y}$. This equation is used to estimate x, when the value of y is known.

Substituting the values of b_{yx} and b_{xy}, the regression equations (9.13.1) and (9.13.2) may also be written as

$$\frac{y - \bar{y}}{\sigma_y} = r\left(\frac{x - \bar{x}}{\sigma_x}\right) \quad \text{and} \quad \frac{x - \bar{x}}{\sigma_x} = r\left(\frac{y - \bar{y}}{\sigma_y}\right) \qquad (9.13.3)$$

respectively.

Example 9.25 *From the following results, obtain the two regression equations and estimate the yield of crops when the rainfall is 22 cms, and the rainfall when the yield is 600 kg:*

	y (Yield in kg.)	*x (Rainfall in cm.)*
Mean	*508.4*	*26.7*
S.D.	*36.8*	*4.6*

Coefficient of correlation between yield and rainfall = 0.52

[C.A., May '76]

Solution For estimating yield (y) we have to use the regression equation of y on x, and for rainfall (x) the regression equation of x on y. Given $\bar{x} = 26.6$, $\bar{y} = 508.4$, $\sigma_x = 4.6$, $\sigma_y = 36.8$, $r = 0.52$,

$$b_{yx} = 0.52 \times 36.8/4.6 = 4.16,$$
$$b_{xy} = 0.52 \times 4.6/36.8 = .065$$

The regression equations are therefore $y - 508.4 = 4.16\ (x - 26.7)$ and $x - 26.7 = .065(y - 508.4)$ respectively, which simplify to

$$y = 4.16x + 397.33, \text{ and } x = .065y - 6.346$$

When $x = 22$, $y = 4.16 \times 22 + 397.33 = 488.8$ kg.

When $y = 600$, $x = .065 \times 600 - 6.346 = 32.7$ cm.

Example 9.26 *The following results were obtained from records of age (x) and systolic blood pressure (y) of a group of 10 women:*

		x	y
Mean	...	*53*	*142*
Variance	...	*130*	*165*

$$\Sigma(x - \bar{x})(y - \bar{y}) = 1220$$

Find the appropriate regression equation and use it to estimate the blood pressure of a women whose age is 45. [I.C.W.A., Jan. '73]

Solution The appropriate equation is the regression equation of y on x, viz. $y - \bar{y} = b_{yx}(x - \bar{x})$, where $b_{yx} = \text{cov}(x, y)/\sigma_x^2$.

But $\text{cov}(x, y) = \Sigma(x - \bar{x})(y - \bar{y})/n = 1220/10 = 122$,

so that $b_{yx} = 122/130 = 0.94$.

The regression equation is therefore $y - 142 = 0.94\ (x - 53)$; or, $y = 0.94x + 92.18$. When $x = 45$, $y = 0.94 \times 92.18 = 134.5$. The estimated blood pressure is 134.5.

Ans. $y = 0.94x + 92.18$; 134.5

Example 9.27 *Marks obtained by 12 students in the college test (x) and the university test (y) are as follows:*

x	41	45	50	68	47	77	90	100	80	100	40	43
y	60	63	60	48	85	56	53	91	74	98	65	43

What is your estimate of the marks a student could have obtained in the university test if he obtained 60 in the college test but was ill at the time of the university test?

[B.U., B.A. (Econ) '70]

Solution For estimating y, we use the regression equation of y on x, *viz.* $y - \bar{y} = b_{yx}(x - \bar{x})$, where $b_{yx} = \text{cov}(x, y)/\sigma_x^2$. Therefore, we have to find $\bar{x}$, $\bar{y}$, cov(x, y) and σ_x^2.

Table 9.13 Calculations for Regression

x	y	$X = x - 60$	$Y = y - 60$	X^2	XY
41	60	– 19	0	361	0
45	63	– 15	3	225	– 45
50	60	– 10	0	100	0
68	48	8	– 12	64	– 96
47	85	– 13	25	169	– 325
77	56	17	– 4	289	– 68
90	53	30	– 7	900	– 210
100	91	40	31	1,600	1,240
80	74	20	14	400	280
100	98	40	38	1,600	1 520
40	65	– 20	5	400	– 100
43	43	– 17	– 17	289	289
Total 781	796	61	76	6,397	2,485

$\bar{x} = 781/12 = 65.08$, $\bar{y} = 796/12 = 66.33$

Since the variance and covariance are unaffected by changes of origin,

$$\text{cov}(x, y) = \text{cov}(X, Y) = \Sigma XY/n - (\Sigma X/n)(\Sigma Y/n)$$
$$= 2485/12 - (61/12)(76/12) = 25184/144$$
$$\sigma_x^2 = \Sigma X^2/n - (\Sigma X/n)^2 = 6397/12 - (61/12)^2 = 73043/144$$

Therefore, $$b_{yx} = \frac{25184/144}{73043/144} = \frac{25184}{73043} = 0.345$$

The regression equation is $y - 66.33 = 0.345(x - 65.08)$;

i.e. $$y = 43.88 + 0.345x$$

When $x = 60$, $y = 43.88 + 0.345 \times 60 = 64.58 = 65$ (approx.)

Ans. $y = 43.88 + 0.345x$; 65

[**Note:** In Examples 9.26 and 9.27, the use of formula $b_{yx} = r\sigma_y / \sigma_x$ would involve laborious calculations for r, and must be avoided.]

Example 9.28 *Derive the regression line which you consider more important from the following series of observations :*

Output (thousand)	5	7	9	11	13	15
Profit per unit of Output (Rs)	1.70	2.40	2.80	3.40	3.70	4.40

[B.U., B.A.(Econ) '66; C.U., B.A. (Econ) '71; I.C.W.A., June '73]

Solution Let x represent 'Output (thousand)' and y represent 'Profit per unit of output (Rs.)'. It is considered that the regression line of y on x is the more important; because, using the equation it will be possible to find the expected profit based on output of the firm.

[**Note:** Here the successive values of x are equidistant. Also, since the number of pairs of observations is even, viz. 6, we apply the transformation (8.6.4) at p. 252. for simplifying the calculations]

Table 9.14 Calculations for Regression

x	y	$u = x - 10$	u^2	uy
5	1.70	− 5	25	− 8.50
7	2.40	− 3	9	− 7.20
9	2.80	− 1	1	− 2.80
11	3.40	1	1	3.40
13	3.70	3	9	11.10
15	4.40	5	25	22.00
Total 60	18.40	0	70	18.00

Proceeding as in Example 9.27, $\bar{x} = 60/6 = 10$, $\bar{y} = 18.406 = 3.07$

$$\text{cov}(x, y) = \text{cov}(u, y) = \Sigma uy/n - (\Sigma u/n)(\Sigma y/n)$$
$$= 18.00/6 - (0/6)\,(18.40/6) = 3.00$$
$$\sigma_x^2 = \sigma_u^2 = \Sigma u^2/n - (\Sigma u/n)^2 = 70/6 - (0/6)^2 = 70/6$$
$$b_{yx} = 18.00/70 = 0.257$$

The regression equation of y on x is $y - \bar{y} = b_{yx}\,(x - \bar{x})$;

i.e., $y - 3.07 = 0.257(x - 10)$; or, $y = 0.257x + 0.50$

Example 9.29 *Find the equation of the line of regression of x on y for the following data:*

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	5.3	5.7	6.3	7.2	8.2	8.7	8.4

[C.U., B.A. (Econ) '72]

Solution The regression equation of x on y is $x - \bar{x} = b_{xy}\,(y - \bar{y})$, where $b_{xy} = \text{cov}(x, y)/\sigma_y^2$. For simplifying the calculations, let us make a change of origin and scale for both the variables

$$u = \frac{x - 2.5}{0.5}, \quad v = \frac{y - 7.0}{0.1}$$

[**Note:** Here the values of x are equidistant and n is odd. So, we use the transformation (8.6.3), page 252].

$$\bar{x} = 17.5/7 = 2.5, \quad \bar{y} = 49.8/7 = 7.11$$
$$\text{cov}(u, v) = \Sigma uv/n - (\Sigma u/n)\,(\Sigma v/n) = 172/7 - (0/7)(8/7) = 172/7$$
$$\sigma_v^2 = \Sigma v^2/n - (\Sigma v/n)^2 = 1140/7 - (8/7)^2 = 7916/49$$

Using (9.6.4), $\text{cov}(x, y) = d.d.' \text{cov}(u, v) = (0.5)\,(0.1)(172/7)$

Table 9.15 Calculations for Regression

x	y	u	v	v^2	uv
1.0	5.3	− 3	− 17	289	51
1.5	5.7	− 2	− 13	169	26
2.0	6.3	− 1	− 7	49	7
2.5	7.2	0	2	4	0
3.0	8.2	1	12	144	12
3.5	8.7	2	17	289	34
4.0	8.4	3	14	196	42
Total 17.5	49.8	0	8	1,140	172

Using (6.7.4), $\sigma_y^2 = d^2.\ \sigma_v^2 = (0.1)^2\ (7916/49)$

Hence, $$b_{xy} = \frac{\text{cov}(x,\ y)}{\sigma_y^{\ 2}} = \frac{0.5 \times 0.1 \times (172/2)}{(0.1)^2 \times (7916/49)} = 0.76$$

Therefore, the regression equation of x on y is

$$x - 2.5 = 0.76(y - 7.11);\quad \text{i.e.}\quad x = 0.76y - 2.90.$$

Example 9.30 *From the following bivariate frequency distribution, calculate (i) the coefficient of correlation and (ii) the regression equation of marks (y) on age (x)–*

Marks \ Age (years)	15–20	20–25	25–30	30–35	35–40	Total
0–19	4	2				6
20–39	6	5	3	1		15
40–59		9	4	2	1	16
60–79		7	4	1		12
80–99			1			1
Total	10	23	12	4	1	50

Solution Using the results from Table 9.16, we have

$$\sigma_u^{\ 2} = \frac{\Sigma f_1 u^2}{N} - \left(\frac{\Sigma f_1 u}{N}\right)^2 = \frac{71}{50} - \left(\frac{-37}{50}\right)^2 = 0.8724$$

Table 9.16 Calculations for Correlation and Regression

x_u / y_v	17.5 / −2	22.5 / −1	27.5 / 0	32.5 / 1	37.5 / 2	Totals f_2	f_2v	f_2v^2	fuv
9.5–2	4	2				6	− 12	24	20
	(16)	(4)							
29.5–1	6	5	3	1		15	− 15	15	16
	(12)	(5)	(0)	(− 1)					
49.5–0		9	4	2	1	16	0	0	0
		(0)	(0)	(0)	(0)				
69.5 – 1		7	4	1		12	12	12	− 6
		(− 7)	(0)	(1)					
89.5 – 2			1			1	2	4	0
			(0)						
Totals									
f_1	10	23	12	4	1	50 = N	− 13	55	30
f_1u	− 20	−23	0	4	2	− 37			
f_1u^2	40	23	0	4	4	71			
fuv	28	2	0	0	0	30			

Explanation of symbols used in Table 9.16:

(i) x and y represent mid-values of the class intervals of age and marks respectively.

(ii) f_1 and f_2 represent marginal frequencies of the distribution of x and y respectively.

(iii) $u = \dfrac{x - 27.5}{5}$ and $v = \dfrac{y - 49.5}{20}$.

(iv) Number within brackets in each cell is the product of that cell frequency and the corresponding values of u and v.

(v) fuv represent totals of the numbers within brackets, mentioned at (iv) above, in any row or column.

$$\sigma_v^2 = \frac{\Sigma f_2 v^2}{N} - \left(\frac{\Sigma f_2 v}{N}\right)^2 = \frac{55}{50} - \left(\frac{-13}{50}\right)^2 = 1.0324$$

$$\text{cov}(u, v) = \frac{\Sigma fuv}{N} - \left(\frac{\Sigma f_1 u}{N}\right)\left(\frac{\Sigma f_2 v}{N}\right) = \frac{30}{50} - \left(\frac{-37}{50}\right)\left(\frac{-13}{50}\right) = 0.4076$$

$$\sigma_u = \sqrt{0.8724} = 0.93\,,\ \sigma_v = \sqrt{1.0324} = 1.02$$

$$\therefore \quad r_{xy} = r_{uv} = \frac{\text{cov}\,(u, v)}{\sigma_u\,\sigma_v} = \frac{0.4076}{0.93 \times 1.02} = 0.43$$

Again, since $u = (x - 27.5)/5$ and $v = (y - 49.5)/20$,

$$\bar{x} = 27.5 + 5(-37/50) = 23.8$$

$$\bar{y} = 49.5 + 20(-13/50) = 44.3$$

Applying (9.6.4) and (6.6.3),

$$\text{cov}(x, y) = 5 \times 20 \times \text{cov}(u, v) = 100 \times 0.4076$$

$$\sigma_x^2 = 5^2 \times \sigma_u^2 = .25 \times 0.8724$$

$$\therefore \quad b_{yx} = \frac{\text{cov}(x, y)}{\sigma_x^2} = \frac{100 \times 0.4076}{25 \times 0.8724} = 1.87$$

The regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

i.e. $y - 44.3 = 1.87(x - 23.8)$

or, $y = 1.87x - 0.21$

Ans. $r = +0.43$, $y = 1.87x - 0.21$

Example 9.31 *Let the lines of regression concerning two variables x and y be given by $y = 32 - x$ and $x = 13 - 0.25y$. Obtain the values of the means and the correlation coefficient.* [C.U., B.Sc. '71]

Solution Since the regression lines intersect at $(\bar{x}, \bar{y})$, the means will be obtained by solving the two equations. Solving $y = 32 - x$ and $x = 13 - 0.25y$, we get $x = 6.7$ and $y = 25.3$. So, $\bar{x} = 6.7$, $\bar{y} = 25.3$.

$y = 32 - x$ is the regression equation of y on x; therefore the 'regression coefficient of y on x' is the coefficient of x on the right; i.e. $b_{yx} = -1$. Similarly, $x = 13 - 0.25y$ being the regression equation of x on y, the 'regression coefficient of x on y' is the coefficient of y on the right; i.e. $b_{xy} = -0.25$.

$$r^2 = b_{yx} \times b_{xy} = (-1)(-0.25) = 0.25;$$

$$r = \sqrt{0.25} = \pm 0.5$$

But since the regression coefficients are negative (note that both must have the same sign), the correlation coefficient also must be negative, i.e. $r = -0.5$.

Ans. $\bar{x} = 6.7$, $\bar{y} = 25.3$, $r = -0.5$.

Example 9.32 *If x and y satisfy the relationship $y = -5 + 6x$, what is the product-moment correlation?*

[B.U., B.A. (Econ) '73]

Solution Since the relationship between x and y is exact and *linear*, there is perfect correlation between the variables; i.e. $r = \pm 1$. But the slope of the straight line (viz. + 6) is positive. Hence r must be positive, and $r = +1$.

[**Note:** We speak of two types of relationship between the variables—(i) 'exact' or 'functional' relationship, and (ii) 'statistical' relationship. If to any specified value of one variable there corresponds some definite value or values of another variable, the relationship is exact. However, in many practical situations, no such relationship exists, e.g. between Height and Weight, or Income and Expenditure, etc. We then try to find the best possible estimates of one variable by means of an equation connecting them. This equation, known as 'Regression Equation', shows the statistical relationship between the variables.]

Example 9.33 *Regression of savings(s) of a family on income (y) may be expressed as $s = a + \frac{y}{m}$, where a and m are constants. In a random sample of 100 families the variance of savings in one-quarter of the variance of incomes and the correlation is found to be 0.4. Obtain the estimate of m.*

[I.C.W.A., June '74-old]

Solution The given regression equation of s on y can be written as $s = a + (1/m)y$, which is the equation of a straight line. Therefore the regression coefficient of s on y is the coefficent of y on the right;

i.e. $$b_{sy} = 1/m \qquad \text{(i)}$$

But $$b_{sy} = r\frac{\sigma_s}{\sigma_y} \qquad \text{(ii)}$$

However, we are given that $\sigma_s^2 = (1/4)\,\sigma_y^2$ and $r = 0.4$, so that $\frac{\sigma_s}{\sigma_y} = \frac{1}{2}$

Hence, using (ii), $$b_{sy} = (0.4) \times \frac{1}{2} = 0.2$$

Now, from (i), $$0.2 = 1/m; \quad \therefore \quad m = 1/0.2 = 5$$

Example 9.34 *You are given that the variance of x is 9. The regression equations are $8x - 10y + 66 = 0$ and $40x - 18y = 214$. Find (i) Average values of x and y, (ii) Correlation coefficient between the two variables, (iii) Standard deviation of y.*

[C.A., Nov. '77; I.C.W.A., June '77]

Solution

[**Note:** The regression equations of y on x and of x on y are usually shown in the forms $y = a + bx$ and $x = a' + b'y$ respectively, and in such cases it is not difficult to find the regression coefficients (Example 9.31). If however the equations are shown otherwise, they should be brought to the usual forms and then the regression coefficients can be easily determined. In the present problem, it is not known which one is the regression equation of y on x and which one the regression equation of x on y. A method has been shown for identifying the appropriate equations.]

Since the regression lines always intersect at the point $(\bar{x}, \bar{y})$, the averages of x and y will be obtained by solving the two equations [see Example 8.1 (g), p. 248]. Multiplying the first equation by 5 and then subtracting the result from the second equation, we have $32y = 544$, so

that $y = 17$. Now putting this value in any of the given equations and solving we get $x = 13$. Therefore, the averages of x and y are 13 and 17 respectively.

Let us assume that $8x - 10y + 66 = 0$ represents the regression equation of x on y, and $40x - 18y = 214$ the regression equation of y on x. These equations can be re-written as $x = (5/4)y - (33/4)$ and $y = (20/9)x - (107/9)$ respectively. The regression coefficients should then be $b_{xy} = 5/4$ and $b_{yx} = 20/9$.

$$\therefore \qquad r^2 = b_{xy} \cdot b_{yx} = (5/4)(20/9) = 25/9$$

This result is impossible, because r cannot exceed 1 numberically. So, our assumptions must be wrong.

