7

Correlation Analysis

INTRODUCTION

So far we have studied problems relating to one variable only. In business we come across a large number of problems involving the use of two or more than two variables. If two quantities vary in such a way that movements in one are accompanied by movements in the other, these quantities are said to be correlated. For example, there exists some relationship between family income and expenditure on luxury items, price of a commodity and amount demanded, increase in rainfall up to a point and production of rice, an increase in the number of television licences and number of cinema admissions, etc. The statistical tool with the help of which these relationships between two or more than two variables is studied is called **correlation***. The measure of correlation called the coefficient of correlation (denoted by the symbol r) summarizes in one figure the direction and degree of correlation. Thus correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables. A very simple definition of correlation is that given by A.M. Tuttle. He defines correlation as : "An analysis of the covariation of two or more variables is usually called *correlation.*"

The problem of analysing the relation between different series should be broken down into three steps :

(1) Determining whether a relation exists and, if it does, measuring it;

(2) Testing whether it is significant; and

(3) Establishing the cause-and-effect relations, if any.

In this chapter only the first aspect will be discussed. For second aspect a reference may be made to chapter on Tests on Hypothesis. The third aspect in the analysis, that of establishing the cause-effect relation, is beyond the scope of this text. An extremely high and significant correlation between the increase in smoking and increase in lung cancer would not prove that smoking causes lung cancer.

It should be noted that the detection and analysis of correlation (*i.e.,* convariation) between two statistical variables requires relationship of some sort which associates the observation in pairs, one of each pair being a value of each of the two variables. In general, the pairing relationship may be of almost any nature, such as observations at the same time or place or over a period of time or different places.

Significance of the Study of Correlation

The study of correlation is of immense use in practical life because of the following reasons :

1. Most of the variables show some kind of relationship between price and supply, income and expenditure, etc. With the help of correlation analysis we can measure in one figure the degree of relationship existing between the variables.

---

*"When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."—Croxton and Cowden : *Applied General Statistics.*

2. Once we know that two variables are closely related, we can estimate the value of one variable given the value of another. This is done with the help of regression analysis which is discussed in the next chapter.

3. Correlation analysis contributes to the economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connection by which disturbances spread and suggest to him the paths through which stabilising forces become effective.

In business, correlation analysis enables the executive to estimate costs, sales, price and other variables on the basis of some other series with which these costs, sales, or prices may be functionally related. Some of the guesswork can be removed from decisions when the relationship between a variable to be estimated and the one or more other variables on which it depends are close and reasonably invariant.

4. Progressive development in the methods of science and philosophy has been characterised by increase in the knowledge of relationship or correlations. Nature has been found to be multiplicity of inter-related forces.

However, it should be noted that coefficient of correlation is one of the most widely used and also one of the most widely abused statistical measures. It is abused in the sense that one sometimes overlooks the fact that correlation measures nothing but the strength of *linear* relationships and that it does not necessarily imply a relationship.

Correlation and Causation

Correlation analysis helps us in determining the degree of relationship between two or more variables—it does not tell us anything about cause-effect relationship. Even a high degree of correlation does not necessarily mean that a relationship of cause and effect exists between the variables or, simply stated, correlation does not necessarily imply causation or functional relationship though the existence of causation always implies correlation. By itself it establishes only *covariation.* The explanation of significant degree of correlation may be any one, or a combination of the following factors :

1. *The correlation may be due to pure chance, especially in a small sample.* We may get a high degree of correlation between two variables in the sample but in the universe, there may not be any relationship between the variables at all. This is especially so in case of small samples. Such a correlation may arise either because of pure random sampling variation or because of the bias of the investigator in selecting the sample. The following example shall illustrate the point :

Advertisement expenditure (Rs. lakhs)	*Sales (Rs. crores)*
25	120
35	140
45	160
55	180
65	200

The above data show a perfect positive relationship between advertisement expenditure and sales, *i.e.*, as the advertisement expenditure is increasing, the sales are also increasing and the ratio of change between the two variables is the same. However, such a situation is rare in practice.

2. *Both the correlated variables may be influenced by one or more other variables.* It is just possible that a high degree of correlation between the variables may be due to the same causes affecting each variable or different causes affecting each with the same effect. For example, a high degree of correlation between the yield per acre of rice and tea may be due to the fact that both are related to the amount of rainfall. But none of the two variables is the cause of the other.

3. *Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect.* There may be a high degree of correlation between the variables but it may be difficult to pinpoint as to which is the cause and which is the effect. This is especially likely to be so in case of economic variables. For example, such variables as demand and supply, price and production, etc., mutually interact. To take a specific case, it is a well-known principle of economics that as the price of a commodity increases, its demand goes down and so price is the cause and demand the effect. But it is also possible that increased demand of a commodity due to growth of population or other reasons may force its price up. Now the cause is the increased demand, the effect the price. Thus at times it may become difficult to explain from the two correlated variables which is the cause and which is the effect because both may be reacting on each other.

The above points clearly bring out the fact that correlation does not manifest causation or functional relationship. By itself, it establishes only covariation. Correlation observed between variables that could not conceivably be causally related are called *spurious or nonsense correlation.* More appropriately, we should remember that it is the *interpretation* of the degree of correlation that is spurious, not the degree of correlation itself. The high degree of correlation indicates only the mathematical result. We should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. A last word of warning: Errors in correlation analysis include not only reading causation into spurious correlation but also interpreting spuriously a perfectly valid association.

Types of Correlation

Correlation is described or classified in several different ways. Three of the most important are :

(*i*) Positive and negative ;

(*ii*) Simple, partial and multiple ; and

(*iii*) Linear and non-linear.

(*i*) **Positive and Negative Correlation.** Whether correlation is positive (direct) or negative (inverse) would depend upon the direction of change of the variable. If both the variables are varying in the same direction, *i.e.,* if one variable is increasing the other *on an average* is also increasing or, if one variable is decreasing the other *on an average* is also decreasing, correlation is said to be positive. If, on the other hand, the variables are varying in opposite directions, *i.e.,* as one variable is increasing the other is decreasing or *vice versa,* correlation is said to be negative. The following examples would illustrate positive and negative correlation :

POSITIVE CORRELATION

X	Y
10	15
12	20
11	22
18	25
20	37

POSITIVE CORRELATION

X	Y
80	50
70	45
60	30
40	20
30	10

NEGATIVE CORRELATION

X	Y
20	40
30	30
40	22
60	15
80	16

NEGATIVE CORRELATION

X	Y
100	10
90	20
60	30
40	40
30	50

(*ii*) **Simple, Partial and Multiple Correlation.** The distinction between simple, partial and multiple correlation is based upon the number of variables studied. When only two variables are studied it is a problem of simple correlation. When three or more variables are studied it is a problem of either multiple or partial correlation. In multiple correlation three or more variables are studied simultaneously. For example, when we study the relationship between the yield of rice per acre and both the amount of rainfall and the amount of fertilisers used, it is a problem of multiple correlation. Similarly, the relationship of plastic hardness, temperature and pressure is multivariate. In partial correlation we recognise more than two variables. But consider only two variables to be influencing each other, the effect of other influencing variable being kept constant. For example, in the rice problem taken above if we limit our correlation analysis of yield and rainfall to periods when a certain average daily temperature existed, it becomes a problem of partial correlation. In this chapter, we shall study problems relating to simple correlation only.