The correct position must be the other way round; namely. $8x - 10y + 66 = 0$ must be the regression equation of y on x, and $40x - 18y = 214$ the regression equation of x on y. The two equations when expressed in the usual forms give $y = (4/5)x + (33/5)$ and $x = (9/20)y + (107/20)$. The correct values of the regression coefficients are $b_{yx} = 4/5$, $b_{xy} = 9/20$.

i.e., $$r = \sqrt{(4/5)(9/20)} = +0.6$$

(The positive value of the square-root is taken, since the regression coefficients are positive).

Again, given $\sigma_x^2 = 9$, i.e. $\sigma_x = 3$, using the relation $b_{yx} = 4/5$,

we have $$r\frac{\sigma_y}{\sigma_x} = 0.8; \quad \text{or} \quad 0.6 \times \frac{\sigma_y}{3} = 0.8$$

This gives $\sigma_y = 4$.

Ans. 13, 17; + 06; 4

Example 9.35 *Prove that the coefficient of correlation is the geometric mean of the coefficients of regression.*

[I.C.W.A., Jan. '65, '69, June '74, Dec. '76; B.U., B.A., (Econ) '66; C.U., M.Com. '64, '68, '71]

Solution The regression coefficients b_{yx} and b_{xy} may be expressed in terms of the correlation coefficient (r) and the standard deviations of x and y (viz. σ_x and σ_y) by the relations

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}, \quad b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$\therefore \qquad b_{yx} \times b_{xy} = \left(r\frac{\sigma_y}{\sigma_x}\right)\left(r\frac{\sigma_x}{\sigma_y}\right) = r^2$$

or, $$r^2 = b_{yx} \times b_{xy}; \quad \therefore \quad r = \sqrt{b_{yx} \times b_{xy}}$$

Thus, the correlation coefficient r is the geometric mean of the two regression coefficients b_{yx} and b_{xy}.

Example 9.36 *What are regression lines? Explain why we have two regression lines and why these two lines are identical if r the correlation coefficient is + 1 or – 1.*

[C.U., B.A. (Econ) '63; M.Com. '66; I.C.W.A., Jan. '73, Dec. '77]

Solution If bivariate data are plotted as points on a graph papaer, it will be found that the concentration of points follows a certain pattern showing the relationship between the variables. When the trend of points is found to be linear, we determine the best-fitting straight line by the Method of Least Squares. Such straight lines which are used to obtain best estimates of one variable for given values of the other, are called *regression lines.* If x is considered as the independent variable and y the dependent variable, a linear equation of the form $y = a + bx$ is fitted to the pairs of observations (x_1, y_1), (x_2, y_2), ... (x_n, y_n) leading to the equation

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \text{(i)}$$

This is known as the *regression line of y on x*. If however, the variable x is considered to be dependent on y, a linear equation of the form $x = a' + b'y$ is fitted, yielding

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \text{(ii)}$$

This is known as the *regression line of x on y*. Equation (i) is used to obtain best estimates of y for given values of x, and equation (ii) to obtain best estimates of x for given values of y.

There are two regression lines, which are separately used to obtain best estimates of each variable. The same equation can not serve both the purposes; because the two equations (i) and (ii) are derived under two different assumptions. In deriving equation (i) it is assumed that the values of x are known exactly and those of y are subject to error. In deriving equation (ii) the assumption is just the reverse, i.e. the values of y are known exactly and those of x only are subject to error.

Since the two regression lines (i) and (ii) both pass through $(\bar{x}, \bar{y})$, they will be identical if their slopes are equal, i.e. if $b_{yx} = 1/b_{xy}$

or, $r\frac{\sigma_y}{\sigma_x} = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$; simplifying, we get $r = \pm 1$.

9.14 EXPLAINED VARIATION AND UNEXPLAINED VARIATION

If y_i' represents the *estimated* value of y from the regression equation of y on x (note that y_i denotes the *observed* value) when $x = x_i$, i.e., $y_i' - \bar{y} = b_{yx}(x_i - \bar{x})$, then it can be shown that

$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - y_i') + \Sigma(y_i' - \bar{y})^2$$

$\Sigma(y_i - \bar{y})^2$ is called *Total variation* of the observed values of y. $\Sigma(y_i - y_i')^2$ is the sum of the squares of vertical distances of the points on the scatter diagram around the regression line of y on x, and as such is called *Variation around the regression line* or *Unexplained variation* or *Residual variation*. $\Sigma(y_i' - \bar{y})^2$ is called *Variation due to regression* or *Explained variation.*

It can be shown that

$$\frac{\textit{Explained variation}}{\textit{Total variation}} = r^2$$

i.e. *Proportion of Total variation explained by regression* $= r^2$

We have, thus, a new interpretation of the correlation coefficient, viz. the square of the correlation coefficient is equal to the proportion of total variation explained by regression.

When $r^2 = 1$, the whole of the total variation is explained by regression, so that the unexplained variation is zero and each $y_i = y_i'$. All the points on the scatter diagram lie on the regression line of y on x (the two regression lines now coincide), and there is *perfect linear dependence* between the variables. For a given value of x, we have a fixed value of y.

EXERCISES

1. The data given below relate to the heights and weights of 20 persons. You are rquired to form a two-way frequency table with class-intervals 62″ to 64″, 64″ to 66″ and so on, and 115 to 125 lbs., 125 to 135 lbs., and so on.

S. No.	1	2	3	4	5	6	7	8	9	10	11	12	13
Height	70	65	65	64	69	63	65	70	71	62	70	67	63
Weight	170	135	136	137	148	124	117	128	143	129	163	139	122
							14	15	16	17	18	19	20
							68	67	69	66	68	67	67
							134	140	132	120	148	129	152

[C.A., May '66]

2. Calculate r from the following given results: $n = 10$; $\Sigma x = 125$; $\Sigma x^2 = 1585$; $\Sigma y = 80$; $\Sigma y^2 = 650$; $\Sigma xy = 1007$.

3. Find the coefficient of correlation from the following results:

$$\sum_{i=1}^{8} X = 42.2, \sum_{i=1}^{8} Y = 46.4, \sum_{i=1}^{8} X^2 = 291.20,$$

$$\sum_{i=1}^{8} Y^2 = 290.52, \sum_{i=1}^{8} XY = 230.42$$

4. Obtain the correlation coefficient from the following:

$x -$	6	2	10	4	8
$y -$	9	11	5	8	7

[D.S.W., '77]

5. Calculate the coefficient of correlation for the ages of husband and wife :

Age of Husband	23	27	28	29	30	31	33	35	36	39
Age of Wife	18	22	23	24	25	26	28	29	30	32

[I.C.W.A., Jan. '70]

6. Calculate the correlation coefficient r_{xy} from the following:

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

[D.M., '78]

7. Calculate the coefficient of correlation between x and y:

x	155	157	153	151	159	162	158
y	118	129	125	124	129	133	127

[C.U., B.A. (Econ) '75]

8. Calculate Pearson's coefficient of correlation from the following data using 44 and 26 as the origins of X and Y respectively :

X	43	44	46	40	44	42	45	42	38	40	42	57
Y	29	31	19	18	19	27	27	29	41	30	26	10

[C.A., May '78]

9. Specimens of similarly treated alloy steel containing various percentages of nickel are tested for toughness with the following results:

Toughness (arbitrary units)	47	50	52	52	54	56	58	59
Percentage of Nickel	2.7	2.7	2.8	2.8	2.9	3.2	3.2	3.3
			60	60	62	64	65	66
			3.4	3.5	3.5	3.6	3.7	3.8

Find the coefficient of correlation between 'toughness' as measured by the test, and 'percentage content of Nickel' in the alloy steel.

[C.U., M. Com. '65; D.S.W., May '71]

10. Calculate the coefficient of correlation from the following data:

Export of Raw Cotton (Rs crores)	42	44	58	55	89	98	66
Import of Manufactured Goods (Rs crores)	56	49	53	58	65	76	58

Calculate also the standard error of the coefficient of correlation.

[I.C.W.A., July '65]

11. Determine the correlation coefficient between x and y:

x	5	7	9	11	13	15
y	1.7	2.4	2.8	3.4	3.7	4.4

[Dip. Management, '67]

12. From the following figures, calculate the coefficient of correlation between the income and the general level of prices:

Income (X)	360	420	500	550	600	640	680	720	750
General Level of Prices (Y)	100	104	115	160	180	290	300	320	330

[C.U., M. Com. '68]

13. The following data give the hardness (x) and tensile strength (y) for some specimens of a material, in certain units. Find the correlation coefficient and calculate its probable error:

x	23.3	17.5	17.8	20.7	18.1	20.9	22.9	20.8
y	4.2	3.8	4.6	3.2	5.2	4.7	4.4	5.6

[I.C.W. A., Jan. '72]

14. The following table gives the saving-bank deposits in billions of dollars and strikes and lock-outs in thousands over a number of years. Compute the correlation coefficient and comment on the results.

Saving Deposits	5.1	5.4	5.5	5.9	6.5	6.0	7.2
Strickes and Lock-outs	3.8	4.4	3.3	3.6	3.3	2.3	1.0

[I.C.W.A., Jan. '64]

15. Two positively correlated variables x_1 and x_2 have variances σ_1^2 and σ_2^2 respectively. Determine the value of the constant a such that $x_1 + ax_2$ and $x_1 + \frac{\sigma_1}{\sigma_2} x_2$ are uncorrelated.

[B.U., B.A.(Econ) '72]

16. State whether the correlation will be positive or negative in the following cases:
    (i) Age and income;
    (ii) Speed of an automobile and the distance required to stop the car after applying brakes;
    (iii) Sale of woollen garments and day-temperature;
    (iv) Sale of cold-drinks and day-temperature;
    (v) Production and price per unit.

17. Given $\Sigma x = 56$, $\Sigma y = 40$, $\Sigma x^2 = 524$
    $\Sigma y^2 = 256$, $\Sigma xy = 364$, $n = 8$,
    find (i) the correlation coefficient and (ii) the regression equation of x on y.
    [I.C.W.A., July '67]

18. The following sums have been obtained from 100 observations-pairs:
    $\Sigma x = 12{,}500$, $\Sigma y = 8{,}000$, $\Sigma x^2 = 1{,}585{,}000$,
    $\Sigma y^2 = 648{,}100$, $\Sigma xy = 1{,}007{,}425$.
    (i) Find the regression of y on x, and estimate the value of y when $x = 130$.
    (ii) Compute the correlation coefficient (r) between x and y and state what you learn from the value of r obtained by you.
    [C.U., B.A. (Econ) '76]

19. Given the following totals for 10 pairs of observtions on two characters x and y, obtain the two regression equations and hence calculate the correlation coefficient:
    $\Sigma x = 12$, $\Sigma y = 4$, $\Sigma x^2 = 16.20$, $\Sigma y^2 = 1.96$, $\Sigma xy = 5.2$
    [M.B.A. '79, D.M. '77]

20. Estimate from the information given below, the probable crop yield, when rainfall is 29 inches:

	Mean	*S.D.*
Rainfall in inches	25	3
Yield in units per acre	40	6

    Coefficient of correlation between the variables: 0.65
    [C.U., B.SC. '73]

21. The correlation coefficient between two variates x and y is $r = 0.60$. If $\sigma_x = 1.50$, $\sigma_y = 2.00$, $\bar{x} = 10$, $\bar{y} = 20$, find the equations of the regression lines of (i) y on x, (ii) x on y.
    [I.C.W.A., Dec. '77]

22. The following data pertain to the marks in two subjects, say A and B.
    Mean marks in $A = 39.5$, Mean marks in $B = 47.5$
    S.D. of marks in $A = 10.8$, S.D. of marks in $B = 16.8$
    Coefficient of correlation between marks in A and $B = 0.42$
    Obtain the equations of the two regression lines and then estiamte the marks in B for candidates who secured 50 marks in A.
    [I.C.W.A., June '78]

23. Given the following results of the height and weight of 1000 men students: $\bar{x} = 68$ inches, $\bar{y} = 150$ lbs., $r = 0.60$, $\sigma_x = 2.50$ inches, $\sigma_y = 20.000$ lbs. John

Doe weights 200 lbs., Richard Roe is five feet tall. Estimate the height of Doe from his weight, and weight of Roe from his height.

[C.U., M. Com. '72, '76]

24. From the following data find the coefficient of linear correlation between X and Y. Determine also the regression line of Y on X, and then make an estimate of the value of Y when $X = 12$.

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

[I.C.W.A., June '75]

25. Obtain the lines of regression for the following data:

(X)	1	2	3	4	5	6	7	8	9
(Y)	9	8	10	12	11	13	14	16	15

[C.U., M.Com. '68; D.S.W. '78; I.C.W.A., Dec. '78]

26. Find the two lines of regression from the following data:

Age of Husband (x)	25	22	28	26	35	20	22	40	20	18
Age of Wife (y)	18	15	20	17	22	14	16	21	15	14

Hence, estimate (i) the age of hushand when the age of wife is 19, (ii) the age of wife when the age of hushand is 30.

[C.U., M.Com. '70]

27. From the following data, obtain the two regression equations:

Sales	91	97	108	121	67	124	51	73	111	57
Purchases	71	75	69	97	70	91	39	61	80	47

[C.A., May '77]

28. Obtain the equation of the line of regression of yield of rice (y) on water (x) from the data given in the following table:

Water in Inches (x)	12	18	24	30	36	42	48
Yield in Tons (y)	5.27	5.68	6.25	7.21	8.02	8.71	8.42

Estimate the most probable yield of rice for 40 inches of water.

[C.U., M.Com. '64; I.C.W. A., Dec. '76]

29. If the regression equation of y on x be $Y = 0.57x + 6.93$ and the regression equation of x on y be $X = 1.12y - 2.46$, find the correlation coefficient between x and y.

[B.U., B.A. (Econ) '72]

30. For some bivariate data the following results were obtained. The mean value of $X = 53.2$, the mean value of $Y = 27.9$, the regression coefficient of Y on $X = -1.5$, and the regression coefficient of X on $Y = -0.2$. Find the (i) most probable value of Y when $X = 60$, (ii) r the coefficient of correlation between X and Y.

[C.U., M.Com '74]

31. The regression equation calculated from a given set of observations are $x = -0.2y + 4.2$, $y = -0.8x + 8.4$. Calculate (i) x and y, (ii) r, (iii) the estimated value of y when $x = 4$. *[I.C.W.A., Jan. '68; C.U., B Com. (Hons) '69]*

32. The two regression lines involving the two variables x and y are $Y = 5.6 + 1.2x$ and $X = 12.5 + 0.6y$. Find the means of x and y and their correlation coefficient.

[W.B.H.S., '78]

33. Two vairates have the least squares regression lines $x + 4y + 3 = 0$ and $4x + 9y + 5 = 0$. Find their mean values and the correlation coefficient.

[I.C.W.A., July '70; M.B.A. '78]

34. Two lines of regression are given by $x = 2y = 5$ and $2x + 3y = 8$, and $\sigma_x^2 = 12$. Calculate the values of $\bar{x}$, $\bar{y}$, σ_y and r.

[I.C.W.A., Dec '76-old]

35. In order to find the correlation coefficient between two variates x and y from 12 pairs of observations, the following calculations were made:
$\Sigma x = 30$, $\Sigma y = 5$, $\Sigma x^2 = 670$, $\Sigma y^2 = 285$, $\Sigma xy = 334$.
On subsequent varification it was found that the pair ($x = 11$, $y = 4$) was copied wrongly, the correct values being ($x = 10$, $y = 14$). Find the correct value of correlation coefficient.

[I.C.W.A., June '75-old]

36. Obtain the linear regression equation that you consider more relevant for the following set of paired observations and given reasons why you consider it to be so:

Age	56	42	72	36	63	47	55	49	38	42	68	60
Blood Pressure	147	125	160	118	149	128	150	145	115	140	152	155

Also estimate the blood pressure of a person whose age is 45.

[C.U., M.Com., '73]

37. For the variables x and y the equations of two regression linear are $4x - 5y + 33 = 0$ and $20x - 9y = 107$. Identify the regression line of y on x and that of x on y. What is the estimated value of y, when $x = 10$? If this estimate is denoted by y_0, find the estimated value of x when $y = y_0$

[C.U., B.A. (Econ) '75]

38. State the meaing of the terms *explained variation* and *unexplained variation*, used in the theory of regression. If the coefficient of correlation between two variables X and Y be 0.83, what percentage of total variation remains unexplained by the regresion equation?

[I.C.W.A., Dec. '75]

39. Given: Unexplained variation = 19.70, and Explained variation = 19.22, determine the coefficient of correlation.

[I.C.W.A., Dec. '76]

40. In a contest, two judeges ranked eight candidates *A, B, C, D, E, F, G* and *H* in order of their preference, as shown in the following table. Find the rank correlation coefficient.

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
First Judge	5	2	8	1	4	6	3	7
Second Judge	4	5	7	3	2	8	1	6

[I.C.W.A., June '75]

41. Compute the correlation coefficient of the following ranks of a group of students in two examinations. What conclusion do you draw from the result?

Roll Nos.	1	2	3	4	5	6	7	8	9	10
Ranks in B.Com. Exam.	1	5	8	6	7	4	2	3	9	10
Ranks in M.Com. Exam.	2	1	5	7	6	3	4	8	10	9

[C.U., M.Com. '75]

42. Ten competitors in a musical contest were ranked by 3 judges *A*, *B*, *C* in the following order :

Ranks by A	1	6	5	10	3	2	4	9	7	8
Ranks by B	3	5	8	4	7	10	2	1	6	9
Ranks by C	6	4	9	8	1	2	3	10	5	7

Using Rank Correlation method, discuss which pair of judges has the nearest approach to common likings in music.

[I.C.W.A., Dec. '78]

43. Ten students obtained the following marks in Mathematics and Statistics Calculate the rank correlation coefficient.

Student (Roll No.)	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics	78	36	98	25	75	82	90	62	65	39
Marks in Statistics	84	51	91	60	68	62	86	58	53	47

44. Given the following coefficients : $r_{12} = 0.41$, $r_{13} = 0.71$, $r_{23} = 0.5$ Find $r_{12.3}$, $r_{13.2}$ and $R_{1.23}$, where the symbols have their usual significance.