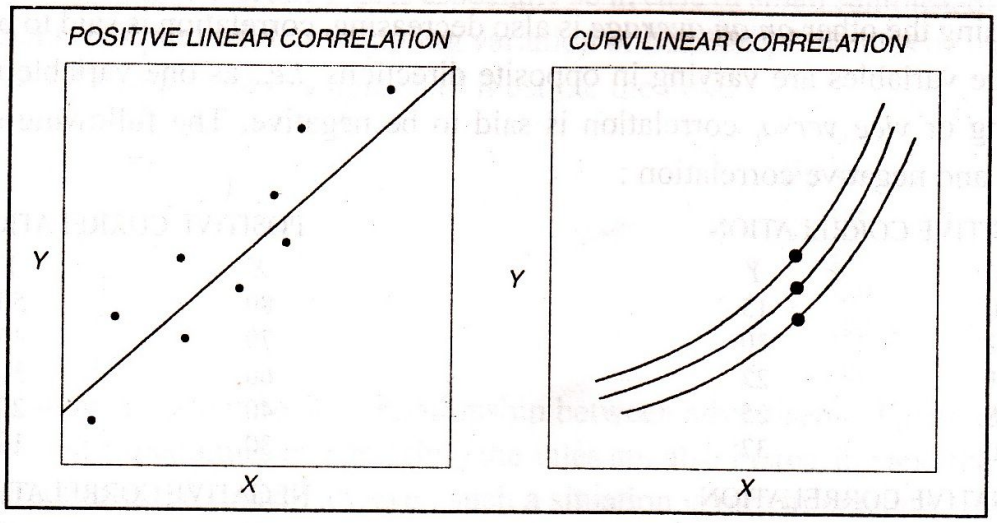
(*iii*) **Linear and Non-linear (Curvilinear) Correlation.** The distinction between linear and non-linear correlation is based upon the constancy of the ratio of change between the variables. If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then the correlation is said to be linear. For example, observe the following two variables X and Y :

X:	10	20	30	40	50
Y:	70	140	210	280	350

It is clear that the ratio of change between the two variables is the same. If such variables are plotted on a graph paper, all the plotted points would fall on a straight line.

Correlation would be called non-linear or curvilinear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable. For example, if we double the amount of rainfall, the production of rice or wheat, etc., would not necessarily be doubled. It may be pointed out that in most practical cases we find a non-linear relationship between the variables. However, since techniques of analysis for measuring non-linear correlation are far more complicated than those for linear correlation, we generally make an assumption that the relationship between the variables is of the linear type.

The following two diagrams will illustrate the difference between linear and curvilinear correlation :



METHODS OF STUDYING CORRELATION

The following are the important methods of ascertaining whether two variables are correlated or not :

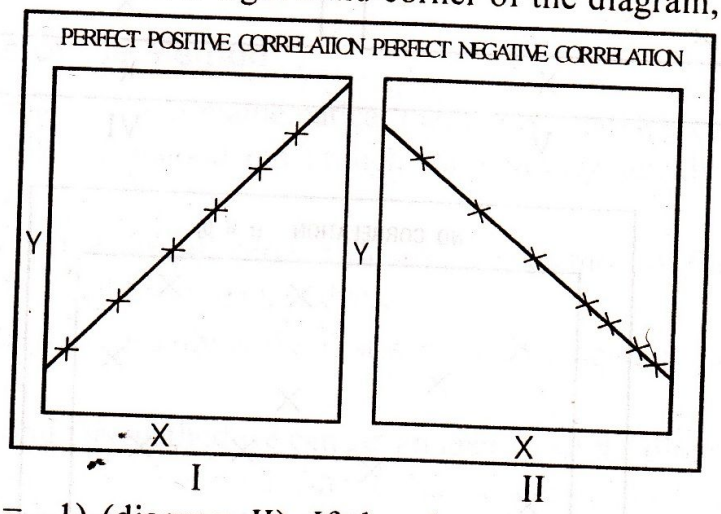
I. Scatter Diagram Method ;

II. Karl Pearson's Coefficient of Correlation ;
III. Spearman's Rank Correlation Coefficient ; and
IV. Method of Least Squares.*

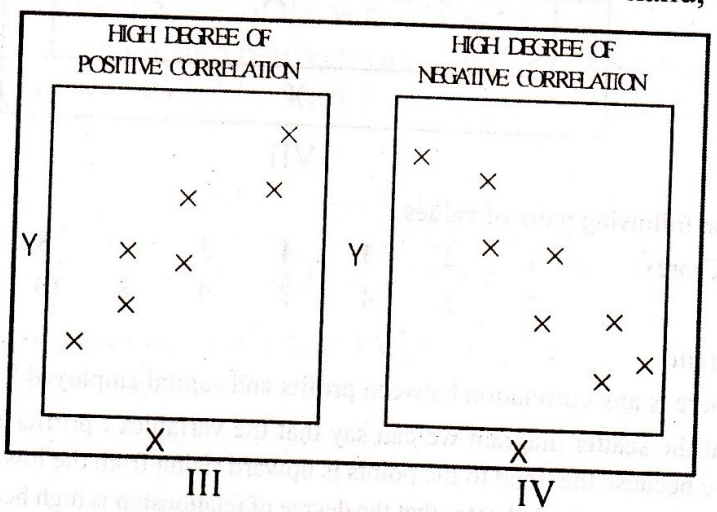
Of these, the first one is based on the knowledge of graphs whereas the others are the mathematical methods. Each of these methods shall be discussed in detail in the following pages.

I. SCATTER DIAGRAM METHOD

The simplest device for studying correlation in two variables is a special type of dot chart called dotogram or scatter diagram. When this method is used, the given data are plotted on a graph paper in the form of dots, *i.e.*, for each pair of X and Y values we put dots and thus obtain as many points as the number of observations. By looking to the scatter of the various points, we can form an idea as to whether the variables are related or not. The more the plotted points "scatter" over a chart, the lesser is the degree of relationship in between the two variables. The more nearly the points come to the line, the higher the degree of relationship. If all the points lie on a straight line falling from the lower left-hand corner to the upper right-hand corner, correlation is said to be perfectly positive (*i.e.*, $r = +1$) (diagram I). On the other hand, if all the points are lying on a straight line rising from the upper left-hand corner to the lower right-hand corner of the diagram, correlation is said to be

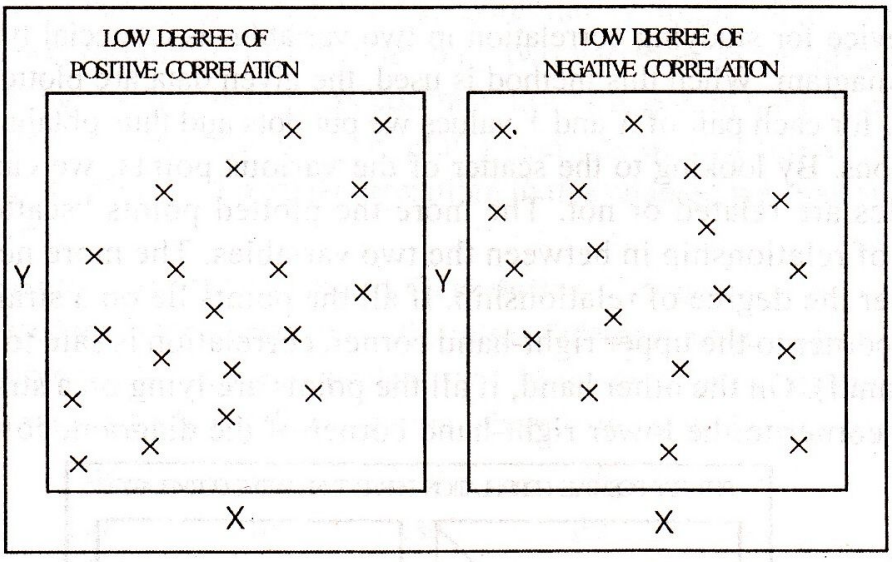


perfectly negative (*i.e.*, $r = -1$) (diagram II). If the plotted points fall in a narrow band, there would be a high degree of correlation between the variables—correlation shall be positive if the points show a rising tendency from the lower left-hand corner to the upper right-hand corner (diagram III) and negative if the points show a declining tendency from upper left-hand corner to the lower right-hand corner of the diagram (diagram IV). On the other hand, if the points are widely

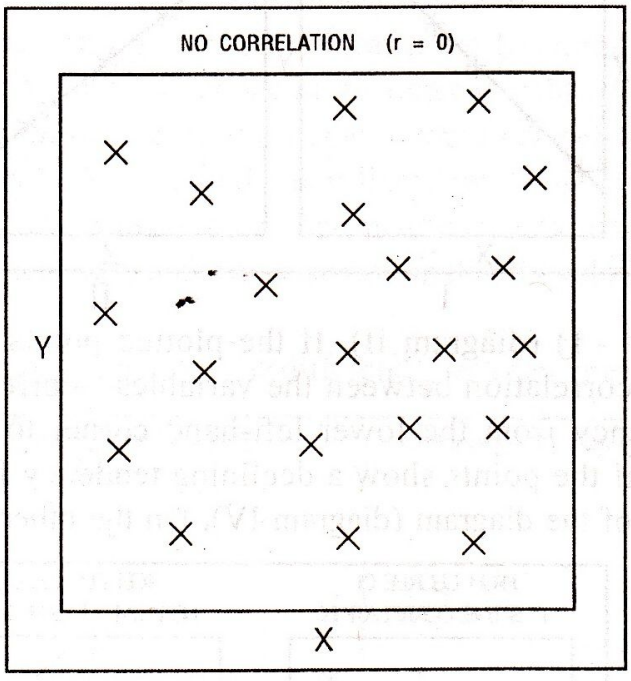


*This method is discussed in detail in Chapter on 'Regression Analysis'.

scattered over the diagrams it indicates very low degree of relationship between the variables—correlation shall be positive if the points are rising from the lower left-hand corner to the upper right-hand corner (diagram V) and negative if the points are running from the upper left-hand side to the lower right-hand side to the diagram (diagram VI). If the plotted points lie on a straight line parallel to the X-axis, or in a haphazard manner, it shows the absence of any relationship between the variables (*i.e.*, $r = 0$) as shown by diagram VII.



V VI



VII

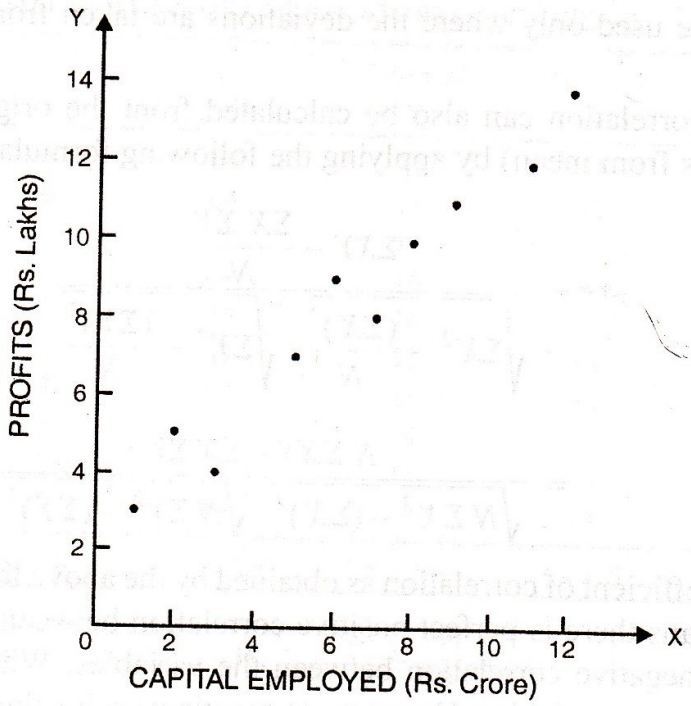
Illustration 1. Given the following pairs of values :

Capital employed (Rs. Crore) :	1	2	3	4	5	7	8	9	11	12
Profits (Rs. Lakhs) :	3	5	4	7	9	8	10	11	12	14

(*a*) Make a scatter diagram.

(*b*) Do you think that there is any correlation between profits and capital employed ? Is it positive ? Is it high or low ?

Solution. By looking at the scatter diagram we can say that the variables : profits and capital employed are correlated. Further, correlation is positive because the trend to the points is upward rising from the lower left-hand corner to the upper right-hand corner of the diagram. The diagram also indicates that the degree of relationship is high because the plotted points are in a narrow band which shows that it is a case of high degree of positive correlation.



Merits and Limitations of the Method

Merits : 1. It is a simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.

2. It is not influenced by the size of extreme values whereas most of the mathematical methods of finding correlation are influenced by extreme values.

3. Making a scatter diagram usually is the first step in investigating the relationship between the variables.

Limitations. By applying this method we can get an idea about the direction of correlation and also whether it is high or low. But we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical method.