[C.U., M.Com. '74]

45. In a three-variate multiple correlation analysis, the following results were found:

$$\bar{x}_1 = 60 \qquad x_2 = 70 \qquad \bar{x}_3 = 100$$
$$s_1 = 3 \qquad s_2 = 4 \qquad s_3 = 5$$
$$r_{12} = 0.7 \qquad r_{13} = 0.6 \qquad r_{23} = 0.4$$

the symbols having their usual singificance. Find the regression of x_1 on x_2 and x_3, and the multiple correlation coefficient $R_{1.23}$.

[B.U., M.A.(Econ) '68]

ANSWERS

1. (See Example 9.2)
2. + 0.47
3. – 0.37
4. – 0.92
5. + 0.996
6. 0.60
7. + 0.71
8. – 0.73
9. + 0.98
10. + 0.90, 0.072
11. + 0.995
12. + 0.94
13. – 0.072, 0.237
14. – 0.82
15. $-\sigma_1/\sigma_2$
16. (i) positive, (ii) positive, (iii) negative, (iv) positive, (v) negative
17. + 0.98; $x = 1.5y - 0.5$
18. $y = 0.33x + 38.75$; 81.65; 0.55
19. $y = 0.222x + 0.133$; $x = 1.11y + 0.756$; 0.50

20. 45.2 units per acre
21. $y = 0.8x + 12$; $x = 0.45y + 1$
22. $y = 0.65x + 21.82$; $x = 0.27y + 26.68$; 54
23. 71.75 inches; 111.60 lbs.
24. + 0.98; $y = 0.64x + 0.52$; 8.2
25. $X = 0.95Y - 6.4$; $Y = 0.95X + 7.25$
26. $x = 2.23y - 12.76$; $y = 0.39x + 7.22$; (i) 30, (ii) 19.
27. $y = 0.613x + 14.83$; $x = 1.360y - 5.2$
28. $y = 3.99 + 0.103x$; 8.11 tons.
29. $r = \sqrt{0.57 \times 1.12} = +0.80$
30. (i) 17.7, (ii) –0.55
31. (i) 3 and 6, (ii) –0.4, (iii) 5.2
32. 56.64, 73.57; + 0.85
33. $\bar{x} = 1$, $\bar{y} = -1$, $r = -0.75$
34. $\bar{x} = 1$, $\bar{y} = 2$, $\sigma_y = 2$, $r = -\sqrt{3/2}$
35. + 0.77
36. Regression equation of blood pressure (y) on age (x) is $y = 1.14x + 80.67$; 132
37. First equation regression line of y on x; $y_0 = 14.6$; 11.92
38. 31%
39. $|r| = 0.70$
40. 2/3
41. +0.64
42. Rank corr. coeffs. between A & B, A & C and B & C respectively are – 0.21, + 0.64, – 0.30. Since + 0.64 is the largest, Judges A and C hae the nearest approach to common likings in music.
43. +0.82
44. +0.09, +0.64, +0.71
45. $x_1 = 8.3 + 0.41x_2 + 0.23x_3$; $R_{1.23} = 0.78$

10 INTERPOLATION

10.1 INTRODUCTION

'*Interpolation*' has been defined as the 'art of reading between the lines of a table', and the term usually denotes the process of finding the intermediate value of a function from a set of given values of that function. For example, using the following given values of x and y

x:	1	3	5	7	9	11
y:	1	15	65	175	369	671

we may be required to find, i.e. *interpolate,* the value of y when $x = 4$. Since the mathematical relationship between x and y is not known, we have necessarily to base our calculations on the given data.

In *Simple Interpolation* (page 63) we use only two pairs of observations on either side of the unknown value, and assume a constant rate of change (i.e. linear relationship) in this interval.

x	y
3	15
$4 \rightarrow$	$\leftarrow ?$
5	65

Using (3.16.1),

$$\frac{4-3}{5-3} = \frac{y-15}{65-15}.$$ Solving, we get $y = 40$.

However, the mathematical relationship between the values of x and y given above is $y = \frac{1}{2}(x^3 + x)$, so that when $x = 4$, the value of y is 34. The discrepancy of the interpolated value $y = 40$ from the true value $y = 34$ is due to the fact that we have not utilised the whole of the given data, but only two pairs of observations. Applying formulae for interpolation, stated in the following pages, it is possible (see Example 10.12) in some cases, to obtain the value of y exactly, and if not to obtain a close approximation to the true value.

10.2 FINITE DIFFERENCES: Δ AND *E* OPERATORS

In problems of interpolation, the independent variable x is often known as '*argument*', and the dependent variable or the function $y = f(x)$ is known as '*entry*'. Let $x_0, x_1, x_2, \ldots, x_n$ denote a set of equidistant values of the argument, i.e.

$$x_1 - x_0 = x_2 - x_1 = \ldots = x_n - x_{n-1} = h$$

where h is a constant; and $y_0, y_1, y_2, \ldots, y_n$ denote the corresponding values of the entry. Differences of the successive values of y, viz.

$$(y_1 - y_0), (y_2 - y_1), (y_3 - y_2), \ldots, (y_n - y_{n-1})$$

are called *finite differences of the first order* (or simply *first differences*), and are denoted by

$$\Delta y_0, \Delta y_1, \Delta y_2, \ldots, \Delta y_{n-1}$$

respectively. The differences of the successive first order differences Δy, namely

$$(\Delta y_1 - \Delta y_0), (\Delta y_2 - \Delta y_1), \ldots, (\Delta y_{n-1} - \Delta y_{n-2})$$

are known as *finite differences of the second order* (or simply *second differences*) and are denoted by

$$\Delta^2 y_0, \Delta^2 y_1, \ldots, \Delta^2 y_{n-2}$$

respectively. Similarly, the *third differences* $\Delta^3 y$, the *fourth differences* $\Delta^4 y$, and differences of higher order may be defined.

Argument (x)	*Entry* (y)	*First Differences* (Δy)	*Second Differences* $(\Delta^2 y)$
x_0	y_0		
		$y_1 - y_0 = \Delta y_0$	
x_1	y_1		$\Delta y_1 - \Delta y_0 = \Delta^2 y_0$
		$y_2 - y_1 = \Delta y_1$	
x_2	y_2		$\Delta y_2 - \Delta y_1 = \Delta^2 y_1$
		$y_3 - y_2 = \Delta y_2$	
x_3	y_3		$\Delta y_3 - \Delta y_2 = \Delta^2 y_2$
		$y_4 - y_3 = \Delta y_3$	
x_4	y_4		

A table which shows the finite differences is known as *Difference Table*.

Table 10.1 Difference Table

Argument x	*Entry* y	*First Differences* Δy	*second Differences* $\Delta^2 y$	*Third Differences* $\Delta^2 y$	*Fourth Differences* $\Delta^4 y$
$x_0 = 1$	$y_0 = 1$				
		14			
$x_1 = 3$	$y_1 = 15$		36		
		50		24	
$x_2 = 5$	$y_2 = 65$		60		0
		110		24	
$x_3 = 7$	$y_3 = 175$		84		0
		194		24	
$x_4 = 9$	$y_4 = 369$		108		
		302			
$x_5 = 11$	$y_5 = 671$				

In symbols, the difference table may be shown as follows:

Table 10.2 Difference Table

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$
x_0	y_0					
		Δy_0				
x_1	y_1		$\Delta^2 y_0$			
		Δy_1		$\Delta^3 y_0$		
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$	
		Δy_2		$\Delta^3 y_1$		$\Delta^5 y_0$
x_3	y_3		$\Delta^2 y_2$		$\Delta^4 y_1$	
		Δy_3		$\Delta^3 y_3$		
x_4	y_4		$\Delta^2 y_3$			
		Δy_4				
x_5	y_5					

The arrangement of differences should be noted. Each difference is written between the two quantities which are used to calculate it.

The initial term y_0 of the entry is called the *leading term* and the initial terms in the difference columns, viz. Δy_0, $\Delta^2 y_0$, $\Delta^3 y_0$ etc. are called *leading differences.*

[**Note:** Δ (*called delta*) is not a quantity multiplied to y^2s, but an 'operator' like the root-sign $\sqrt{}$ or log. These operators separately have no meaning, but when operating on a number, e.g. $\sqrt{9}$ or log 7, have a real significance. Again, Δ^2 does not imply the square of Δ, but the repetition of operation by Δ twice].

Another symbolic operator E is defined as follows:

$$Ey_0 = y_1, \quad Ey_1 = y_2, \quad Ey_2 = y_3, \quad \ldots\ \ldots$$

Both the operators Δ and E can be applied repeatedly, the repeated operations being indicated by Δ^2, Δ^3, ... and E^2, E^3, etc. Thus

$$\begin{aligned}\Delta^2 y_0 &= \Delta(\Delta y_0) = \Delta y_1 - \Delta y_0 \\ &= (y_2 - y_1) - (y_1 - y_0) = y_2 - 2y_1 + y_0 \\ \Delta^3 y_0 &= \Delta(\Delta^2 y_0) = \Delta^2 y_1 - \Delta^2 y_0 \\ &= (y_3 - 2y_2 + y_1) - (y_2 - 2y_1 + y_0) \\ &= y_3 - 3y_2 + 3y_1 - y_0\end{aligned}$$

Similarly, $E^2 y_0 = E(Ey_0) = E(y_1) = y_2$

$E^3 y_0 = E(E^2 y_0) = E(y_2) = y_3$; etc.

From the definitions, we may in general write

$$\Delta y_r = y_{r+1} - y_r \qquad (10.2.1)$$

$$Ey_r = y_{r+1} \qquad (10.2.2)$$

These operators may thus be interpolated in the following manner:

(a) Δ when prefixed to y_r implies that y_r is to be subtracted from the next value of the entry y_{r+1}.

(b) E when prefixed to y_r denotes the next value of the entry y_{r+1}.

From (10.2.1) and (10.2.2) we find that

$$Ey_r = y_r + \Delta y_r$$

or,
$$Ey_r = (1 + \Delta)\, y_r \quad \text{(suppose)} \qquad (10.2.3)$$

Omitting y_r from both sides, we find that the operators E and Δ are connected by the symbolic relation

$$E = 1 + \Delta \qquad (10.2.4)$$

This does not mean that 1 when added to Δ gives E, but that the operation by E is equivalent to the operation by $(1 + \Delta)$. It may be shown that the above relation follows certain algebraic rules.

As shown earlier, we have

$$\begin{aligned} \Delta\ y_0 &= y_1 - y_0 \\ \Delta^2\ y_0 &= y_2 - 2y_1 + y_0 \\ \Delta^3\ y_0 &= y_3 - 3y_2 + 3y_1 - y_0 \\ \Delta^4\ y_0 &= y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 \end{aligned} \qquad (10.2.5)$$

Alternatively, we may write

$$\begin{aligned} \Delta\ y_0 &= Ey_0 - y_0 \\ \Delta^2\ y_0 &= E^2y_0 - 2Ey_0 + y_0 \\ \Delta^3\ y_0 &= E^3y_0 - 3E^2y_0 + 3Ey_0 - y_0 \\ \Delta^4\ y_0 &= E^4y_0 - 4E^3y_0 + 6E^2y_0 - 4Ey_0 + y_0 \end{aligned}$$

With the operators only (removing y_0's from both sides)

$$\begin{aligned} \Delta\ &= E - 1 \\ \Delta^2 &= E^2 - 2E + 1 = (E + 1)^2, \\ \Delta^3 &= E^3 - 3E^2 + 3E - 1 = (E - 1)^3 \\ \Delta^4 &= E^4 - 4E^3 + 6E^2 - 4E + 1 = (E - 1)^4 \end{aligned}$$

Thus, we have developed a convenient method of expressing the finite difference of any order in terms of the entries.

(see Example 10.1)

Example 10.1 *Express $\Delta^5 y_0$ in terms of $y_0, y_1, y_2, \ldots$*

Solution Using (10.2.4), we replace Δ by $E - 1$ and expand $(E - 1)^5$ by the binomial series (see pages 68-69).

$$\begin{aligned} \Delta^5 y_0 &= (E - 1)^5\, y_0 \\ &= (E^5 - 5E^4 + 10E^3 - 10E^2 + 5E - 1)\, y_0 \qquad \text{[Example 3.27 (iii)]} \\ &= E^5y_0 - 5E^4y_0 + 10E^3y_0 - 10E^2y_0 + 5Ey_0 - y_0 \\ &= y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0. \end{aligned}$$

Example 10.2 *If $y = c_0 + c_1x + \frac{c_2}{2!}\, x(x - 1)$ passes through the points $(0, y_0)$, $(1, y_1)$, $(2, y_2)$ calculate c_0, c_1, c_2 in terms of differences of y's.*

[I.C.W.A., July '68]

Solution The statement that the curve passes through the points $(0, y_0)$, $(1, y_1)$ and $(2, y_2)$ implies that when $x = 0, 1, 2$, the values of y are respectively y_0, y_1, y_2.

Putting $x = 0$,

$$y_0 = c_0 + c_1 \times 0 + \frac{c_2}{2!} \times 0(0 - 1)$$

i.e.,
$$y_0 = c_0 \qquad \ldots(i)$$

Solution Since only 4 values are available, we assume a 3rd degree polynomial for u_x so that the 4th and higher order differences are zero. In particular, $\Delta^8 u_0 = 0$; or, $(E-1)^8 u_0 = 0$. Expanding,

$$u_8 - 8u_7 + 28u_6 - 56u_5 + 70u_4 - 56u_3 + 28u_2 - 8u_1 + u_0 = 0$$

or, $(u_8 + u_0) - 8(u_7 + u_1) + 28(u_6 + u_2) - 56(u_5 + u_3) + 70u_4 = 0$

Substituting the values and solving, we get $u_4 = 12$.

10.4 NEWTON'S FORWARD INTERPOLATION FORMULA

Let $y_0, y_1, y_2, \ldots y_n$ be some tabulated values of a function $y = f(x)$ corresponding to the equidistant values $x = x_0, x_1, x_2, \ldots x_n$.

$$x_1 - x_0 = x_2 - x_1 = x_3 - x_2 = \ldots = x_n - x_{n-1} = h \text{ (say)}.$$

It is required to find the value of y corresponding to an intermediate value of x lying near the beginning of the tabulated values. This is given by *Newton's Forward Interpolation Formula*:

$$y = y_0 + u\Delta y_0 + \frac{u(u-1)}{1\times 2}\Delta^2 y_0 + \frac{u(u-1)(u-2)}{1\times 2\times 3}\Delta^3 y_0 + \ldots$$

$$\ldots + \frac{u(u-1)(u-2)\ldots(u-n+1)}{1\times 2\times 3\times \ldots \times n}\Delta^n y_0. \quad (10.4.1)$$

where $u = \dfrac{x - x_0}{h}$

Derivation of Newton's Forward Formula

All interpolation formulae are based on the following principle. Within the range of $n+1$ given values, the tabulated function $y = f(x)$ can be replaced by an n-th degree polynomial $\phi(x)$, which coincides with $f(x)$ when $x = x_0, x_1, x_2, \ldots, x_n$; i.e.

$$\phi(x_0) = y_0,\ \phi(x_1) = y_1,\ \phi(x_2) = y_2, \ldots, \phi(x_n) = y_n \quad \ldots(1)$$

Let us write the n-th degree polynomial $\phi(x)$ in the form

$$\phi(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + a_3(x - x_0)(x - x_1)(x - x_2) + \ldots + a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}). \quad \ldots(2)$$

The $n+1$ constants $a_0, a_1, a_2, \ldots, a_n$ are so determined as to satisfy the $n+1$ relations (1).

Putting $x = x_0$ in (2), we have

$$\phi(x_0) = a_0$$

because all other terms will contain $(x_0 - x_0)$ and therefore vanish.

or, $y_0 = a_0$; i.e. $a_0 = y_0$

Now putting $x = x_1$ in (2)

$$\phi(x_1) = a_0 + a_1(x_1 - x_0) \text{ because the other terms vanish.}$$

or, $y_1 = y_0 + a_1 h$, since $x_1 - x_0 = h$ (given)

or, $a_1 h = y_1 - y_0$

$$\therefore \quad a_1 = \frac{y_1 - y_0}{h} = \frac{\Delta y_0}{h}$$

Again, putting $x = x_2$ in (2),

$$\phi(x_2) = a_0 + a_1(x_3 - x_0) + a_2(x_2 - x_0)(x_3 - x_1)$$

or, $$y_2 = y_0 + \frac{y_1 - y_0}{h}(2h) + a_2(2h^2)(h)$$

or, $$y_2 = y_0 + 2(y_1 - y_0) + a_2(2h^2)$$

Solving, we get

$$a_2 = \frac{y_2 - 2y_1 + y_0}{2h^2} = \frac{\Delta^2 y_0}{2h^2}$$

Now putting $x = x_3$ in (2),

$$\phi(x_3) = a_0 + a_1(x_3 - x_0) + a_2(x_3 - x_0)(x_3 - x_1) + a_3(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)$$

or, $$y_3 = y_0 + \frac{y_1 - y_0}{h}(3h) + \frac{y_2 - 2y_1 + y_0}{2h^2}(3h)(2h) + a_3(3h)(2h)(h)$$

or, $$y_3 = y_0 + 3(y_1 - y_0) + 3(y_2 - 2y_1 + y_0) + a_3(1 \times 2 \times 3)\, h^3$$

Solving, $$a_3 = \frac{y_3 - 3y_2 + 3y_1 - y_0}{(1 \times 2 \times 3)h^3} = \frac{\Delta^3 y_0}{(1 \times 2 \times 3)h^3}$$

Similarly, putting $x = x_4, \ldots, x_n$ successively in (2), we find that

$$a_4 = \frac{\Delta^4 y_0}{(1 \times 2 \times 3 \times 4)h^4}, \ldots a_n = \frac{\Delta^n y_0}{(1 \times 2 \times \ldots \times n)\, h^n}$$

Substituting the values of $a_0, a_1, a_2, \ldots, a_n$ in (2),

$$y = \phi(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{(1 \times 2)h^2}(x - x_0)(x - x_1)$$

$$+ \frac{\Delta^3 y_0}{(1 \times 2 \times 3)h^3}(x - x_0)(x - x_1)(x - x_2) + \ldots$$

$$+ \frac{\Delta^n y_0}{(1 \times 2 \times \ldots \times n)h^n}(x - x_0)(x - x_1) \ldots (x - x_{n-1}) \qquad \ldots(3)$$

If we write $\frac{x - x_0}{h} = u$, i.e. $x = x_0 + hu$, then

$$x - x_0 = hu,$$
$$x - x_1 = (x_0 + hu) - (x_0 + h) = h(u - 1),$$
$$x - x_2 = (x_0 + hu) - (x_0 + 2h) = h(u - 2),$$
$$\ldots \qquad \ldots \qquad \ldots$$
$$x - x_{n-1} = (x_0 + hu) - \{x_0 + (n - 1)h\} = h(u - n + 1)$$

Substituting these values in (3), we have

$$y = y_0 + \Delta y_0 . u + \frac{\Delta^2 y_0}{1 \times 2}u(u - 1) + \frac{\Delta^3 y_0}{1 \times 2 \times 3}u(u - 1)(u - 2)$$

$$+ \ldots \ldots + \frac{\Delta^n y_0}{1 \times 2 \times \ldots \times n}u(u - 1)(u - 2) \ldots (u - n + 1).$$

Rearranging the terms on the right, we get Newton's forward interpolation formula (10.4.1).