II. KARL PEARSON'S COEFFICIENT OF CORRELATION

Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely used in practice. The coefficient of correlation is denoted by the symbol r. It is one of the very few symbols that is used universally for describing the degree and direction of relationship between two variables. If the two variables under study are X and Y, the following formula suggested by Karl Pearson can be used for measuring the degree of relationship.

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\sqrt{\Sigma(Y - \bar{Y})^2}} \quad ...(i)$$

where $\bar{X}$ and $\bar{Y}$ are the respective means of X and Y variable.

The above formula can be written as :

$$r^* = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \quad ...(ii)$$

where $x = (X - \bar{X})$ and $y = (Y - \bar{Y})$.

This formula is to be used only where the deviations are taken from *actual* means and *not* from assumed means.

The coefficient of correlation can also be calculated from the original set of observations (*i.e.*, without taking deviations from mean) by applying the following formula :

$$r^{**} = \frac{\Sigma XY - \frac{\Sigma X\,\Sigma Y}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}\ \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

$$= \frac{N\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} \qquad ...(iii)$$

The value of the coefficient of correlation as obtained by the above formula shall always lie between ± 1. When $r = +1$, it means there is perfect positive correlation between the variables. When $r = -1$, it means there is perfect negative correlation between the variables. When $r = 0$, it means there is no relationship between the two variables. However, in practice, such value of r as +1, –1, and 0 are rare. We normally get values which lie between +1 and –1 such as 0.8, –0.4, etc. The coefficient of correlation describes not only the magnitude of correlation but also its direction. Thus, +0.8 would mean that correlation is positive because the sign of r is +ve and the magnitude of correlation is 0.8.

The following illustration will clarify the procedure of computing the coefficient of correlation :

Illustration 2. Find correlation coefficient between the sales and expenses from the data given below :

Firm	:	1	2	3	4	5	6	7	8	9	10
Sales (Rs. Lakhs)	:	50	50	55	60	65	65	65	60	60	50
Expenses (Rs. Lakhs)	:	11	13	14	16	16	15	15	14	13	13

*The coefficient of correlation can also be expressed in terms of covariance and variance as given below :
From (*ii*), we have

$$r = \frac{\Sigma xy / N}{\sqrt{\Sigma x^2 / N}\ \sqrt{\Sigma y^2 / N}} = \frac{\text{Cov}\,[x, y]}{\sqrt{\text{Var}\,x,\ \text{Var}\,y}} = \frac{\text{Cov}\,[x, y]}{\sigma_x\,\sigma_y}$$

**This formula is derived from formula (*i*) as follows :

$$r = \frac{\Sigma(X - \bar{X})\,\Sigma(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\ \sqrt{\Sigma(Y - \bar{Y})^2}}$$

Opening the brackets, we get :

$$r = \frac{\Sigma XY - N\bar{X}\,\bar{Y}}{\sqrt{\Sigma X^2 - N\bar{X}^2}\ \sqrt{\Sigma Y^2 - N\bar{Y}^2}}$$

$$= \frac{\Sigma XY - \frac{\Sigma X.\Sigma Y}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}\ \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

$$= \frac{N\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

Solution. CALCULATION OF CORRELATION COEFFICIENT

Firm	*Sales* X	$(X-\bar{X})$ x	x^2	*Expenses* Y	$(Y-\bar{Y})$ y	y^2	xy
1	50	–8	64	11	–3	9	+24
2	50	–8	64	13	–1	1	+8
3	55	–3	9	14	0	0	0
4	60	+2	4	16	+2	4	+4
5	65	+7	49	16	+2	4	+14
6	65	+7	49	15	+1	1	+7
7	65	+7	49	15	+1	1	+7
8	60	+2	4	14	0	0	0
9	60	+2	4	13	–1	1	–2
10	50	–8	64	13	–1	1	+8
$N=10$	$\Sigma X=580$	$\Sigma x=0$	$\Sigma x^2=360$	$\Sigma Y=140$	$\Sigma y=0$	$\Sigma y^2=22$	$\Sigma xy=70$

$$\bar{X}=\frac{\Sigma X}{N}=\frac{580}{10}=58;\ \bar{Y}=\frac{\Sigma Y}{N}=\frac{140}{10}=14$$

$$r=\frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}=\frac{70}{\sqrt{360\times 22}}=\frac{70}{88.994}=0.787$$

To simplify calculation we can solve the question with the help of logarithms also.
Taking logarithms

$$\log r = \log 70 - \tfrac{1}{2}(\log 360 + \log 22)$$
$$= 1.8451 - \tfrac{1}{2}(2.5563 + 1.3424)$$
$$= 1.8451 - \tfrac{1}{2}(3.8987) = 1.8451 - 1.9493 = \bar{1}.8958$$
$$r = AL\ \bar{1}.8958 = 0.787$$

Hence, there is a high degree of positive correlation between the two variables *i.e.*, as the value of sales goes up, the expenses also go up.

When Deviations are taken from an Assumed Mean

When actual means are in fractions, say the actual means of X and Y series are 20.167 and 29.23, the calculation of coefficient of correlation by the method discussed above would involve too many calculations and would take a lot of time. In such cases we make use of the assumed mean method for finding out coefficient of correlation. When deviations are taken from an assumed mean, the following is applicable :

$$r=\frac{N\Sigma d_x d_y-(\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2-(\Sigma d_x)^2}\sqrt{N\Sigma d_y^2-(\Sigma d_y)^2}} \quad \ldots (iv)$$

where d_x refers to deviations of X series from an assumed mean, *i.e.*, $(X–A)$.

Similarly, d_y refers to deviation of Y series from an assumed mean *i.e.*, $(Y–A)$.

It may be noted that this form of formula is same as (*iii*), only difference being that whereas in form (*iii*) we are dealing with original X and Y, in form (*iv*) we are taking deviations of X and Y series from assumed mean.

The following example shall illustrate the application of this formula :

Illustration 3. The following data relate to the age of 10 employees and the number of days which they reported sick in a month :

Age	:	20	30	32	35	40	46	52	55	58	62
Sick days	:	11	12	10	13	14	16	15	17	18	19

Calculate Karl Pearson's coefficient of correlation and interpret its value.

(*MBA, DU, 2003*)

Solution. Let age and sick days be represented by variable X and Y respectively.

CALCULATION OF CORRELATION COEFFICIENT

Age X	$(X-43)$ d_x	d_x^2	*Sick days* Y	$(Y-14)$ d_y	d_y^2	$d_x d_y$
20	–23	529	11	– 3	9	+ 69
30	– 13	169	12	– 2	4	+ 26
32	– 11	121	10	– 4	16	+ 44
35	– 8	64	13	– 1	1	+ 8
40	– 3	9	14	0	0	0
46	+ 3	9	16	+ 2	4	+ 6
52	+ 9	81	15	+ 1	1	+ 9
55	+ 12	144	17	+ 3	9	+ 36
58	+ 15	225	18	+ 4	16	+ 60
62	+ 19	361	19	+ 5	25	+ 95
$\Sigma X = 430$	$\Sigma d_x = 0$	$\Sigma d_x^2 = 1712$	$\Sigma Y = 145$	$\Sigma d_y = 5$	$\Sigma d_y^2 = 85$	$\Sigma d_x d_y = 353$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{10\times 353 - (0)(-5)}{\sqrt{10\times 1712 - (0)^2}\sqrt{10\times 85 - (-5)^2}}$$

$$= \frac{3530}{\sqrt{17120}\sqrt{825}} = \frac{3530}{130.85\times 28.72} = 0.939$$

Thus, there is a very high degree of positive correlation between age and sick days taken. Hence, we can conclude that as the age of an employee increases, he is liable to be sick more often than others.

Illustration 4. Find the coefficient of correlation by Karl Pearson's method between X and Y and interpret its value.

X :	57	42	40	33	42	45	42	44	40	56	44	43
Y :	10	60	30	41	29	27	27	19	18	19	31	29

(MBA, M.D. Univ., 2001)

Solution. CALCULATION OF KARL PEARSON'S CORRELATION COEFFICIENT

X	$(X-44)$ d_x	d_x^2	Y	$(Y-30)$ d_y	d_y^2	$d_x d_y$
57	+ 13	169	10	– 20	400	– 260
42	– 2	4	60	+ 30	900	– 60
40	– 4	16	30	0	0	0
33	– 11	121	41	+ 11	121	– 121
42	– 2	4	29	– 1	1	+ 2
45	+ 1	1	27	– 3	9	– 3
42	– 2	4	27	– 3	9	+ 6
44	0	0	19	– 11	121	0
40	– 4	16	18	– 12	144	+ 48
56	+ 12	144	19	– 11	121	– 132
44	0	0	31	+ 1	1	0
43	– 1	1	29	– 1	1	+ 1
$\Sigma X = 528$	$\Sigma d_x = 0$	$\Sigma d_x^2 = 480$	$\Sigma Y = 340$	$\Sigma d_y = -20$	$\Sigma d_y^2 = 1828$	$\Sigma d_x d_y = -519$

$$r = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{12(-519) - (0)(-20)}{\sqrt{12(480) - (0)^2}\sqrt{12(1828) - (-20)^2}}$$

$$= \frac{-6228}{\sqrt{5760}\sqrt{21536}} = \frac{-6228}{11137.66} = -0.559.$$

Therefore, it is a case of moderate degree of negative correlation.

Correlation of Bivariate Grouped Data

When we have to find coefficient of correlation from a bivariate grouped data table, the following formula is applicable :

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

This formula is the same as that of *(iv)*. The only difference is that here the deviations are also multiplied by the frequencies.

The following illustration shall explain the application of this formula :

Illustration 5. Find the coefficient of correlation between the age and the sum assured from the following table :

Sum assured (in Rs.)

Age group	10,000	20,000	30,000	40,000	50,000
20–30	4	6	3	7	1
30–40	2	8	15	7	1
40–50	3	9	12	6	2
50–60	8	4	2	—	—

(MBA, Delhi Univ., 1999)

Solution. Let the sum assured be denoted by X and the age group by Y.

CALCULATION OF COEFFICIENT OF CORRELATION

Y \ X			10,000	20,000	30,000	40,000	50,000				
		d_x	−2	−1	0	1	2	f	fd_y	fd_y^2	fd_xd_y
20–30	m.p. 25	−2	16 / 4	12 / 6	0 / 3	−14 / 7	−4 / 1	21	−42	84	+10
30–40	35	−1	4 / 2	8 / 8	0 / 15	−7 / 7	−2 / 1	33	−33	33	+3
40–50	45	0	0 / 3	0 / 9	0 / 12	0 / 6	0 / 2	32	0	0	0
50–60	55	+1	−16 / 8	−4 / 4	0 / 2	—	—	14	+14	14	−20
		f	17	27	32	20	4	$N = 100$	$\Sigma fd_y = -61$	$\Sigma fd_y^2 = 131$	$\Sigma fd_xd_y = -7$
		fd_x	−34	−27	0	20	−8	$\Sigma fd_x = -33$			
		fd_x^2	68	27	0	20	16	$\Sigma fd_x^2 = 131$			
		fd_xd_y	4	16	0	−21	−6	$\Sigma fd_xd_y = -7$			

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{100(-7) - (-33)(-61)}{\sqrt{100(131) - (-33)^2}\sqrt{100(131) - (-61)^2}}$$

$$r = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}} = \frac{12(-519) - (0)(-20)}{\sqrt{12(480) - (0)^2}\sqrt{12(1828) - (-20)^2}}$$

$$= \frac{-6228}{\sqrt{5760}\sqrt{21536}} = \frac{-6228}{11137.66} = -0.559.$$

Therefore, it is a case of moderate degree of negative correlation.

Correlation of Bivariate Grouped Data

When we have to find coefficient of correlation from a bivariate grouped data table, the following formula is applicable :

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

This formula is the same as that of (*iv*). The only difference is that here the deviations are also multiplied by the frequencies.

The following illustration shall explain the application of this formula :

Illustration 5. Find the coefficient of correlation between the age and the sum assured from the following table :

	Sum assured (in Rs.)				
Age group	10,000	20,000	30,000	40,000	50,000
20–30	4	6	3	7	1
30–40	2	8	15	7	1
40–50	3	9	12	6	2
50–60	8	4	2	—	—

(*MBA, Delhi Univ., 1999*)

Solution. Let the sum assured be denoted by X and the age group by Y.

CALCULATION OF COEFFICIENT OF CORRELATION

Y \ X			10,000	20,000	30,000	40,000	50,000				
		d_x / d_y	−2	−1	0	1	2	f	fd_y	fd_y^2	fd_xd_y
20–30	m.p. 25	−2	16 / 4	12 / 6	0 / 3	−14 / 7	−4 / 1	21	−42	84	+10
30–40	35	−1	4 / 2	8 / 8	0 / 15	−7 / 7	−2 / 1	33	−33	33	+3
40–50	45	0	0 / 3	0 / 9	0 / 12	0 / 6	0 / 2	32	0	0	0
50–60	55	+1	−16 / 8	−4 / 4	0 / 2	—	—	14	+14	14	−20
		f	17	27	32	20	4	$N = 100$	$\Sigma fd_y = -61$	$\Sigma fd_y^2 = 131$	$\Sigma fd_xd_y = -7$
		fd_x	−34	−27	0	20	−8	$\Sigma fd_x = -33$			
		fd_x^2	68	27	0	20	16	$\Sigma fd_x^2 = 131$			
		fd_xd_y	4	16	0	−21	−6	$\Sigma fd_xd_y = -7$			

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{100(-7) - (-33)(-61)}{\sqrt{100(131) - (-33)^2}\sqrt{100(131) - (-61)^2}}$$

$$= \frac{-700-2013}{\sqrt{13100-1089}\sqrt{13100-3721}} = \frac{-2713}{\sqrt{12011}\sqrt{9379}}$$

$$= \frac{-2713}{109.59 \times 96.85} = -0.256.$$

Hence the age and sum assured are negatively correlated, *i.e.*, as age goes up the sum assured comes down.