Uses—Newton's forward interpolation formula can be used only when the values of the argument are equidistant. Since the formula contains only the leading term y_0 and the leading differences Δy_0, $\Delta^2 y_0$, etc., it is most suitable for interpolation near the beginning of the tabulated values. It may also be used to find the value of the function corresponding to values slightly before the first tabulated value of argument (The process of finding the value of the function outside the range of tabulated values is known as *Extrapolation*).

Example 10.11 *From the following table, find by interpolation the value of $f(x)$ for $x = 24$:*

x	20	25	30	35	40
$f(x)$	30.5	34.5	40	47.75	59.25

Solution Since the given values of x are equidistant and the value $x = 24$ lies near the beginning of these values, we use Newton's forward interpolation formula (10.4.1).

Table 10.3 Difference Table

x	$y = f(x)$	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
20	30.5				
		4.0			
25	34.5		1.5		
		5.5		0.75	
30	40		2.25		0.75
		7.75		1.50	
35	47.75		3.75		
		11.50			
40	59.25				

Hence, $u = \dfrac{x - x_0}{h} = \dfrac{24 - 20}{5} = 0.8$. Applying (10.4.1),

$$y = 30.5 + 0.8(4.0) + \frac{0.8(0.8 - 1)}{2}(1.5) + \frac{0.8(0.8 - 1)(0.8 - 2)}{6}(0.75) + \frac{0.8(0.8 - 1)(0.8 - 2)(0.8) - 3)}{24}(0.75)$$

$$= 30.5 + 3.2 - 0.12 + 0.024 - 0.0132$$

$$= 33.5908 = 33.59$$

Ans. 33.59

Example 10.12 *Given the following table, find $(3.24)^3$:*

x	31	32	33	34	35
x^3	29791	32768	35937	39304	42875

Solution Since the successive values of x have a common difference, we shall use Newton's forward interpolation formula and find $y = (32.4)^3$, from which the value of $(3.24)^3$ can be easily obtained by shifting the decimal point three places to the left.

Table 10.4 Difference Table

x	$y = x^3$	Δy	$\Delta^2 y$	$\Delta^3 y$
31	29,791			
		2,977		
32	32,768		192	
		3,169		6
33	35,937		198	
		3,367		6
34	39,304		204	
		3,571		
35	42,875			

Here $u = (x - x_0)/h = (32.4 - 31)/1 = 1.4$. Using (10.4.1),

$$y = (32.4)^3 = 29791 + 1.4(2977) + \frac{1.4(1.4-1)}{2}(192) + \frac{1.4(1.4-1)(1.4-2)}{6} \quad (6)$$

$$= 29791 + 4167.8 + 53.76 - 0.336 = 34012.224$$

Hence, $(3.24)^3 = 34.012,224$. This result is exact.

10.5 NEWTON'S BACKWARD INTERPOLATION FORMULA

As in Section 10.4, let $x_0, x_1, x_2, ..., x_n$ be some equidistant values of the argument and $y_0, y_1, y_2, ..., y_n$ the corresponding entries. It is required to find the value of y corresponding to a specified value of x lying near the end of the tabulated values. This is given by *Newton's Background Interpolation Formula*:

$$y = y_n + v\Delta y_{n-1} + \frac{v(v+1)}{1\times 2}\Delta^2 y_{n-2} + \frac{v(v+1)(v+2)}{1\times 2\times 3}\Delta^3 y_{n-3} + ...$$

$$... + \frac{v(v+1)(v+2)...(v+n-1)}{1\times 2\times 3\times ...\times n}\Delta^n y_0. \quad (10.5.1)$$

where $v = \dfrac{x - x_n}{h}$.

Derivation of Newton's backward formula

As before, within the range of the given values, the tabulated function $y = f(x)$ is replaced by an n-th degree polynomial $\phi(x)$, which coincides with $f(x)$ when $x = x_0, x_1, ..., x_n$; i.e.

$$\phi(x_0) = y_0, \phi(x_1) = y_1, \phi(x_2) = y_2, ..., \phi(x_n) = y_n \quad ...(1)$$

We write the n-th degree polynomial $\phi(x)$ in the form

$$\phi(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + a_3(x - x_n)(x - x_{n-1})(x - x_{n-2})$$
$$+ ... + a_n(x - x_n)(x - x_{n-1}) ... (x - x_1). \quad ...(2)$$

The $n + 1$ constants $a_0, a_1, a_2, ... a_n$ are so determined as to satisty the $n + 1$ relations (1).

Putting $x = x_n$ in (2), we have

$$\phi(x_n) = a_0$$

because all other terms contain the factor $(x_n - x_n)$ and therefore vanish.

or, $y_n = a_0$; i.e. $a_0 = y_n$

Now putting $x = x_{n-1}$ in (2),

$$\phi(x_{n-1}) = a_0 + a_1 (x_{n-1} - x_n), \text{ because the other terms vanish;}$$

or, $$y_{n-1} = y_n + a_1 (-h), \quad \text{since} \quad x_n - x_{n-1} = h$$

or, $$a_1 h = y_n - y_{n-1} = \Delta y_{n-1}$$

$$\therefore \quad a_1 = \frac{y_n - y_{n-1}}{h} = \frac{\Delta y_{n-1}}{h}$$

Again, putting $x = x_{n-2}$ in (2),

$$\phi(x_{n-2}) = a_0 + a_1(x_{n-2} - x_n) + a_2(x_{n-2} - x_n)(x_{n-2} - x_{n-1})$$

or, $$y_{n-2} = y_n + \frac{y_n - y_{n-1}}{h}(-2h) + a_2(-2h)(-h)$$

or, $$y_{n-2} = y_n - 2(y_n - y_{n-1}) + a_2(2h^2)$$

Solving, $$a_2 = \frac{y_n - 2y_{n-1} + y_{n-2}}{2h^2} = \frac{\Delta^2 y_{n-2}}{2h^2}$$

Similarly, putting $x = x_{n-3}, x_{n-4}, \ldots, x_0$ successively in (2), we get

$$a_3 = \frac{\Delta^3 y_{n-3}}{(1 \times 2 \times 3)h^3},$$

$$a_4 = \frac{\Delta^4 y_{n-4}}{(1 \times 2 \times 3 \times 4)h^4}, \ldots a_n = \frac{\Delta^n y_0}{(1 \times 2 \times \ldots \times n)h^n}$$

Substituting the values of $a_0, a_1, a_2, \ldots, a_n$ in (2),

$$y = y_n + \frac{\Delta y_{n-1}}{h}(x - x_n) + \frac{\Delta^2 y_{n-2}}{(1 \times 2)h^2}(x - x_n)(x - x_{n-1})$$

$$+ \frac{\Delta^3 y_{n-3}}{(1 \times 2 \times 3)h^3}(x - x_n)(x - x_{n-1})(x - x_{n-2}) + \ldots$$

$$\ldots + \frac{\Delta^n y_0}{(1 \times 2 \times \ldots \times n)h^n}(x - x_n)(x - x_{n-1}) \ldots (x - x_1) \qquad \ldots(3)$$

If we write $\frac{x - x_n}{h} = v$, i.e. $x = x_n + hv$, then

$$x - x_n = hv$$
$$x - x_{n-1} = (x_n + hv) - (x_n - h) = h(v + 1)$$
$$x - x_{n-2} = (x_n + hv) - (x_n - 2h) = h(v + 2)$$
$$\ldots \qquad \ldots$$
$$x - x_1 = (x_n + hv) - \{x_n - (n-1)h\} = h(v + n - 1)$$

When these values are substituted in (3), we have

$$y = y_n + \Delta y_{n-1} v + \frac{\Delta^2 y_{n-2}}{1 \times 2} v(v+1) + \frac{\Delta^2 y_{n-3}}{1 \times 2 \times 3} v(v+1)(v+2)$$

$$+ \ldots \ldots + \frac{\Delta^n y_0}{1 \times 2 \times \ldots \times n} v(v+1)(v+2) \ldots (v+n-1).$$

Rearranging the terms on the right, we get Newton's backward interpolation formula (10.5.1).

Uses—Newton's backward interpolation formula can be used only when the values of the argument are equidistant. Since the formula contains only the terms at the bottom of the difference table (Table 10.2), it is most suitable for interpolation near the end of the tabulated values. It may also be used to extrapolate the value of the function corresponding to an argument slightly beyound the last tabulated valued.

[**Note:** (i) Usually, each term in formulae (10.4.1) and (10.5.1) is less than the preceding. Hence, in numerical problems it is not necessary to calculate all the terms given in the formula. Only the first few terms are sufficient to obtain the required estimate of y to the same degree of accuracy as the given entries.

(ii) If all the terms in formulae (10.4.1) and (10.5.1) are used, Newton's forward formula and backward formula would give exactly the same result.]

Example 10.13 *Given the following table, estimate y when $x = 0.35$, by using Newton's backward interpolation formula.*

x:	*0*	*0.1*	*0.2*	*0.3*	*0.4*
y:	*1*	*1.095*	*1.179*	*1.251*	*1.310*

(Give answer correct to 3 decimal places). [I.C.W.A., June '78]

Solution (For simplicity, we shall ignore the decimal point while computing the differences).

Table 10.5 Difference Table

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0	1				
		95			
0.1	1.095		− 11		
		84		− 1	
0.2	1.179		− 12		0
		72		− 1	
0.3	1.251		− 13		
		59			
0.4	1.310				

Here $v = (x - x_n)/h = (0.35 - 0.4)/0.1 = -0.5$

Using 3 decimals,

$$y = 1310 + (-0.5) \times 59 + \frac{(-0.5)(-0.5+1)}{2} \times (-13) + \frac{(-0.5)(-0.5+1)(-0.5+2)}{6} \times (-1)$$

$$= 1310 - 29.5 + 1.6 + 0.1 = 1282.2$$

$$= 1282$$

Adjusting the decimal point, the required value is 1.282.

EXERCISES

1. The following data show the monthly average number of deaths under one year in a certain large city. Find the missing term:

Year	1960	1961	1962	1963	1964
Number of Deaths (monthly average)	940	?	907	843	798

[I.C.W.A., Jan. '72]

2. The following gives the amount y of cement in thousants of tons manufactured in India in the year x. Find the missing term.

x	1946	1948	1950	1952	1954	1956
y	39	85	?	151	264	388

[I.C.W.A., July '69]

3. The growth of population in India, according to the decemial census, is shown below:

Year	1901	1911	1921	1931	1941	1951
Population (lakh)	2384	2552	2514	2791	...	3613

The census figure for 1941 is not given here. Give an estimate of the actual population for 1941.

[C.U., B.Sc. '71]

4. Below are given the values of a function U_x for certain values of x:

x	0	1	2	3	4
U_x	1	0	5	22	57

Construct the table of differences. What does the table suggest? Use this table to find U_5.

[I.C.W.A., Dec.'76]

5. Form a difference table and find the values of y_3 and y_9 from the following: $y_4 = 135$, $y_5 = 122$, $y_6 = 1015$, $y_7 = 2016$, $y_8 = 3591$.
6. (a) Given $y_0 = 8$, $y_1 = 6$, $y_2 = 4$, $y_4 = 24$, find the value of y_3.
   (b) Given $y_2 = 5$, $y_5 = 122$, $y_6 = 193$, find the values of y_3 and y_4.
   (c) Given $u_5 = 3$, $u_{10} = 0$, $u_{20} = -24$, $u_{25} = -33$, estimate u_{15} under suitable assumption.
7. Find the missing term:

x	0	1	2	3	4
$f(x)$	1	3	9	?	81

8. Find y for $x = 2$, from the following table:

x	...	0	1	3	4	5
y	...	39	85	151	264	388

[I.C.W.A., Jan. '69]

9. Find $f(5)$ from the following data:
   $f(3) = 4$, $f(4) = 13$, $f(6) = 43$

10. Find the polynomial function $f(x)$ from the following values:
$f(3) = -1, \quad f(4) = 5, \quad f(5) = 15$

11. Given the following table, find the function $f(x)$, assuming it to be a polynomial of the 3rd degree in x:

x	0	1	2	3
$f(x)$	1	2	11	34

[I.C.W.A., June '75]

12. u_x is a polynomial in x. Given the following table, find u_x.

x	0	1	2	5
u_x	2	3	12	147

13. Below are given the values of a function $f(x)$ for certain values of x. Find $f(2)$, stating your assumption.

x	0	1	3	4
$f(x)$	5	6	50	105

[I.C.W.A., Dec. '75]

14. Find $f(x)$, given that
$$f(0) = -3, \quad f(1) = 6, \quad f(2) = 8, \quad f(3) = 12.$$
(State your assumptions, if any). Hence find $f(6)$. *[I.C.W.A., June '76]*

15. Using any algebraic method, find the value of y when $x = 6$,

x	3	7	9	10
y	16.8	12.0	7.2	6.3

[I.C.W.A., Dec. '75-old]

16. For a certain polynomial function y_x it is known that
$$y_1 = -1, \quad y_2 + y_3 = -1, \quad y_4 + y_5 + y_6 = 61.$$
Find y_x, and hence the value of y_3.

17. Given $u_0 + u_6 = -107$, $u_1 + u_5 = -36$, $u_2 + u_4 = -3$, find the value of u_3.

18. From the following table estimate by interpolation the number of units of a commodity supplied when the price is Rs 4:

Price in Rs	1	3	5	7	9
No. of units Supplied	256	625	935	1201	1433

19. The following table gives the expectation of life e_x° at age x. Calculate the expectation of life at age 12 by using Newton's forward interpolation formula.

x	10	15	20	25	30	35
e_x°	35.4	32.2	29.1	26.0	23.1	20.4

[I.C.W.A., Dec. '77]

20. The following shows the values of a function $y = f(x)$ for a number of values of x:

x:	0.5	0.6	0.7	0.8	0.9
y	0.35207	0.33322	0.31225	0.28969	0.26609

Obtain the value of y when $x = 0.58$, using a suitable interpolation formula.
[C.U., B.A. (Econ) '76]

21. The table below gives the average number of years of life remaining to persons who survive to exact age x, for male African population of Belgian Congo:

x	0	5	10	15	20
$^{\circ}e_x$	37.64	44.04	41.40	37.78	34.41

Obtain $^{\circ}e_2$ approximately. *[I.C.W.A., Dec. '73]*

22. State Newton's Forward Interpolation formula, and use it to find $\sqrt{5.5}$, given that $\sqrt{5} = 2.236, \sqrt{6} = 2.449, \sqrt{7} = 2.646$ and $\sqrt{8} = 2.828$. *[I.C.W.A., June '74]*

23. The following table shows the number of earners earning incomes exceeding different amounts during a certain period:

Income (Rs)	50,000	75,000	100,00	125,000	150,000
No. of Earners	412	304	225	147	88

Find the number of earners earning more than Rs 60,000 by linear interpolation and also by using Newton's forward formula. *[C.U., B.A. (Econ) '77]*

24. Using suitable interpolation formulae, calculate the values of y, when (i) $x = 10$, and (ii) $x = 25$.

x	7	11	15	19	23	27
y	20,256	20,625	21,296,	22,407,	24,098	26,511

25. Find the value of sin 48° from the following table:

x (degrees)	30	35	40	45	50
sin x	.5000	.5736	.6428	.7071	.7660

26. Using Newton's interpolation formula, find the number of factories earning less than Rs 65,000 as profits, from the following data:

Profits (Rs '000)	30–40	40–50	50–60	60–70	70–80
No. of Factories	34	43	56	39	29

[I.C.W.A., Dec. '75]

27. Apply the appropriate interpolation formula to find log 3.146, given
log 3.141 = 0.497, 0679 log 3.144 = 0.497, 4825
log 3.142 = 0.497, 2062 log 3.145 = 0.497, 62.05
log 3.143 = 0.497, 3444
(Find correct upto 7 decimal places). *[I.C.W.A., Dec. '78]*

28. Use Stirling's interpolation formula to find the value of the probability integral (P) when $X = 1.52$:

X...	1.3	1.4	1.5	1.6	1.7
P = prob. integral	.90320	.91924	.93319	.94520	.95543

29. Given the following cube-roots, find by Bessel's interpolation formula, the cube root of 102.5:

Number	101	102	103	104
Cube-root	4.657,0095	4.672,3287	4.687,5481	4.702,6694

[B.U., B.A. (Econ) '65]

30. Find by using Bessel's interpolation formula, the expectation of life at age 22 from the following data:

Age (x)	10	15	20	25	30	35
Exp. of Life (y)	35.4	32.2	29.1	26.0	23.1	20.4

31. The following table gives the normal weight of a baby during the first six months of life:

Age in Months	0	2	3	5	6
Weight in lbs.	5	7	8	10	12

Estimate the weight of a baby at the age of 4 months.