Illustration 6. Calculate the coefficient of correlation from the following bivariate frequency distribution :

Sales Revenue	*Advertising Expenditure (Rs. '000)*			
(Rs. lakhs)	*5–10*	*10–15*	*15–20*	*20–25*
75–125	4	1	—	—
125–175	7	6	2	1
175–225	1	3	4	2
225–275	1	1	3	4

Solution. Let sales revenue be denoted by Y and advertising expenditure by X.

CALCULATION OF COEFFICIENT OF CORRELATION

Y \ X		*m.p.*	5–10 7.5	10–15 12.5	15–20 17.5	20–25 22.5				
		d_x / d_y	−1	0	+1	+2	f	fd_y	fd_y^2	fd_xd_y
75–125	*m.p.* 100	−2	8 (4)	0 (1)	—	—	5	−10	20	8
125–175	150	−1	7 (7)	0 (6)	−2 (2)	−2 (1)	16	−16	16	3
175–225	200	0	0 (1)	0 (3)	0 (4)	0 (2)	10	0	0	0
225–275	250	+1	−1 (1)	0 (1)	3 (3)	8 (4)	9	9	9	10
		f	13	11	9	7	N = 40	$\Sigma fd_y = -17$	$\Sigma fd_y^2 = 45$	$\Sigma fd_xd_y = 21$
		fd_x	−13	0	+9	+14	$\Sigma fd_x = 10$			
		fd_x^2	13	0	9	28	$\Sigma fd_x^2 = 50$			
		fd_xd_y	14	0	1	6	$\Sigma fd_xd_y = 21$			

$$r = \frac{N\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{40x21 - 10(-17)}{\sqrt{40x50 - (10)^2}\sqrt{40 \times 45 - (-17)^2}}$$

$$= \frac{840+170}{\sqrt{1900 \times 1511}} = \frac{1010}{1694.373} = 0.596$$

There is a moderate degree of positive correlation between sales revenue and advertising expenditure.

Assumptions of the Pearsonians Coefficient

The Karl Pearson's coefficient of correlation is based on the following assumptions :

1. There is linear relationship between the variables, *i.e.*, when the two variables are plotted on a scatter diagram, a straight line will be formed by the points so plotted.

2. The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc., are affected by such forces that a normal distribution is formed.

3. There is a cause-and-effect relationship between the forces affecting the distribution of the items in the two series. If such a relationship is not formed between the variables, *i.e.,* if the variables are independent there cannot be any correlation. For example, there is no relationship between income and height because the forces that affected these variables are common.

Properties of the Coefficient of Correlation

The following are the important properties of the coefficient of correlations, r :

1. The coefficient of correlation lies between –1 and +1. Symbolically, $-1 \le r \le +1$ or $|r| \le 1$.

Proof. $$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2}\sqrt{\Sigma (Y - \bar{Y})^2}}$$

Let $$a = \frac{(X - \bar{X})}{\sqrt{\Sigma (X - \bar{X})^2}}, \quad b = \frac{(Y - \bar{Y})}{\sqrt{\Sigma (Y - \bar{Y})}}$$

Then $$\Sigma (a + b)^2 = \Sigma a^2 + 2\Sigma ab + \Sigma b^2$$

$$= 1 + 2r + 1 = 2(1 + r) \ge 0 \quad \text{or} \quad 1 + r \ge 0 \qquad ...(i)$$

Similarly, $$\Sigma (a - b)^2 = \Sigma a^2 - 2\Sigma ab + \Sigma b^2$$

$$= 1 - 2r + 1 = 2(1 - r) \ge 0 \quad \text{or} \quad 1 - r \ge 0 \qquad ...(ii)$$

From (i) and (ii), $-1 \le r \le 1$.

2. The coefficient of correlation is independent of change of origin and scale.

Proof. By change of origin we mean subtracting some constant from the given value of X and Y and by change of scale we mean dividing or multiplying every value of X and Y by some constant.

We know that $$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2 \; \Sigma (Y - \bar{Y})^2}} \qquad ...(i)$$

where $\bar{X}$ and $\bar{Y}$ refer to the acutal means of X and Y series.

Let us now change the origin and scale. Deduct a fixed quantity a from X and b from Y. Also divide X and Y series by a fixed value i and c.

Let the new values be denoted by u and v.

$$u = \frac{X - a}{i} \qquad v = \frac{Y - b}{c}$$

$$X = a + iu, \qquad Y = b + cv$$

$$\bar{X} = a + i\bar{u}, \qquad \bar{Y} = b + c\bar{v}$$

$$X - \bar{X} = i(u - \bar{u}) \qquad Y - \bar{Y} = c(v - \bar{v})$$

Substituting these values in (i), we get

$$\frac{\Sigma (u - \bar{u})(v - \bar{v})}{\sqrt{\Sigma (u - \bar{u})^2}\sqrt{\Sigma (v - \bar{v})^2}}$$

Thus the formula for r remains unchanged. Hence the value of r is independent of change of origin and scale.

3. The coefficient of correlation is the geometric mean of two regression coefficients.*

Symbolically : $r = \sqrt{b_{xy} \times b_{yx}}$

4. If X and Y are independent variables then coefficient of correlation is zero. However, the converse is not true.

Interpreting the Coefficient of Correlation

The coefficient of correlation measures the degree of relationship between two sets of figures. As the reliability of estimates depends upon the closeness of the relationship it is imperative that utmost care be taken while interpreting the value of coefficient of correlation otherwise fallacious conclusions be drawn.

Unfortunately, the interpretation of the coefficient of correlation depends very much on experience. The full significance of r will only be grasped after working out a number of correlation problems and seeing the kind of data that give rise to various values of r. The investigator must know his data thoroughly in order to avoid errors of interpretation. He must be familiar, or become familiar, with all the relationships and theory which bear upon the data and should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. However, the following general guidelines are given which would help in interpreting the value of r.

1. When $r = +1$, it means there is perfect positive correlation between the variables.
2. When $r = -1$, it means there is perfect negative correlation between the variables.
3. When $r = 0$, it means there is no correlation between the variables, *i.e.*, the variables are uncorrelated.
4. The closer r is to +1 or –1, the closer the relationship between the variables and the closer r is to 0, the less closer the relationship. Beyond this is not safe to go. The full interpretation of r depends upon circumstances, one of which is the size of the sample. All that can really be said that when estimating the value of one variable from the value of another; the higher the value of r, the better the estimate.
5. The closeness of the relationship is not proportional to r. If the value of r is 0.8, it does not indicate a relationship twice as close as that of 0.4. It is in fact very much closer.

Coefficient of Correlation and Probable Error

The probable error of the coefficient of correlation helps in interpreting its value. With the help of probable error it is possible to determine the liability of the value of the coefficient in so far as it depends on the condition of random sampling. The probable error of the coefficient of correlation is obtained as follows :

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

where r is the coefficient of correlation and N the number of pairs of items.

1. If the value of r is less than the probable error, there is no evidence of correlation, *i.e.*, the value of r is not at all significant.

---

*See chapter on Regression Analysis.

*If 0.6745 is omitted from the formula of probable error, we get the standard error from the coefficient of correlation. The standard error of r, therefore, is

$$\text{S.E.}r = \frac{1 - r^2}{\sqrt{N}}$$

2. If the value of r is more than six times the probable error, the existence of correlation is practically certain, *i.e.*, the value of r is significant.

3. By adding and subtracting the value of probable error from the coefficient of correlation we get respectively the upper and lower limits within which coefficient of correlation in the population can be expected to lie. Symbolically,

$$\rho = r \pm \text{P.E.}r$$

where ρ (rho) denotes correlation in the population.

Carrying out the computation of the probable error, assuming a coefficient of correlation of 0.80 computed from a sample of 16 pairs of items, we have

$$\text{P.E.}r = 0.6745 \frac{1-(0.8)^2}{\sqrt{16}} = 0.06$$

The limits of the correlation in the population should be $r \pm \text{P.E.} = 0.8 + 0.06 = 0.74 - 0.86$.

Instances are quite common wherein a correlation coefficient of 0.5 or even 0.4 is obviously considered to be a fairly high degree of correlation by a research worker. Yet a correlation coefficient of 0.5 means that only 25 per cent of the variation is explained. A correlation coefficient of 0.4 means that only 16 per cent of the variation is explained.

Conditions for the Use of Probable Error

The measure of probable error can be properly used only when the following three conditions exist :

1. The data must approximate to a normal frequency curve (bell-shaped curve).
2. The statistical measure for which the P.E. is computed must have been calculated from a sample.
3. The sample must have been selected in an unbiased manner and the individual items must be independent.

However, these conditions are generally not satisfied and as such the reliability of the correlation coefficient is determined largely on the basis of exterior tests of reasonableness which are often of a statistical character.

Illustration 7. If $r = 0.6$ and $N = 64$, find out the probable error of the coefficient of correlation and determine the limits for r.

Solution :

$$\text{P.E.}r = 0.6745 \frac{1-r^2}{\sqrt{N}}; \qquad r = 0.6 \text{ and } N = 64$$

$$\text{P.E.}r = 0.6745 \frac{1-(0.6)^2}{\sqrt{64}} = \frac{0.6745 \times 0.64}{8} = 0.054$$

limits of $r = 0.6 \pm 0.054$ or $= 0.546$ to 0.654.

Merits and Limitations of the Pearsonian Coefficient

Amongst the mathematical methods used for measuring the degree of relationship, Karl Pearson's method is most popular. The correlation coefficient summarizes in one figure not only the degree of correlation but also the degree, *i.e.*, whether correlation is positive or negative.

However, the utility of the coefficient depends in part on a wide knowledge of the meaning of this 'yardstick' together with its limitations. The chief *limitations* of the method are :

1. The correlation coefficient always assumes linear relationship regardless of the fact whether that assumption is true or not.

2. Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted.

3. The value of the coefficient is unduly affected by the extreme values.

4. As compared to other methods of finding correlation, this method is more time-consuming.

Coefficient of Determination*

One very convenient and useful way of interpreting the value of coefficient of correlation between two variables is to use the square of coefficient of correlation, which is called coefficient of determination. The coefficient of determination thus equals r^2. The coefficient, r^2 expresses the proportion of the variance in Y determined in X, that is, the ratio of the explained variance to the total variance.

Therefore, the coefficient of determination expresses the proportion of the total variation that has been "explained", or the relative reduction in variance when measured about the regression equation rather than about the mean of the dependent variable. If the value of $r = 0.9$, r^2 will be 0.81 and this would mean that 81 per cent of the variation in the dependent variable has been explained by the independent variable. The maximum value of r^2 is unity because it is possible to explain all of the variation in Y, but it is not possible to explain more than all of it.

It is much easier to understand the meaning of r^2 and r and, therefore, the coefficient of determination is to be preferred in presenting the result of correlation analysis. Tuttle has beautifully pointed out that "the coefficient of correlation has been grossly overrated and is used entirely too much. Its square coefficient of determination is a much more useful measure of the linear covariation of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before coming to any conclusion about the extent of the linear relationship between the two correlated variables."

The relationship between r and r^2 may be noted—as the value of r decreases from its maximum value of 1, the value of r^2 decreases much more rapidly. r will of course always be larger than r^2, unless $r^2 = 0$ or 1, when $r = r^2$,

r	r^2	r	r^2
0.90	0.81	0.60	0.36
0.80	0.64	0.50	0.25
0.70	0.49	0.40	0.16

Thus the coefficient of correlation is 0.707 when just half the variance in Y is due to X.

It should be clearly noted that the fact that a correlation between two variables has a value of $r = 0.60$ and the correlation between two other variables has a value of $r = 0.30$ does not demonstrate that the first correlation is twice as strong as the second. The relationship between the two given values of r can better be understood by computing the value of r^2. When $r = 0.6$, $r^2 = 0.36$ and when $r = 0.30$, $r^2 = 0.09$.

The coefficient of determination is a highly useful measure. However, it is often misinterpreted. The term itself may be misleading in that it implies that the variable X stands in a determining of casual relationship of the variable Y. The statistical evidence itself never establishes the existence of such causality. All that statistical evidence can do it to define covariation, that term being used in a perfectly

*$1 - r^2$ is known as the coefficient of non-determination.

neutral sense. Whether causality is present or not and which way it runs if it is present, must be determined on the basis of evidence other than the quantitative observations.

III. RANK CORRELATION COEFFICIENT

This method of finding out covariability or the lack of it between two variables was developed by the British psychologist Charles, Edward Spearman in 1904. This measure is especially useful when quantitative measure of certain factors (such as in the evaluation of leadership ability or the judgement of female beauty) cannot be fixed, but the individuals in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank* in the group. In any event, the rank correlation coefficient is applied to a set of ordinal rank numbers, with 1 for the individual ranked first in quantity or quality, and so on, N for the individual ranked last in a group of N individuals (or N pairs of individuals). Spearman's rank correlation coefficient is defined as :

$$R = 1 - \frac{6\Sigma D^2}{N\left(N^2 - 1\right)} \text{ or } 1 - \frac{6\Sigma D^2}{\left(N^3 - N\right)}$$

where R denotes rank coefficient of correlation and D refers to the difference of ranks between paired items in two series.