[I.C.W.A., Jan. '70]

32. State Lagrange's interpolation formula. Use it to find $f(x)$ when $x = 0$, given

x	–1	–2	2	4
$f(x)$	–1	–9	11	69

[I.C.W.A., Dec. '74]

33. State Lagrange's interpolation formula. Use it to find the value of U_4 of a function U_x, given that

$U_1 = 10, \quad U_2 = 15, \quad U_5 = 42.$ *[I.C.W.A., Dec. '76]*

34. Using Lagrange's formula or otherwise, obtain the value of log 95 approximately from the following table:

x	95	97	98	99
log x	1.977,7236	1.986,7717	1.991,2261	1.995,6352

[C.U., B.Com. (Hons) '66, '69]

35. Given $\log_{10} 654 = 2.8156$, $\log_{10} 658 = 2.8182$
$\log_{10} 659 = 2.8189$, $\log_{10} 661 = 2.8202$,
find by Lagrange's interpolation formula $\log_{10} 656$ (Retain 4 decimal places in your answer). *[I.C.W.A., June '78]*

36. Find the value of x for which $y = 40$:

x	10	12	15	20
y	25	32	35	45

ANSWERS

1. 952
2. 96.4
3. 3322 lakh
4. 3rd differences constant; hence, U_x may be given by a 3rd degree polynomial. $U_5 = 116$.
5. 16; 5920
6. (a) 8; (b) 28, 67; (c) –11
7. 31
8. 96.4
9. 26
10. $2x^2 - 8x + 5$
11. $x^3 + x^2 - x + 1$
12. $x^3 + x^2 - x + 2$
13. 19; assuming $f(x)$ to be a 3rd degree polynomial.

14. 126 : assuming $f(x)$ to be a 3rd degree polynomial.

$$f(x) = \frac{1}{2}(3x^3 - 16x^2 + 31x - 6)$$

15. 14.7; (use *Second method*, Example 10.8).
16. $2x^2 - 7x + 4$; 1
17. 3.2
18. 786
19. 34.1
20. 0.33718
21. 42.08
22. 2.345
23. 369; 362
24. 20,510; 25,205
25. 0.7431
26. 156 (use Cumulative frequency distribution)
27. 0.497, 7584
28. 0.93574
29. 4.679, 9508
30. 27.85
31. $8\frac{8}{9}$ lbs.
32. 1
33. 31
34. 1.9822, 7115 (see Examples 10.6 and 10.15)
35. 2.8168
36. 19.56

13 SAMPLING THEORY

13.1 MEANING AND OBJECTS OF 'SAMPLING'

'*Sampling*' denotes the selection of a part of the aggregate statistical material with a view to obtain information about the whole. This aggregate or totality of statistical information on a particular character of all the members covered by an investigation, is called *Population* or *Universe.* The selected part, which is used to ascertain the characteristics of population is called *sample* (Section 1.4, page 6). While taking a sample, the population is assumed to be composed of individual *units* or *members,* some of which are included in the sample. The total number of members of the population and the number included in the sample, called '*population size*' and '*sample size*' respectively.

A statistical population is '*finite*' or '*infinite*' according to its size. When the number of members of the population can be expressed as a definite quantity, the population is said to be 'finite'. Otherwise the population is 'infinite'. In particular, if a sample is known to have been drawn from a continuous probability distribution, then the population is infinite. A complete list of all the units in finite population, properly numbers for identification, is called a *Sampling Frame,* e.g., a list of all households in a given region as used in population census.

Again, the population may be '*existent*' or '*hypothetical*'. The population of incomes of all persons in a country is an example of existent population; because the members of the population really exist, though they may not be actually known because of practical difficulties. The population of points obtained in all possible throws of a die is an example of 'hypothetical' population.

If a sample of an adequate size is properly chosen and analysed, it is most likely to reveal the characteristics of the whole population, and the results obtained therefore can be fairly relied upon as if they were based on all members of the population. The possibility of reaching valid conclusions concerning a population by means of a properly chosen sample is based on two important laws: (i) Law of Statistical Regularity, and (ii) Law of Inertia of Large Numbers. The former states that a sample of reasonably large size when selected at random, is almost sure to represent the characteristics of the population (page, 11). The second law states that samples of large size show a high degree of stability, i.e., the results obtained theorem are expected to be very close to the population characteristics.

Some practical examples of sampling in Commerce and Industry are as follows:

(1) Sampling methods are used in Market Research for assessing customer behaviour, especially during launching of new products in the market, and also to gauge their opinion about the existing products.

(2) In industry sampling is done for Statistical Quality Control (Chapter 19). During manufacture, a few consecutive items are picked out from the production line at regular intervals of time, (say, half an hour) and these items are thoroughly tested. If the results are beyond certain permissible limits, the production process is shut down to find out the causes of variation.

(3) Sampling is also used by Stores people for incoming lots from suppliers. These are subjected to 'sampling inspection' lot by lot and the decision about the acceptance or otherwise of a lot is decided by examining only samples.

(4) Auditors use test check and test audit to cover only a small fraction of all entries in the books of accounts in order to verify the accuracy of all the entries, where either full verification is not possible or considered unnecessary.

There are 2 methods of collection of statistical data—(i) *Census method* or *Complete Enumeration,* and (ii) *Sample Survey method.* 'Census' denotes collection of data from each and every unit of the statistical population; while 'sample survey' refers to the study of population based only on a sample (Section 1.5).

Example 13.1 *Discuss the advantages of sampling method over census method of collection of statistical information.*

[C.U., M. Com, '73, '78, '80; B.A. (Econ) '75; I.C.W.A. June '74; W.B.H.S. '78, '82]

Solution (See Example 1.9)

Example 13.2 *What are the main objects of sampling?*

[I.C.W.A., June '75]

Solution The main objects of sampling are:

(i) to obtain the maximum information about the population with the minimum effort; and

(ii) to state the limits of accuracy of estimates based on samples.

In the study of a population by means of sample observations, we are generally interested in one or more variables. The first requirement towards the analysis of a sample is to form an idea of the probability distribution of the variable in the population. This probability distribution involves some constants (known as *parameters*), and consequently our next task is to form estimates of these unknown parameters on the basis of available data. For example if the probability distribution is assumed to be normal, there are two parameters μ and σ to be estimated. Usually there are several methods of estimating the same parameter, and naturally the question arises as to which among these estimates will be the most suitable under given circumstances. The study of various criteria for food estimators and the methods of estimation are treated in what is known as *Theory of Estimation.*

Again, on the basis of samples we may be interested to known whether a certain population parameter has a specified value, or to compare the parameters of two populations on the basis of two samples drawn from these populations. Such problems are treated in what is known as *Theory of Testing Statistical Hypothesis.*

The second object of sampling is to determine how precise the estimate would be, i.e., to state the degree of confidence that we may put on the estimates obtained. Since the sample is only a part of the population and is obtained by chance, any inference regarding population

cannot be hundred per cent correct. Therefore a measure of precision of the estimate is given in terms of probability. This is done by using the probability distribution called *Sampling Distribution* of the estimate.

Example 13.3 *"Sampling is a necessity under certain conditions"—Explain this with illustrative examples.*

Solution For the collection of statistical data, sample survey method is usually preferred to census, because the former takes less time, money and labour. But many situations arise in which sampling is not an alternative to census, but is a must; because census is either impossible or useless.

Illustrations:

(i) In order to find the acreage under rice in India by the census method, each and every plot of land on which rice is grown within the geographical limits of Indian Union has to be visited and area under rise cultivation has to be measured. The method is quite impracticable, because of the enormous task of visiting the plots and the consequential time and money involved. Even if such a step is taken, it will take years before the data are completely available and the whole purpose of enquiry may be vitiated. However, estimates a of crop area are now obtained by the methods of sample survey within a short period of time.

(ii) Again, the rice merchant cannot afford to examine every single grain of rice he purchases. He has to depend on only a sample based on which he forms an idea about the quality of rice in the whole consignment. Thus when the population size is extremely large, e.g., in counting the number of 'sal' trees in a forest area sampling the only practical method.

(iii) A municipal corporation, which purchases thousands of electric bulbs for street lighting, in an attempt to know the average life of a lot supplied by a company, cannot go on testing each and every bulb until all of them burn out; because in that case nothing will be left for use. Instead, by testing only a few bulbs from the whole lot, an estimate of the average length of life is known. Similarly, in other destructive tests, e.g. measuring the resistance of glass sheets, or strength of bullets, there is no alternative but to resort to sampling.

(iv) When a new mine is discovered and it is necessary to know the quality of its contents, only a few ounces of ore, drilled out from here and there, are sufficient for chemical analysis. It is impossible and unnecessary to dig out the whole mine and then examine the quality.

(v) When the population is hypothetical, e.g. in order to test whether a coin is biased or not, it is tossed a number of times and decisions have to be arrived at on the basis of the results of the sample. Here sampling is the only method of studying the characteristics of the population.

13.2 SAMPLING ERROR AND BIAS

Example 13.4 *Explain what is meant by 'random error'.*

[C.U., M.Com., '73; I.C.W.A. June '74]

Solution Sampling means the selection of a part of the aggregate statistical material, called *population*, with a view to obtaining information about the whole aggregate.

However, the composition of sample is a matter of chance and hence even if a proper process of selection is employed, the results obtained from sample cannot be expected to represent the characteristics of population in toto. Some error, i.e., discrepancy, is inevitable. This is known as 'sampling error'. When random sampling is used for obtaining the sample, it is called *Random Sampling Error.* The magnitude of such errors depends on the size of the sample and the variability of the population. Random sampling error depends on chance, and if the proper process of selection is followed, the possible limits of the magnitude of such error can be theoretically calculated.

Example 13.5 *What is 'bias' and how does it arise in sampling? Discuss the various methods of reducing bias.* [C.A. May '78]

Solution In sampling, a properly selected part only of the universe is analysed and the results are made applicable to the whole. It cannot be expected that these results will agree in full with those of the universe, e.g. the mean of sample may differ from the mean of universe. Since a part can never represent the whole in toto, a certain amount of discrepency is involved in the sampling process. These discrepancies or errors are of two types—(i) *Sampling Error,* and (ii) *Non-sampling Error* or *Bias.*

Sampling error is inherent in the method of sampling and arises because the information collected relates only to a part. In actual practice, certain inaccuracies arise out of complex reaction between enumerators and informants, or in processing the collected data, which affect the conclusions derived therefrom. This kind of error is known as 'bias'. Bias arises solely due to human factor and is difficult to detect. It occurs even in census, where sampling error is completely absent. Bias may arise in several ways:

1. Bias may arise in selecting the sample itself. If the proper process of selection is not strictly adhered to, the investigator's desire or motive to obtain a certain result from the sample survey may influence his selection of the sample consciously or unconsciously. This type of bias is father serious in as much as its existence may not be immediately apparent. Such bias can be avoided only by following the scientific process of selection. For example, white drawing a random sample, the universally accepted random numbers should be used, instead of resorting to haphazard selection by the investigator.
2. Bias may arise because of substitution of a selected unit by another. For example, in a house-to-house survey, the next house is interviewed if no reply is forthcoming from the selected house. Bias introduced in such cases can only be avoided, if the selected sample is strictly adhered to.
3. Bias may arise from incomplete investigation, i.e. because of not investigating the whole of the selected sample. In a house-to-house survey, even if there be no substitution of the selected house by the next house, bias may arise if the selected house is not visited a second time, or if necessary, several times till the desired information is obtained. This type of bias arises particularly in postal questionnaires. Such bias can only be eliminated if the entire selected sample is investigated.
4. Bias may arise in the process of estimation of parameters. For instance, forecasts of crop-yield (say, rice production) in a State are computed on the basis of estimates of average yields per acre obtained for several regions and sub-regions. If the simple average of the recorded yields are used, instead of the weighted average, the ultimate result will be biased.

4504	9523	3282	3756	0653
1014	7894	9307	0458	9983
1575	2458	9200	8566	7302

Of course, any haphazard arrangement of the digits 0 to 9 will not give random numbers. The random numbers have been prepared by special devices and tested for randomness, Some of the famous series of random number are those of Tippett (41,600 digits), Kendall and Babington Smith (100,000 digits). Fisher and Yates (15,000 digits), Rand Corporation (1,000,000 digits). Random numbers are used for drawing random samples.

Example 13.7 *The following table gives the grades of 100 students in mathematics. Draw a random sample of size ten from the group of students and estamte the mean grade from the sample:*

75, 58, 80, 86, 76, 65, 75, 76, 72, 86, 83, 82, 80, 88, 68,
66, 63, 60, 69, 80, 66, 87, 73, 58, 76, 74, 85, 96, 60, 72,
86, 79, 95, 84, 41, 76, 78, 74, 74, 56, 50, 80, 66, 96, 80,
68, 79, 73, 72, 73, 87, 77, 60, 87, 40, 82, 77, 87, 76, 82,
66, 81, 84, 72, 63, 59, 76, 52, 57, 78, 79, 92, 80, 65, 90,
75, 74, 98, 64, 45, 68, 57, 79, 79, 83, 35, 76, 88, 67, 75,
60, 52, 63, 80, 94, 34, 78, 64, 58, 56.

(Tables of Random Sampling Numbers to be supplied.)

Solution Here, the population size $N = 100$. Each of the population members is, therefore, allotted a two-digited identity number 01, 02, ..., 99 and 00 (for the 100th member), as shown below:

Table 13.1 Grades in Mathematics and Their Identity Numbers

Identity No.	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Grade	75	58	80	86	76	65	75	76	72	86	83	82	80	88	68	66	63	60	69	80
Identity No.	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Grade	66	87	73	58	76	74	85	96	60	72	86	79	95	84	41	76	78	74	74	56
Identity No.	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Grade	50	80	66	96	80	68	79	73	72	73	87	77	60	87	40	82	77	87	76	82
Identity No.	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
Grade	66	81	84	72	63	59	76	52	57	78	79	92	80	65	90	75	74	98	64	45
Identity No.	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00
Grade	68	57	79	79	83	35	76	88	67	75	60	52	63	80	94	34	78	64	58	56

We now take down random numbers in groups of two, starting from the digit 5 in the 2nd row and 15th column (this was done without looking at the table) and moving horizontally, using the random number table at page 511. The numbers obtained are

58, 99, 83, 15, 75, 24, 58, (rejected), 92, 00, 85, 66

The 7th number 58 is rejected, because it has already appeared once. The ten random numbers remaining and the grades having these as identity numbers (Table 13.1) are shown below:

Random Numbers	58	99	83	15	75	24	92	00	85	66
Grades	87	58	79	68		38	52	56	83	59

The random sample of size ten is, therefore, composed of grades

87, 58 79, 68, 90, 58, 52, 83, 59

The mean of the sample, which gives an estimate of the population mean, is

$$\bar{x} = (87 + 58 + \ldots + 83 + 59) \div 10 = \frac{690}{10} = 69$$

Incidentally, the population mean of the 100 grades is 72.66.

Example 13.8 *Draw a random sample of size 10 (without replacement) from the following data, stating clearly the procedure followed by you:*

45	24	43	17	5	28	27	21	11	46	
33	26	24	14	34	21	25	48	35	38	
26	27	35	8	30	26	30	28	21	27	
20	13	23	36	38	20	25	31	24	18	12

You may use the random sampling numbers given below:

5967	8941	7989	3335	7577	9735
3042	8409	7053	5364	5872	1143

[I.C.W.A., July '72]

Solution There are altogether $N = 41$ members and these are serially numbered horizontally from 1 to 41. Sicne the population size $N = 41$ is two-digited, the identity numbers must also be two-digited. But there are 100 two-digited numbers possible, viz., 00, 01, 02, ..., 99.

If we allot only one identity number to each member, many random numbers will be rejected before we get the required sample. Let us allot 2 identity numbers to each member as follows:

Serial Nos	1	2	3	...	...	39	40	41
Identity	01	02	03	...	...	39	40	41
Numbers	42	43	44	...	...	80	81	82

[**Note:** It will be noticed that identity numbers upto 41 are the same as the corresponding serial numbers, and the other identity numbers when divided by 41 leave remainders which are equal to the serial numbers. With identity number 82, however, the remainder is zero, and the last observation is referred to.]

Starting from the frist digit of the given random sampling numbers and proceeding row-wise, we now taken down groups of 2 consecutive digits.

59, 67, 89 (rejected), 41, 79, 89 (rejected), 33, 35, 75, 77

97 (rejected), 35, 30, 42, 84 (rejected), 09, 70, etc.

The numbers greater than 82 have been rejected, because the identity numbers are 01 to 82 only (If 00 apeared, it will be also have been rejected). The numbers greater than 41 are replaced by their remainders on division by 41, and we now have

18, 26, 41, 38, 33, 35, 34, 36, 35 (rejected), 30, 1, 9, 29 etc.

Since the drawing are *without* replacement, the second 35 occurring above, has been rejected. Of the remaining numbers, the first ten are

18, 26, 41, 38, 33, 35, 34, 36, 30 & 1

(Note that these numbers all lie between 1 and 41, and none is repeated). The members having these as serial numbers are respectively

48, 26, 12, 31, 23, 38, 36, 20, 27, 45

These observations comprise the required random sample of size 10.

(*Alternative method*) [This is a modification of the previous method. Note that 100 identity numbers 00, 01, 02, ..., 99 and 00 can be allotted equally among 50 (slightly higher than 41) members.]

Let the population members be serially numbered horizontally from 1 to 41, as before; and two-digited random numbers are taken down. These numbers are divided by 50 (not 41) and the remainers obtained. If a number is less than 50, the number itself is taken as the remainders. Any remainders between 42 and 49, and also 00, if any are rejected. Remainders which occur more than once are retained only at the first appearane: the repetitions are rejected. Thus the two-digited random numbers

59, 67, 89, 41, 79, 89, 33, 35, 75, 77, 97, 35, 30, 42, 84, 09 leave the remainders

9, 17, 39, 41, 29, 39 (rejected, because repeated), 33, 35, 25, 27, 47 (rejected because more than 41), 35 (rejected, because repeated), 30, etc.

From the numbers of now remaining the frist ten are only retained

9, 17, 39, 41, 29, 33, 35, 25, 27, 30

The members of the population which correspond to these as serial numbers, viz.

11, 25, 24, 12, 21, 23, 38, 30, 30, 27

constitute the required random sample.

Example 13.9 *Describe in detail ow you will select without replacement a random sample of 3 units from a population of 121 units using a procedure which does not involve rejection of a large number of random numbers.* [C.U., B.A. (E.con) '75]

Solution The method is described in the following steps:

1. All the 121 units of the population are serially numbered from 1 to 121.
2. Let us now take a page of random sampling numbers, and with closed eyes select a digit from the page. Starting from this and proceeding horizontally, consecutive digits are taken down in groups of three (because of largest serial number 121 is three-digited), giving several 3-digited numbers.
3. The random numbers are divided by 125 and the remainders obtained (Note that 1000 three-digited numbers are possible, viz. 001, 002, ..., 999 and 000; and 1000 is exactly divisible 125). If a number is less than 125, that number itself is taken as the remainder.
4. Any remainders between 122 and 124, and also 000, if any occurring, are rejected. If any remainder appears more than one, all subsequent repetitions are rejected.
5. From the numbers now remaining, only the first three numbers are taken and the population units corresponding to these as serial numbers give the random sapmle.

Example 13.9A *Draw a random sample of size 10 with replacement from the following frequency distribution and compare the sample mean with the population mean:*

Annual Sales (Rs '000)	*30–39*	*40–49*	*50–59*	*60–69*	*70–79*	*80–89*	*90–99*	*100–109*	Total
No. of Firms	*14*	*36*	*47*	*66*	*41*	*24*	*10*	*2*	*240*

Solution All the 240 firms considered in the frequency distribution are serially numbered—14 observations in the class interval 30–39 are numbered 1 to 14, 36 in the next interval are numbered 15 to 50, and so on, as shown in the table below.

Table 13.1A Calculation of Cumulative Frequencies

Class Interval	Frequency	Mid value	Cumulative Frequency	Serial Nos. allotted
30–39	14	34.5	14	1–14
40–49	36	44.5	50	15–50
50–59	47	54.5	97	51–97
60–69	66	64.5	163	98–163
70–79	41	74.5	204	164–204
80–89	24	84.5	228	205–228
90–99	10	94.5	238	229–238
100–109	2	104.5	240	239–240

Using the random sampling numbers shown in the Appendix (Table VI), we start from (Row 6, Col. 1) an take down 3-digited numbers proceeding horizontally. The numbers 250 and above are replaced by their remainders on division by 250. Any number between 241–249 or 000, if occurring, is now rejected.

Random Numbers	489	572	665	890	501	154	786	475	888	750	147
Remainders	239	72	165	140	1	154	36	225	138	000*	147

(*The remainder 000 is rejected). The first 10 of the numbers are now retained, and the firms corresponding to these as serial numbers constitute the sample of firms. (Note that there is no necessity of rejecting any number which is repeated, because the sample is drawn *with* replacement.)

The random sample is given by the mid-values of classes in which the selected firms are placed. This can be done by looking at the last column and then the 3rd col. of Table 13.1A.

Random Sample (with replacement)

Serial Nos	239	72	165	140	1	154	36	225	138	147	Total
Sample Values (Rs '000)	104.5	54.5	74.5	64.5	34.5	64.5	44.5	84.5	64.5	64.5	655.0

The sample mean is 655.0 ÷ 10 = 65.5 (Rs '000). The mean calculated from the the given frequency distribution (calculations not shown here) is 63.08 (Rs '000).

Sample mean ($\bar{x}$) = 65.5, Population Mean (μ) = 63.08.

13.5 SAMPLING DISTRIBUTION

Example 13.10 *Distinguish between Statistic and Parameter, giving examples.* [W.B.H.S. '78, '82; I.C.W.A. June '73, '76, '79; Dec. '81; C.U., B.A. (Econ) '81]

Solution Any statistical measure calculated on the basis of sample observations is called a *Statistic*; e.g., sample mean, sample s.d., the proportion of defectives observed in the sample, etc. Any statistical measure based on all units in the population is called a *Parameter;* e.g., population mean, population s.d., proportion of defectives in the whole lot, etc.

In all sampling techniques (except Purposive Sampling), the composition of sample depends only on chance. Therefore, if many samples of a fixed size the drawn from a given

population, the group of units constituting the sample also varies from one sample to another. Hence, the value of a statistic varies from sample to sample; but the parameter remains a constant. This variation in the value of a statistic is known as '*sampling fluctuation*'. The parameter has no sampling fluctuation. Usually parameters are unknown and statistics are used as estimates of parameters. The probability distribution of a statistic is called its '*sampling distribution*' and the standard deviation in the sampling distribution is called '*standard error*' of the statistic. However, since the parameter is constant it has neither a sampling distribution nor a standard error.

The following notations will be used to distinguish between statistic and parameter:

	Statistic (from Sample Values)	*Parameter (from all Population Values)*
Mean	$\bar{x}$	μ
Standard Deviation	S	σ
Proportion	p	P
rth Raw Moment	m_r'	μ_r'
rth Central Moment	m_r	μ_r

Example 13.11 *Explain clearly the concept of Sampling Distribution of a statistic.*

[W.B.H.S. '78, '82; C.U., B.A. (Econ) '74, '81; C.A., Nov. '78; I.C.W.A., June '73, '76, '78, '79, Dec. '74, '81]

Solution *Sampling Distribution* of a statistic may be defined as the probability law which the statistic follows, if repeated random samples of a fixed size are drawn from a specified population.

Let us consider a random sample $x_1, x_2, ..., x_n$ of size n drawn from a population containing N units. Let us further suppose that we are interested in the sampling distribution of the statistic $\bar{x}$ (i.e., sample mean), where $\bar{x} = (x_1 + x_2 + ... + x_n)/n$.

If the population size N is finite, there is a finite number (say K) of possible ways of drawing n units in the sample out of a total of N units in the population. For each of these K samples we calculate the value of $\bar{x}$ (see Tables 13.3 and 13.5. Here $N = 4$, $n = 2$, and $K = N^n = 16$ for SRSWR and $K = {}^NC_n = 6$ for SRSWOR). Although the K samples are distinct, the sample means may not be all different; but each of these occurs with equal probability. Thus, we can construct a table shwoing the set of possible values of the statistic $\bar{x}$ and also the probability that $\bar{x}$ will take each of these values. The probability distribution of the statistic $\bar{x}$, will be called '*sampling distribution*' of sample mean (Tables 13.4 and 13.6). The above method is quite general, and the sampling distribution of any other statistic, say, median, or standard deviation of the sample, may be obtained.

If, however, the number (N) of units in the population is large, the number (K) of possible distinct samples being even larger, the above method of finding the sampling distribution cannot be applied. In this case, the values of $\bar{x}$ obtained from a large number of samples may be arranged in the form of a relative frequency distribution. The limiting form of this relative frequency distribution, when the number of samples considered becomes infinitely large, will be called *sampling distribution* of the statistic. When the population is specified by a theoretical distribution (e.g., binomial, or normal), the sampling distribution can be theoretically obtained. The knowledge of sampling distribution is necessary in finding 'confidence limits' for parameters and in 'testing statistical hypotheses'.

13.6 TWO IMPORTANT SAMPLING DISTRIBUTION (LARGE SAMPLE)

(I) Sampling Distribution of Sample Mean

If $\bar{x}$ represents the mean of a random sample of size n, drawn from a population with mean μ and standard deviation (s.d.) σ, then the sampling distribution of $\bar{x}$ is approximately a *normal distribution* with

mean = μ and s.d. = standard error of $\bar{x}$ (13.6.1)

provided the sample size n is sufficiently large.

(II) Sampling Distribution of Sample Proportion

If p represents the proportion of defectives in a random sample of size n drawn from a lot with proportion of defectives P, then the sampling distribution of p is approximately a *normal distribution* with

mean = P and s.d. = standard error of p, (13.6.2)

provided the sample size n is sufficiently large.

[**Note:** Usually a sample of size 30 or more is considered as a 'large sample'. However, the larger the value of n the better is the approximation.]

13.7 STANDARD ERROR (S.E.)

Standard Error of a statistic is the standard deviation calcualted from the sampling distribution of the statistic. Some importnat formula for standard error are shown below:

1. For random samples of size n,

$$\text{S.E. of sample mean } (\bar{x}) = \frac{\sigma}{\sqrt{n}} \tag{13.7.1}$$

$$\text{S.E. of sample proportion } (p) = \sqrt{\frac{PQ}{n}} \tag{13.7.2}$$

where σ denotes the population standard deviation (s.d.) and P the population proportion ($P + Q = 1$).

If the random sample is drawn without replacement from a finite population of size N, then the above formulae are modified on multiplication by the correction factor $\sqrt{\frac{N-n}{N-1}}$ (see Example 13.14)

2. If $\bar{x}_1$ and $\bar{x}_2$ denote the means calculated from independent random samples of sizes n_1 and n_2 drawn from the populations with standard deviations σ_1 and σ_2 respectively, then

$$\text{S.E. of } (\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{13.7.3}$$

In particular, if the poulations have the same s.d. σ, then

$$\text{S.E. of } (\bar{x}_1 - \bar{x}_2) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{13.7.3a}$$

3. If p_1 and p_2 denote the proportions calcualted from independent random samples of sizes n_1 and n_2 and drawn from two populations with proportions P_1 and P_2 respectively, then

$$\text{S.E. (of } p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}} \tag{13.7.4}$$

In particular, if it is assume that the two population proportions P_1 and P_2 are equa, say $P_1 = P_2 = P$, then

$$\text{S.E. of } (p_1 - p_2) = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{13.7.4a}$$

4. For random sample from a normal population with s.d. σ,

$$\text{S.E. of sample s.d. } (S) = \frac{\sigma}{\sqrt{2n}} \tag{13.7.5}$$

$$\text{S.E. of sample variance } (S^2) = \sigma^2\sqrt{\frac{2}{n}} \tag{13.7.6}$$

Formulae (13.7.5) and (13.7.6) are however approximate and used only when the sample size n is large (say greater than 50).

Example 13.12 *Discuss the concept of 'Standard error' of a statistic. What does the standard error of a statistic measure?*

[C.U. M.Com. '71 '77,; C.U. B.A. (Econ) '8; C.A., Nov. '78; I.C.W.A., June '73, '75, '76, '78; Dec. '81]

Solution Let $x_1, x_2, ..., x_n$ be a random sample of size n drawn from a specified population. On the basis of this sample, let us calculate the value of certain statistic, say, sample mean

$$\bar{x} = \frac{(x_1 + x_2 + \ldots x_n)}{n}$$

We repeat the process of drawing a random sample of the fixed size n a large number of times, and calcualte the value of the statistic (here, sample mean) for each sample. The relative frequency distribution of these sample means, when the number of samples considered is infinitely large, is called the *sampling distribution* of sample mean.

The sampling distribution of any statistic will have its own mean, standard deviation, moments, etc. *The standard deviation calculated from the sampling distribution of a statistic is called its 'Standard Error'*. The standard error gives a measure of dispersion of the concerned statistic. It depends on the sample size n and goes on diminishng as the sample size increases. It is used to set up 'confidence limits' for population parameters and in 'tests of significance'. Thus, the standard errors of sample mean ($\bar{x}$) and sample proportion (p) are used to find confidence limits for the population mean (μ) and the population proportion (P) respectively.

Example 13.13 *State the formulae for standard error of sample mean and sample proportion.* [M.B.A. '77; I.C.W.A., June '74]

Solution

(I) Standard Error of Sample Mean

If $\bar{x}$ represents the mean calcualted from a random sample of size n, and the standard deviation of the population be σ, then *Standard Error (S.E.) of sample mean* is given by

(i) $$\text{S.E.} = \frac{\sigma}{\sqrt{n}} \qquad (13.7.7)$$

when either the population size is infinitely large, or the sample is drawn *with replacement* (i.e., in simple sampling).

(ii) $$\text{S.E.} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad (13.7.8)$$

whent he population size N is finite, but the sample is drawn *without replacement.*

(II) Standard Error of Sample Proportion

If a random sample of size n is drawn from a population containing a proportion P of the units belonging to a certain category (e.g., proportion of defectives in a batch of articles), then *Standard Error (S.E.) of sample proportion* is given by

(i) $$\text{S.E.} = \sqrt{\frac{PQ}{n}} \qquad (13.7.9)$$

when either the population size is infinitely large, or the sample is drawn *with replacement.* ($P + Q = 1$, i.e., $Q = 1 - P$).

(ii) $$\text{S.E.} = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}} \qquad (13.7.10)$$

when the population size N is finite, but the sample is drawn *without replacement.*

It may be noted that the formulae for S.E. in simple sampling are easier. In sampling without replacement from a finite population S.E. contains an exra factor $\sqrt{\frac{N-n}{N-1}}$, which is called *finite population correction* (f.p.c.)

Example 13.14 *A simple random sample of size 5 is drawn without replacement from a finite population consisting of 41 units. If the population standard deviation is 6.25, what is the standard error of sample mean? (Use finite population correction).*

[B.U., B.A. (Econ) '73]

Solution Here, $n = 5$, $N = 41$, $\sigma = 6.25$. Using the formulae

$$\text{S.E. of mean} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$= \frac{6.25}{\sqrt{5}}\sqrt{\frac{41-5}{41-1}}$$

$$= \frac{6.25 \times 6}{10\sqrt{2}} = 2.65$$

Example 13.15 *The safety limit of a crane is known to be 32 tons. The mean weight and the standard deviation of a large number of iron rods are 0.3 ton and 0.2 ton respectively. 100 rods are lifted at a time. Find the probability of an accident.*

Solution The accident will occur when the total weight of 100 rods exceeds 32 tons, i.e., the mean weight ($\bar{x}$) is greater than 0.32 ton. So, we have to find the probability that the mean of a random sample of size 100 exceeds 0.32.

The weight (x) of a single rod has population mean $\mu = 0.3$ and s.d. $\sigma = 0.2$. Since the sample size $n = 100$ is large, the sample mean ($\bar{x}$) follows normal distribution approximately (see Section 13.6) with

$$\text{mean} = \sigma = 0.3$$

$$\text{s.d.} = \frac{\sigma}{\sqrt{n}} = \frac{0.2}{\sqrt{100}} = 0.02$$

Therefore, the standardised value of the sample mean 0.32 is

$$z = \frac{0.32 - 0.3}{0.02} = 1$$

Probability that $\bar{x}$ is greater than 0.32

= Area under the standard normal curve to the right of the ordinate at $z = 1$

= 0.1587

(Also see Example 13.26 and 13.27, p. 173)

Example 13.16 *It has been found that 2% of the tools produced by a certain machine are defective. What is the probability that in a shipment of 400 such tools, 3% or more will be defective? (Probability that the normal deviate lies between 0 and 1.43 is 0.4236).* [B.U., B.A. (Econ) '72]

Solution Since the sample size $n = 400$ is lare, the sample proportion (p) is approximately normally distributed (see Section 13.6) with

$$\text{mean} = P = 2\% = .02$$

$$\text{s.d.} = \sqrt{\frac{PQ}{n}}\sqrt{\frac{.02 \times .98}{400}} = .007$$

Probability that the sample proportion p exceds .03

= Area under the standard normal curve to the right of the ordinate at the standardised value $z = \dfrac{.03 - .02}{.007} = 1.43$

= (Area to the right of 0) – (Area between 0 and 1.43)

= 0.5 – 0.4236 = .0764 *Ans.* .0764

Example 13.17 *Define Standard Error of the mean. How is it calculated ? Explain its uses.* [C.U. B.A. (Econ) '79; C.A., Nov. '73]

Solution *Standard Error (S.E.) of mean* may be defined as the standard deviation in the sampling distribution of mean. If several random samples, each having the same number of

members, are drawn from a statistical population, it will be found that the means of these smaples are in general different from the mean of the population. A measure of the extent of this discrepancy, or 'error' in the sample mean, is provided by Standard Error of mean.

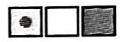
Applying the principles of probability, the formula for S.E. of mean can be theoretically derived. In fact, if σ denotes the standard deviation of the population and n represents the number of members included in the sample, then (see Examples 13.21 and 13.24)

$$\text{S.E. of mean} = \frac{\sigma}{\sqrt{n}}$$

If, however, the random sample is drawn without replacement from a finite population of size N, thenS.E. of mean

$$= \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}.$$

S.E. of mean is used to measure the extent of sampling error in the mean. If a sufficiently large number of members is included in the sample, the means obtained from many samples (all of equal size) will closly follow a pattern known as 'normal distribution' whatever be the nature of distribution in the population. S.E. of mean would then give a good account of the proportion of sample means different by given amounts from the population mean—about 2/3rd of them will differ by less than the S.E., about 95% by less than twice the S.E., and almost all sample means will differ from the population mean by less than 3 times the S.E.



Example 13.18 *How do you distinguish between 'standard error' and 'standard deviation' ?* [I.C.W.A., Dec. '78]

Solution *Standard Deviation* of a given set of values $x_1, x_2, \ldots, x_n$ is a measure of their variability or dispersion, and is defined as $\sigma = \sqrt{\frac{\Sigma(x-x)^2}{n}}$. It is irrelevant to ask whether these values relate to a sample or the the population. We may calculate the standard deviation of sample observations or of all members in the population.

The term '*Standard Error*' is however, used in relation to a statistic, which is a measure based on the sample observations. As such, we may speak of the standard error only in connection with sampling. As the group of observations constituting a sample is likely to be different from that of another sample, the value of a statistic varies from sample to sample. A measure of this variabiity of the statistic is called *standard error.* Of course, the variability is measured by standard deviation. Thus, standard error is the standard deviation of all possible values of statistic in repeated samples of a fixed size from a given population. In other words, standard error is the standard deviation calculated from the sampling distribution.

Standard error depends on (i) the sampel size, (ii) the natue of the statistic e.g., mean variance, etc., (iii) the mathematical form of the sampling distribution, and (iv) the values of some of the parameters used in the sampling distribution. The reciprocal of standard error is sometimes used to measure the precision of the statistic as an estimate of a parameter. For example, since

$$\text{S.E. of } \bar{x} = \frac{\sigma}{\sqrt{n}}, \text{ the precision of } \bar{x} \text{ is } \frac{\sqrt{n}}{\sigma}.$$

Hence, the precision of $\bar{x}$, which is used as an estimate of the population mean (μ) is directly proportional to the square-root of the sample size (n). This implies that to double the precision of the estimate we have to make the sample size four-fold.

Example 13.19 *For a population of six units, the values of a characteristic x are given below:*

$$3, \quad 9, \quad 6, \quad 5, \quad 7, \quad 10$$

Consider all possible samples of size two from the above population and show that the mean of the sample means is exactly equal to the population mean.

[C.U., M.Com. '73]

Solution The population mean is

$$\mu = \frac{(3+9+6+5+7+10)}{6} = \frac{20}{3}$$

The possible distinct samples (here random sampling *without replacement* is implied) and their means are shown below:

Table 13.2 Random Samples of Size Two (Without Replacement)

Serial No.	Sample Values	Total	Sample Mean	Serial No.	Sample Values	Total	Sample Mean
1	3, 9	12	6.0	9	9, 10	19	9.5
2	3, 6	9	4.5	10	6, 5	11	5.5
3	3, 5	8	4.0	11	6, 7	13	6.5
4	3, 7	10	5.0	12	6, 10	16	8.0
5	3, 10	13	6.5	13	5, 7	12	6.0
6	9, 6	15	7.5	14	5, 10	15	7.5
7	9, 5	14	7.0	15	7, 10	17	8.5
8	9, 7	16	8.0			Total	100.0

$$\therefore \quad \text{Mean of sample means} = \frac{100.0}{15} = \frac{20}{3}$$

This is exactly equal to the population mean $\mu = \frac{20}{3}$, show above.

[**Note:** If random sampling *with replacement* is applied, then also the mean of sample means will be exactly equal to the population mean]

Example 13.20 *A population consists of the four members 3, 7, 11, 15. Consider all possible samples of size two which can be drawn with replacement from this population. Find*

(i) the population mean,
(ii) the population standard deviation,
(iii) the mean of the sampling distribution of means,
(iv) the standard deviation of the sampling distribution of mean.

Verify (iii) and (iv) directly from (i) and (ii) by use of suitable formulae (which you are to members).

Solve this problem if sampling is without replacement! [I.C.W.A., Dec. '76]

Solution For the 4 members of the population

$$\Sigma x = 3 + 7 + 11 + 15 = 36$$
$$\Sigma x^2 = 3^2 + 7^2 + 11^2 + 15^2 = 404$$

(i) Population mean $(\mu) = \frac{36}{4} = 9$

(ii) Population variance $(\sigma^2) = \frac{404}{4} - \left(\frac{36}{4}\right)^2 = 101 - 81 = 20$

$\therefore$ Population standard deviation $(\sigma) = \sqrt{20}$.

Case I

Simple Random Sampling With Replacement (SRSWR):

The possible random samples of size 2 (with replacement) and the sample means are shown below:

Table 13.3 Possible Samples (With Replacement)

Serial No.	Sample Values	Total	Sample Mean	Serial No.	Sample Values	Total	Sample Mean
1	3, 3	6	3	9	11, 3	14	7
2	3, 7	10	5	10	11, 7	18	9
3	3, 11	14	7	11	11, 11	22	11
4	3, 15	18	9	12	11, 15	26	13
5	7, 3	10	5	13	15, 3	18	9
6	7, 7	14	7	14	15, 7	22	11
7	7, 11	18	9	15	15, 11	26	13
8	7, 15	22	11	16	15, 15	30	15

$N^n = 16$ possible samples of size 2 (with replacement) are possible, and hence each of the 16 values of sample mean $(\bar{x})$, as shown in the last column, occurs with probability $\frac{1}{16}$. The distinct possible values of $\bar{x}$, are arranged in order of magnitude and the corresponding probabilities shown against each.

Table 13.4 Sampling Distribution of Mean (With Replacement)

Sample Mean $(\bar{x})$	3	5	7	9	11	13	15	Total
Probability (p)	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	1

(iii) Mean of the sampling distributon of means:

$$E(\bar{x}) = \frac{1}{16}\times 3+\frac{1}{8}\times 5+\frac{3}{16}\times 7+\frac{1}{4}\times 9+\frac{3}{16}\times 11+\frac{1}{8}\times 13+\frac{1}{16}\times 15$$

$$= \frac{3}{16}+\frac{5}{8}+\frac{21}{16}+\frac{9}{4}+\frac{33}{16}+\frac{13}{8}+\frac{15}{16}$$

$$= 9 \qquad \ldots(A)$$

(iv) Variance of the sampling distribution of means:

$$\text{Var}(\bar{x}) = E(\bar{x}^2) - \{E(\bar{x})\}^2$$

$$= \left(\frac{1}{16}\times 3^2 + \frac{1}{8}\times 5^2 + \frac{3}{16}\times 7^2 + \frac{1}{4}\times 9^2 + \frac{3}{16}\times 11^2 + \frac{1}{5}\times 13^2 + \frac{1}{16}\times 15^2\right) - 9^2$$

$$= \left(\frac{9}{16} + \frac{25}{8} + \frac{147}{16} + \frac{81}{4} + \frac{363}{16} + \frac{169}{8} + \frac{225}{16}\right) - 81$$

$$= 91 - 81 = 10$$

$\therefore$ Standard Deviation of the sampling distribution of means

$$\text{S.D.}(\bar{x}) = \sqrt{10}$$

Using formulae (13.6.1) and (13.7.7), ...(B)

$$E(\bar{x}) = \mu = 9$$

$$\text{S.E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{20}}{\sqrt{2}} = \sqrt{10}$$

These results agree exactly with those at (A) and (B) abve, obtained by direct calculation from the sampling distribution of mean.

Case II

Simple Random Sampling Without Replacement (SRSWOR):

Table 13.5 Possible Samples (Without Replacement)

Serial No.	*Sample Values*	*Total*	*Sample mean*	*Probability*
1	3, 7	10	5	1/6
2	3, 11	14	7	1/6
3	3, 15	18	9	1/6
4	7, 11	18	9	1/6
5	7, 15	22	11	1/6
6	11, 15	26	13	1/6

Table 13.6 Sampling Distribution of Mean (Without Replacement)

Sample Mean ($\bar{x}$)	5	7	9	11	13	Total
Probability (p)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	1

Mean of the sampling distribution of means:

$$E(\bar{x}) = \frac{1}{6} \times 5 + \frac{1}{6} \times 7 + \frac{1}{3} \times 9 + \frac{1}{6} \times 11 + \frac{1}{6} \times 13$$

$$= \frac{5}{6} + \frac{7}{6} + 3 + \frac{11}{6} + \frac{13}{6} = 9$$

Variance of the sampling distribution of means:

$$\text{Var}(\bar{x}) = E(\bar{x}^2) - \{E(\bar{x})\}^2$$

$$= \left(\frac{1}{6} \times 5^2 + \frac{1}{6} \times 7^2 + \frac{1}{3} \times 9^2 + \frac{1}{6} \times 9^2 + \frac{1}{6} \times 11^2 + \frac{1}{6} \times 13^2\right) - 9^2$$

$$= \left(\frac{25}{6} + \frac{49}{6} + 27 + \frac{121}{6} + \frac{169}{6}\right) - 81$$

$$= \frac{263}{3} - 81 = \frac{20}{3}$$

Standard deviation of the sampling distribution of means:

$$\text{S.D.}(\bar{x}) = \sqrt{\frac{20}{3}}$$

Using formulae (13.6.1) and (13.7.8),

$$E(\bar{x}) = \mu = 9$$

$$\text{S.E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$= \frac{\sqrt{20}}{\sqrt{2}} \sqrt{\frac{4-2}{4-1}} = \sqrt{\frac{20}{3}}$$

These results agree exactly with those obtained above by direct calculation from the sampling distribution of $\bar{x}$.

Example 13.21 *Show that the mean and standard error of sample mean ($\bar{x}$) from simple samples of size n are:*

$$E(\bar{x}) = \mu$$

$$S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

which μ and σ denote the mean and s.d. of the population.

Solution Let $x_1, x_2, \ldots, x_n$ denote the sample observations. The sample mean is

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum x_i$$

where x_i is the i-th member of the sample (i.e., the member obtained at the i-th drawing).

In simple sampling, each of the sample members has the same probability distribution as the variable x in the population. Hence

$$E(x_i) = \mu = \text{Population mean}$$
$$\text{Var}(x_i) = E(x_i - \mu)^2 = \sigma^2 = \text{Population variance}$$

Therefore,

$$E(\bar{x}) = E\left[\frac{1}{n}(x_1 + x_2 + \ldots + x_n)\right]$$

$$= \frac{1}{n} E(x_1 + x_2 + \ldots + x_n)$$

$$= \frac{1}{n}\left[E(x_1) + (x_2) + \ldots + (x_n)\right]$$

i.e.,

$$E(\bar{x}) = \frac{1}{n}\sum E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu$$

$$= \frac{1}{n} \cdot n\mu = \mu$$

Now,

$$\text{Var}(\bar{x}) = E(\bar{x} - \mu)^2$$

$$= E\left[\frac{1}{n}\sum (x_i - \mu)\right]^2 = \frac{1}{n^2} \cdot E\left[\sum (x_i - \mu)\right]^2$$

$$= \frac{1}{n^2} E\left[\sum (x_i - \mu)^2 + \sum_{i<j} 2(x_i - \mu)(x_j - \mu)\right]$$

$$= \frac{1}{n^2}\left[\sum E(x_i - \mu)^2 + \sum_{i<j} 2E(x_i - \mu)(x_j - \mu)\right]$$

[**Note:** $(a_1 + a_2 + a_3)^2 = (a_1^2 + a_2^2 + a_3^2) + (2a_1a_2 + 2a_1a_3 + 2a_2a_3)$

$$= \sum_i a_i^2 + \sum_{i<j} 2a_ia_j]$$

But $E(x_i - \mu)^2 = \sigma^2$

$E\{(x_i - \mu)(x_j - \mu)\} = E(x_i - \mu) \cdot E(x_j - \mu)$

since x_i and x_j are independent.

$= 0 \times 0 = 0$

Substituting these results,

$$\text{Var}(\bar{x}) = \frac{1}{n^2}\left[\sum_{i=1}^{n} \sigma^2 + \sum_{i<j} 2 \times 0\right]$$

$$= \frac{1}{n^2}\left[n\sigma^2 + 0\right] = \frac{\sigma^2}{n}$$

$$\therefore \quad \text{S.E.}(\bar{x}) = \sqrt{\text{Var}(\bar{x})} \quad \frac{\sigma}{\sqrt{n}}$$

[**Note:** In random sampling without replacement from a finite population, the result $E(\bar{x}) = \mu$ holds; because each of hte sample members x_i has individual the same probability distribution as x in the population.

But, S.E.$(\bar{x}) \neq \frac{\sigma}{\sqrt{n}}$, since x_i and x_j are *not independent* in this case.]

Example 13.22 *If $x_1, x_2, \ldots x_n$ is a simple random sample of size n from a finite population of N units, show that*

(i) in drawing "with" replacement (SRSWR).

$$P(x_i = X_l) = \frac{1}{N}$$

$$P(x_i = X_k, x_j = X_l) = \frac{1}{N^2}$$

(ii) in drawing "without" replacement (SRSWOR),

$$P(x_i = X_k) = \frac{1}{N}$$

$$P(x_i = X_k, x_j = X_l) = \frac{1}{N^2}(N - 1)$$

where x_i and x_j denote the i-th and the j-th members of the sample, and X_k, X_l are two specified members of the poipulation.

Solution The simple random sample is drawn one by one and x_i, x_j denote sample members obtained at the i-th and j-th drawings respectively. Since the selection of a member in any drawing depends purely of chance, x_i and x_j may be considered as random variables.

(i) In SRSWR, the size and composition of the population remains exactly the same throughout the sampling process, and hence during the i-th drawing any of the N population members is likely to appear. These N cases are mutually exclusive, exhaustive and equally likely (because all the members are given 'equal chance'). The probability that a particular member X_k appears at the i-th drawing is therefore, by the classical definition, $\frac{1}{N}$. That is

$$P(x_i = X_k) = \frac{1}{N}$$

Since the sample is drawn "with" replacement, the selection is made from all the population members in any drawing, and hence the probability of selection of any member at the j-th drawing remains $\frac{1}{N}$, irrespective of what has actually been obtained in the i-th drawing. In other words, the random variables x_i and x_j are independent.

$$P(x_i = X_k, x_j = X_l) = P(x_i = X_k)\,(x_j = X_l)$$
$$= \left(\frac{1}{N}\right)\cdot\left(\frac{1}{N}\right) = \frac{1}{N^2}$$

(ii) In SRSWOR, the sample $x_1, x_2, \ldots x_n$ of size n may be drawn out of N units in the population in ${}^{N}P_{n}$ ways; because we may think of the n places of the sample units (corresponding to the n drawings) as being filled up b n things out of N in the population. These ${}^{N}P_{n}$ cases are mutually exclusive, exhaustive and equally likely. The number of cases favourable to the event that X_k is selected in the i-th drawing (i.e., $x_i = X_k$) may be obtained by considering X_k as having occupied the i-th place and filling up the remaining $(n-1)$ places by $(n-1)$ things out of the remaining $(N-1)$ things left in the population (excluding X_k), viz., ${}^{N-1}P_{n-1}$. Hence, by the classical definition of probability,

$$P(x_i = X_k) = \frac{{}^{N-1}P_{n-1}}{{}^{N}P_{n}} = \frac{(N-1)(N-2)\ldots(N-n+1)}{N(N-1)\ldots(N-n+1)} = \frac{1}{N}$$

Similarly, the number of cases favourable to the event that X_k is selected at the i-th drwing and X_l at the j-th drawing (i.e., $x_i = X_k, x_j = X_l$) is the by the number of permutations of $(n-2)$ things out of $(N-2)$, viz. ${}^{N-2}P_{n-2}$. Hence by the classical definition,

$$P(x_i = X_k, x_j = X_l) = \frac{{}^{N-2}P_{n-2}}{{}^{N}P_{n}} = \frac{1}{N(N-1)}$$

Example 13.23 *If $x_1, x_2, \ldots x_n$ is a simple random samle of size n from a finite population of N units with mean μ and variance σ^2, show that*

(a) $E(x_i) = \mu$

(b) $Var(x_i) = \sigma^2$

(c) $Cov(x_i, x_j) = 0$ *in SRSWR*

$= \frac{-\sigma^2}{N-1}$ *in SRSWOR*

where x_i and x_j denote the sample units obtained at the i-th and j-th drawings respectively.

Solution Let $X_1, X_2, \ldots, X_N$ denote the members of the population from which the sample is drawn. Then

$$\text{Population Mean } (\mu) = \frac{(X_1 + X_2 + \ldots + X_N)}{N} \qquad \ldots(1)$$

$$\text{Population variance } (\sigma^2) = \frac{\Sigma(X_k - \mu)^2}{N}$$

where the summation extends over all values of $k = 1, 2, \ldots N$. ...(2)

(a) In SRSWR, any of the population members may appear at the i-th drawing; i.e., the random variable x_i take the possible values $X_1, X_2, \ldots, X_N$ with equal probability $\frac{1}{N}$. (Example 13.22)

x_i :	X_1	X_2	...	...	X_N	Total
Prob. :	$\frac{1}{N}$	$\frac{1}{N}$	...	...	$\frac{1}{N}$	1

Therefore, by the definition of Expectation (page 385),

$$E(x_i) = \left(\frac{1}{N}\right) X_1 + \left(\frac{1}{N}\right) X_2 + \ldots\ldots + \left(\frac{1}{N}\right) X_N$$

$$= \frac{(X_1 + X_2 + \ldots + X_n)}{N}$$

$$= \mu, \text{ by (1)}$$

(b) $\text{Var } (x_i) = E\{x_i - E(x_i)\}^2$, by definition

$= E(x_i - \mu)^2$, since $E(x_i) = \mu$

$$= \sum_{k=1}^{N} (X_k - \mu)^2 \cdot P(x_i = X_k)$$

$$= \sum_{k} (X_k - \mu)^2 \cdot \left(\frac{1}{N}\right), \text{ from Example 13.22}$$

$$= \frac{\sum (X_k - \mu)^2}{N}$$

$= \sigma^2$, from (2)

It may be noted that the results (a) and (b) hold both for SRSWR and SRSWOR.

(c) In SRSWR, the random variable x_i and x_j are independent. Hence,

$$\text{Cov } (x_i, x_j) = 0$$

In SRSWOR, the random variable x_i and x_j are not independent.

$\text{Cov } (x_i, x_j) = E[\{x_i - E(x_j)\}\ \{x_i - E(x_j)\}]$, by defintion

$= E[(x_i - \mu)\ (x_j - \mu)]$

$$= \sum_{k \neq l} (X_k - \mu)\ (X_l - \mu).\ P(x_i = X_k,\ x_j = X_l)$$

$$= \sum_{k \neq l} (X_k - \mu)\ (X_l - \mu).\ \frac{1}{N(N-1)}, \text{ by Example 13.22}$$

$$= \frac{1}{N(N-1)} \sum_{k=1}^{N} (X_k - \mu) \sum_{1 \neq k = 1}^{N} (X_l - \mu)$$

$$= \frac{1}{N(N-1)} \sum_{k} (X_k - \mu) \left\{ \sum_{l=1}^{N} (X_l - \mu) - (X_k - \mu) \right\}$$

$$= \frac{1}{N(N-1)} \sum_{k} (X_k - \mu)\ \{0 - (X_k - \mu)\}$$

$$= \frac{1}{N(N-1)} \sum \{-X_k - \mu)^2\}$$

$$= \frac{-1}{N-1} \cdot \frac{1}{N} \Sigma (X_k - \mu)^2$$

$$= \frac{-1}{N-1} \cdot \sigma^2$$

$$= \frac{-\sigma^2}{N-1}$$

Example 13.24 *Derive the formulae for standard error of sample mean in both simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR) from a finite population.* [W.B.H.S. '80]

Solution Let $x_1, x_2, ..., x_n$ denote a simple random sample of size n from a finite population of size N, with mean μ and standard deviation σ. The sampel mean is

$$\bar{x} = \frac{(x_1 + x_2 + ... + x_n)}{N} = \sum_{i=1}^{n} \frac{x_i}{n}$$

where x_i is the i-th member of the sample.

$$\therefore \quad E(\bar{x}) = E\left(\frac{1}{n} \Sigma x_i\right) = \frac{1}{n} \Sigma E(x_i)$$

But $E(x_i) = \mu$, in both SRSWR and SRSWOR (see Example 13.23)

$$\therefore \quad E(\bar{x}) = \sum_{i=1}^{n} \frac{\mu}{n} = \frac{n\mu}{n} = \mu \qquad ...(1)$$

The result holds for both SRSWR and SRSWOR. Hence

$$\text{Var}(\bar{x}) = E(\bar{x} - \mu)^2$$

$$= E\left[\sum \frac{(x_i - \mu)}{n}\right]^2$$

$$= \frac{1}{n^2} \cdot E\left[\sum (x_i - \mu)\right]^2$$

$$= \frac{1}{n^2} \cdot E\left[\sum_i (x_i - \mu)^2 + \sum\sum_{i \neq j} (x_i - \mu)(x_j - \mu)\right]$$

because $$\left(\sum a_i\right)^2 = \sum a_i^2 + \sum\sum_{i \neq j} a_i a_j$$

$$= \frac{1}{n^2} \cdot \left[\Sigma E(x_i - \mu)^2 + \Sigma\Sigma E\{(x_i - \mu)(x_j - \mu)\}\right]$$

But, $E(x_i - \mu)^2 = \text{Var}(x_i)$, $E\{(x_i - m)(x_j - m)\} = \text{Cov}(x_i, x_j)$

$$\therefore \quad \text{Var}(\bar{x}) = \frac{1}{n^2}\left[\sum_i \text{Var}(xi) + \sum_{i \neq 1}\sum \text{Cov}(x_i, x_j)\right] \qquad ... (2)$$

This result holds for both the case SRSWR and SRSWOR.

Case I

Simpl random sampling "with" replacement (SRSWR):

Here x_i and x_j are independent.

$$\text{Var}(x_i) = \sigma^2$$

$$\text{Cov}(x_i, x_j) = 0$$ (see Example 13.23)

Substituting these results in (2), we have

$$\text{Var}(\bar{x}) = \frac{1}{n^2}\left[\sum_i \sigma^2 + \sum\sum_{i \neq j} 0\right]$$

$$= \frac{1}{n^2}\left[n\sigma^2 + 0\right] = \frac{\sigma^2}{n}$$

Thus, in SRSWR, the standard error of sample mean is

$$\text{S.E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Case II

Simple random sampling "without" replacement (SRSWOR).

Here x_i and x_j are not independent

$$\text{Var}(x_i) = \sigma^2$$

$$\text{Cov}(x_i, x_j) = \frac{-\sigma^2}{N-1} \quad \text{(see Example 13.23)}$$

Substituting in (2),

$$\text{Var}(\bar{x}) = \frac{1}{n^2}\left[\sum_i \sigma^2 + \sum\sum_{i \neq j}\left(\frac{-\sigma^2}{N-1}\right)\right]$$

$$= \frac{1}{n^2}\left[n\sigma^2 - n(n-1)\frac{\sigma^2}{N-1}\right]$$

because there are $n(n-1)$ possible pairs of values $i \neq j$.

$$\text{Var}(\bar{x}) = \frac{n\sigma^2}{n^2}\left[1 - \frac{n-1}{N-1}\right] = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

Thus, in SRSWOR, the standard error of sample mean is

$$\text{S.E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Example 13.25 *Find the expectation and standard error of sample proportion in both simple random sampling with replacement and simple random sampling without replacement from a finite population.*

Solution Let a simple random simple of size n be drawn from a population in which the proportion of members belonging to a particular category (e.g., proportion of defective articles) be P. If x denotes the number of such members included in the sample, then

$$\text{Sample proportion } (p) = \frac{x}{n}$$

where $x = 0, 1, 2, \ldots, n$. Therefore

$$E(p) = \frac{E(x)}{n} \quad \ldots(1)$$

$$\text{Var}(p) = \text{Var}\,\frac{(x)}{n^2} \quad \ldots(2)$$

(Note that Sample proportion = p, Population proportion = P).

Case I

Simple random sampling "with" replacement (SRSWR).

Here the probability of "success" at each trial is a constant P, and the n trials are independent. Therefore, x follows *binomial* distribution (p. 433) with

$$E(x) = np$$
$$\text{Var}(x) = nPQ$$

where $Q = 1 - P$. Substituting these results in (1) and (2),

$$E(p) = \frac{nP}{n} = P$$
$$\text{Var}(p) = \frac{nPQ}{n^2} = \frac{PQ}{n}$$

Thus, in SRSWR,

$$E(p) = P$$
$$\text{S.E.}(p) = \sqrt{\frac{PQ}{n}}$$

Case II

Simple random sampling "without" replacement (SRSWOR).

Here the drawing is "without" replacement, and hence x follows *Hypergeomtric* distribution (page 447) with

$$E(x) = nP$$
$$\text{Var}(x) = nPQ\left(\frac{N-n}{N-1}\right)$$

Substituting in (1) and (2),

$$E(p) = P$$
$$\text{Var}(p) = \frac{PQ}{n} \cdot \left(\frac{N-n}{N-1}\right);$$

Thus, in SRSWOR,

$$E(p) = P$$
$$\text{S.E.}(p) = \sqrt{\frac{PQ}{n} \cdot \left(\frac{N-n}{N-1}\right)}$$

Example 13.26 *The mean of a certain normal distribution is equal to the standard error of the mean of samples of 25 from the distribution. Find the probability that the mean of a sample of 49 from the distribution will be negative. (Given: Area under standard normal curve to the left of the ordinate at 1.4 is 0.9192).*

Solution Let μ and σ denote the mean and s.d. of the normal distribution. Here, it is stated that

$$\text{Mean} = \text{S.E.}(\bar{x}) \text{ in samples of size } 25$$

i.e.,
$$\mu = \frac{\sigma}{\sqrt{25}}, \text{ since S.E.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

or,
$$\sigma = 5\mu \quad \ldots \text{(i)}$$

We have to find the probability $P(\bar{x} < 0)$, when $n = 40$. Since the population is normal,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \text{ is a standard normal variate.}$$

$$\therefore \quad P(\bar{x} < 0) = P\left\{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{0 - \mu}{\sigma/\sqrt{n}}\right\}, \text{ when } n = 49,$$

$$= P\left\{z < \frac{-\mu\sqrt{49}}{5\mu}\right\}, \text{ since } \sigma = 5\mu, \text{ from (i)}$$

$$= P(z < -1.4)$$

$$= \phi(-1.4) = 1 - \phi(1.4)$$

$$= 1 - 0.9192 = 0.0808.$$

Ans. 0.0808

Example 13.27 *The guaranteed average life of a certain type of electric light bulbs is 1000 hours with a standard deviation of 125 hours. It is proposed to sample the output so as to assure that 90% of the bulbs do not fall short of the guaranteed average by more than 2.5 per cent. What should be the minimum size of the sample? (The area under standard normal curve from $z = 0$ to $z = 1.28$ is 0.4000).*

[I.C.W.A., Dec. 82]

Solution Let x denote the life (in hours) of the electric light bulbs. Assuming normal distribution, we have $\mu = 1000$, $\sigma = 125$. It is required to find the value of sample size (n) such that

$$P\{\bar{x} - \mu > -2.5\% \text{ of } \mu\} = 0.90$$

or, $$P\{\bar{x} - \mu > -25\} = 0.90, \quad \text{since } \mu = 1000$$

or, $$P\left\{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{-25}{\sigma/\sqrt{n}}\right\} = 0.90$$

or, $$P\{z > -\sqrt{n}/5\} = 0.90, \quad \text{since } \sigma = 125$$

or, $$P\{z < -\sqrt{n}/5\} = 1 - 0.90 = 0.10$$

or, $$P\{z > \sqrt{n}/5\} = 0.10, \quad \text{... (i)}$$

since the curve is symmetrical about $z = 0$

From the given area under standard normal curve, we find that the area to the *right* of $z = 1.28$ is $0.5 - 0.4000 = 0.10$. That is $P(z > 1.28) = 0.10$.

Comparing this with (i), we find that

$$\sqrt{n}/5 = 1.28, \quad \text{Hence } n = 40.96 = 41.$$

Ans. 41

(Also see Example 14.38).

13.7A PROBABLE ERROR (P.E.)

The *probable error* of a statistic may be defined to be a quantity such that exactly half of the possible values of the statistic differ from its expected value by more than that quantity. If t is a statistic and θ its expected value, then the probability that the difference of t and θ exceeds the probable error (P.E.) is exactly $\frac{1}{2}$.

$$P\{|t - \theta| > \text{P.E.}\} = \frac{1}{2} \qquad (13.7\text{ A.1})$$

Relations between Standard Normal, χ^2, t and F distribution

(a) A chi-square variate with n degrees of freedom (d.f.) is the sum of the squares of n independent standard normal variates.

$$\chi^2 = \sum_{i=1}^{n} z_i^2$$

(b) A t-variate with n d.f. is the ratio of a standard normal variate and the square-root of an independent chi-square variate when divided by its degree of freedom.

$$t = \frac{z}{\sqrt{y/n}}$$

(c) An F-variate with (n_1, n_2) d.f. is the ratio of two independent chi-square variates each divided by its own d.f. (chi-square with d.f. n_1 appears in the numerator).

$$F = \frac{y_1/n_1}{y_2/n_2}$$

(d) An F-variate with d.f. $(1, n)$ is equivalent to the square of a t-variate with n d.f. The realation between upper percentage points is

$$F_{p,\,(1,\,n)} = t^2_{p/2,\,n}$$

(e) When the the d.f. n is large, $\sqrt{2\chi^2} - \sqrt{2n-1}$ is approximately a standard normal variate.

(f) When the d.f. n is large, a t-variate is approximately a standard normal variate. The relation between percentage points is

$$z_p = t_{p,\,\infty}$$

(g) For large degrees of freedom, both the chi-square and the F distribution tend to nromal distributions.

EXERCISES

1. What are the main objects of sampling? Compare and contrast the merits and drawbacks of Sample and Census studies. *[C.U., M.Com. '80]*
2. Distinguish between a *population* and a *sample*. What is a random sample ? Describe some method of drawing such a sample from a finite population. *[C.U., M.Com '72]*
3. What are sampling and non-sampling errors? *[D.S.W. '78]*
4. Define simple random sampling. Distinguish between simple random sampling with replacement and simple random sampling without replacement from a finite population. *[W.B.H.S. '80, '82]*
5. Define 'simple random sampling' and 'stratified random sampling'. What are 'random numbers' and how can you use them? *[M.B.A. '78]*
6. Distinguish between simple random sampling and systematic sampling. *[I.C.W.A., June '81]*
7. Describe briefly any three types of sampling commonly used. *[C.A., Nov. '81]*

8. What are 'random number' and how can you use them in drawing a random sample from a finite population? Give the expression of the standard error of the sample mean. *[D.M. '78]*
9. (a) From a population with 20 members a random sample without replacement of 2 members is taken. Which of the following is the possible number of all such samples—400, 40, 380, 190? *[W.B.H.S. '78]*
   (b) A random sample of 2 individuals is to be drawn from a population of size 43. What is the possible number of distinct samples when sampling is (i) with replacement order of drawing to be taken into account), and (ii) without replacement (order of drawing to be ignored). *[W.B.H.S. '79]*
10. How can you draw a random sample of 10 patients from a hospital accommodating 310 patients? What do you mean by the sampling distribution of $\bar{x}$ and its standard error? *[D.S.W. '77]*
11. Explain how you will select without replacement a random sample of 3 units from a population of 124 units using a procedure which is fairly simple and does not involve the rejection of a large number of random number. *[(C.U., B.A., (Econ) '74]*
12. Describe in each of the following cases an appropriate method of drawing a random sample:
    (i) Drawing a random sample of 2 members with replacement from a family of 10.
    (ii) Drawing a random sample of 15 answer-scripts *without* replacement from 1243 answer-scripts.
    (iii) Drawing a random sample of 3 houses *without* replacement from a locality in which there are 4 streets with 10, 12, 4 and 7 houses respectively. *[W.B.H.S. '78]*
13. There are 220 students in a school. Their Roll Nos. are from 1 to 220. Explain the method of drawing a simple random sample of size 5 without replacement from the above population with the help of random sampling number series table. *[W.B.H.S. '81]*
14. Describe a procedure for drawing a random sample of size 8 from a population of size 24 (without replacement method). *[I.C.W.A., June '81]*
15. How can you draw a random sample of 100 workers from a population of 1000 workers? If in your sample the mean and S.D. of weakly wages come out as Rs 48 and Rs 11 respectively, how will you get the standard error of the sample mean ? *[D.S.W. '76]*
16. What is meant by 'stratified random sampling? Explain the procedure and advantages of stratification. *[I.C.W.A., July '70]*
17. Describe the techniques of 'systematic sampling' and 'stratified random sampling', and state the situations where each of them will be suitable. *[I.C.W.A., June '75]*
18. Define "statistic" and its "sampling distribution". *[C.U., M.Com. '79]*
19. Explain what is meant by 'sampling fluctuation' and 'sampling distribution' of a statistic. *[W.B.H.S. '82]*
20. What is meant by a 'statistic' and its 'standard error? Give expressions for the standard error of the sample mean and the sample proportion. *[M.B.A., '77]*

21. Explain briefly the two terms: Standard Deviation and Standard Error. Mention at least two utilities of Standard Error in the Sampling theory.
[I.C.W.A., Dec. '78]

22. The values of a characteristic x of a population containing six units are given below: 2, 6, 5, 1, 7, 3. Take all possible samples of size two and verify that the mean of the population is exactly equal to the mean of the sample means.
[C.U., M.Com. '71]

23. A population consists of 5 numbers (2, 3, 6, 8, 11). Consider all possible samples of size two which can be drawn with replacement from this population. Calculate the S.E. of sample means. *[I.C.W.A., June '78]*

24. Define simple random sampling with replacement (SRSWR). What do you know about the sampling distribution of the sample mean when sampling is SRSWR? What do you know about its standard error?
[C.U., B.A.(Econ) '76]

25. Compute the standard error of the mean and construct the sampling distribution of the mean for simple random samples of two families each from a population of 5 families which is given below:

Family	A	B	C	D	E
Family Size	4	3	2	5	7

[C.U., M.Com. '74]

26. Ages of 5 persons have been recorded (in years) as 14, 19, 17, 20, 25. For random samples of size 3 drawn without replacement from this population, obtain the sampling distribution of $\bar{x}$. Show that the mean of $\bar{x}$ equals the population mean and obtain the S.E. of $\bar{x}$, directly from the sampling distribution and also by using the formula. *[W.B.H.S., '79]*

27. The mean μ of a certain population is equal to the standard error of mean of random samples of size 100 from that population. Find out the standard error of the mean of random samples of size 36 from the population, in terms, of μ. (Assume that the population size is very large.)

28. If it costs one rupee to draw one member of a sample, how much will it cost to take a random sample of such a size that the standard error of sample mean is 0.1% of population mean, when the population is very large-sized one with, its mean = 100 and s.d. = 1. How will you modify your answer, if the population size is 100, instead of very large?

29. Define standard error of a statistic and derive it when the statistic is sample proportion in a random sample of size n, drawn without replacement.
[W.B.H.S., '82]

30. Explain each of the following statements, and state (giving reasons) whether it is true or false:
    (a) Random sampling is sampling in a haphazard manner.
    (b) The larger the sample (in random sampling), the smaller is the standard error of the sample mean.
    (c) The value of a statistic in a random sample must be near the value of the corresponding parameter. *[I.C.W.A., Dec. '82]*

31. Pick, out the correct alternative:
    (a) Sampling distribution is the distribution of : Parameter : Statistic : Sample : Population.
    (b) In random sample with replacement from a population with standard deviation σ, if the sample size is equal to the population size ($= N$), then the standard error of sample mean will be: 0 ; σ ; $\sigma/\sqrt{N}$.
32. If a random variable x is distributed normally around a mean 20 with a s.d. 3, describe the important characteristics of the probability distribution of $y = (x - 20)/3$. *[M.B.A. '78]*
33. A normal population has a mean 0.1 and a standard deviation 2.1. Find the probability that the mean of a sample of size 900 will be negative, it being given that the probability that the absolute value of a standard normal variate exceeds 1.43 is 0.153. *[C.U., B.Sc. (Math) '78]*
34. The variable x is normally distributed with mean 68 inches and s.d. 2.5 inches. What is the size of the sample whose mean shall not differ from the population mean by more than one inch, with probability 0.95? (Given that the area under standard normal curve to the right of the ordinate at 1.96 is 0.025).
35. A random sample of size 25 is taken from a $N(\mu, \sigma^2)$ with $\mu = 30$ and $\sigma^2 = 16$. Would the probability that the sample mean would lie between 25 and 35 be greater than 0.99? *[C.U., B.Sc. (Econ) '82]*
36. If $x_1, x_2, \ldots x_n$ be a random sample from a normal population with mean zero and variance σ^2, define "chi-square" as a function of these observations.
37. Describe the important characteristics of t and F distributions.
38. If X_1, X_2, X_3 be a random sample from $N(0, \sigma^2)$ population, what is the distribution of $(X_1^2 + X_2^2 + X_3^2)/\sigma^2$? State (without derivation) the sampling distributions of the statistics.

$$\frac{\sqrt{2}X_1}{\sqrt{(X_2^2 + X_3^2)}} \quad \text{and} \quad \frac{X_1^2}{X_2^2}$$

mentioning the appropriate degrees of freedom in each case.

ANSWERS

9. (a) 190, (b) 1849; 903.
15. 1.04 Rs
23. $\sqrt{5.4} = 2.32$
25. (i) Sampling Distribution of Sample Mean

Mean ($\bar{x}$)	2.5	3	3.5	4	4.5	5	5.5	6	Total
Probability	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.1	1.0

    (ii) Standard Error of mean = $\sqrt{1.11}$
26. S.E. = $\sqrt{2.2}$ years
27. (5/3) μ

28. Rs 100; Rs 91
30. (a) False; it is sampling with equal probability.
    (b) True; as n increases, SE($\bar{x}$) decreases.
    (c) False; it depends on the sampling distribution of the statistic and its standard error.
31. (a) Statistic, (b) $\sigma / \sqrt{N}$
32. Standard normal
33. 0.0765
34. 24
35. Yes; $P(25 < \bar{x} < 35) = P(-6.25 < z < 6.25) > 0.99$
36. $\Sigma x_i^2/\sigma^2$
38. Chi-square distribution with 3 d.f.; t distribution with 2 d.f.; F distribution with (1, 1) d.f.