The value of this coefficient also lies between +1 and –1. When R is +1, there is complete agreement in the order of the ranks and the ranks are in the same direction. When R is –1, there is complete agreement in the order of the ranks and they are in opposite directions. This shall be clear from the following :

R_1	R_2	D $(R_1 - R_2)$	D^2	R_1	R_2	D $(R_1 - R_2)$	D^2
1	1	0	0	1	3	–2	4
2	2	0	0	2	2	0	0
3	3	0	0	3	1	2	4
			$\Sigma D^2 = 0$				$\Sigma D^2 = 8$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 0}{3^3 - 3} = 1 - 0 = 1 \qquad R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 8}{3^3 - 3} = 1 - 2 = -1$$

In rank correlation we may have two types of problems :

A. Where actual ranks are given.

B. Where ranks are not given.

A. Where Actual Ranks are Given.

Where actual ranks are given, the steps required for computing rank correlation are :

(*i*) Take the differences of the two ranks, *i.e.*, $(R_1 - R_2)$ and denote these differences by D.

(*ii*) Square these differences and obtain the total ΣD^2.

(*iii*) Apply the formula :

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

*The rank transformation for a sample of n observation replaces the smallest observation by the integer 1 (called) the rank, the next by rank 2 and so on until the largest observation is replaced by rank n.

Illustration 7. Two managers are asked to rank a group of employees in order of potential for eventually becoming top managers. The rankings are as follows :

Employees	*Ranking by Manager I*	*Ranking by Manager II*
A	10	9
B	2	4
C	1	2
D	4	3
E	3	1
F	6	5
G	5	6
H	8	8
I	7	7
J	9	10

Compute the coefficient of rank correlation and comment on the value.

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

Employees	*Rank by Manager I* R_1	*Rank by Manager II* R_2	$(R_1 - R_2)^2$ D^2
A	10	9	1
B	2	4	4
C	1	2	1
D	4	3	1
E	3	1	4
F	6	5	1
G	5	6	1
H	8	8	0
I	7	7	0
J	9	10	1
$N = 10$			$\Sigma D^2 = 14$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 14}{990} = 1 - 0.085 = 0.915$$

Thus we find that there is a high degree of positive correlation in the ranks assigned by the two managers.

Illustration 8. Two housewives, Geeta and Rita, asked to express their preference for different kinds of detergents, gave the following replies :

Detergent	*Geeta*	*Rita*
A	4	4
B	2	1
C	1	2
D	3	3
E	7	8
F	8	7
G	6	5
H	5	6
I	9	9
J	10	10

To what extent the preferences of these two ladies go together?

Solution. In order to find out how far the preferences for different kinds of detergents go together, we will calculate rank correlation coefficient.

CALCULATION OF RANK CORRELATION COEFFICIENT

Detergent	*Rank by Geeta* R_1	*Rank by Rita* R_2	$(R_1 - R_2)^2$ D^2
A	4	4	0
B	2	1	1
C	1	2	1
D	3	3	0
E	7	8	1
F	8	7	1
G	6	5	1
H	5	6	1
I	9	9	0
J	10	10	0
$N = 10$			$\Sigma D^2 = 6$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 6}{990} = 1 - 0.036 = 0.964.$$

Thus the preferences of these two ladies agree very closely as far as their opinion on detergents is concerned.

B. Where Ranks are not Given.

When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value, we must follow the same method in case of all the variables.

Illustration 9. Calculate the rank correlation coefficient for the following data of marks of 2 tests given to candidates for a clerical job.

Preliminary test :	92	89	87	86	83	77	71	63	53	50
Final test :	86	83	91	77	68	85	52	82	37	57

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

Preliminary test X	R_1	*Final test* Y	R_2	$(R_1 - R_2)^2$ D^2
92	10	86	9	1
89	9	83	7	4
87	8	91	10	4
86	7	77	5	4
83	6	68	4	4
77	5	85	8	9
71	4	52	2	4
63	3	82	6	9
53	2	37	1	1
50	1	57	3	4
$N = 10$				$\Sigma D^2 = 44$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 44}{990} = 1 - 0.267 = 0.733.$$

Thus, there is a high degree of positive correlation between preliminary and final test.

Equal Ranks or Tie in Ranks

In some cases it may be found necessary to assign equal rank to two or more individuals or entries. In such a case, it is customary to give each individual or entry an average rank. Thus if two individuals are ranked equal at fifth place, they are each given the rank $\frac{5+6}{2}$, that is 5.5 while if

three are ranked equal at fifth place, they are given the rank $\frac{5+6+7}{3}=6$. In other words, where two or more individuals are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of the ranks which these individuals would have got had they differed slightly from each other.

Where equal ranks are assigned to some entries, an adjustment in the above formula for calculating the rank coefficient of correlation is made.

The adjustment consists of adding $\frac{1}{12}(m^3-m)$ to the value of ΣD^2, where m stands for the number of items whose ranks are common. If there are more than one such group of items with common rank, this value is added as many times as the number of such groups. The formula can thus be written as :

$$R=1-\frac{6\{\Sigma D^2+\frac{1}{12}(m_1^3-m_1)+\frac{1}{12}(m_2^3-m_2)+.....\}}{N^3-N}$$

Illustration 10. An examination of eight applicants for a clerical post was taken by a firm. From the marks obtained by the applicants in the Accountancy and Statistics papers, compute rank coefficient of correlation.

Applicant	:	A	B	C	D	E	F	G	H
Marks in Accountancy	:	15	20	28	12	40	60	20	80
Marks in Statistics	:	40	30	50	30	20	10	30	60

(MBA, Delhi Univ., 2009)

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

Applicants	*Marks in Accountancy* X	*Rank assigned* R_1	*Marks in Statistics* Y	*Rank assigned* R_2	$(R_1-R_2)^2$ D^2
A	15	2	40	6	16.00
B	20	3.5	30	4	0.25
C	28	5	50	7	4.00
D	12	1	30	4	9.00
E	40	6	20	2	16.00
F	60	7	10	1	36.00
G	20	3.5	30	4	0.25
H	80	8	60	8	0.00
$N=8$					$\Sigma D^2=81.5$

$$R=1-\frac{6\{\Sigma D^2+\frac{1}{12}(m_1^3-m_1)+\frac{1}{12}(m_2^3-m_2)\}}{N^3-N}$$

The item 20 is repeated 2 times in series X and hence $m_1=2$. In series Y, the item 30 occurs 3 times and hence $m_2=3$. Substituting these values in the above formula :

$$R=1-\frac{6\left\{81.5+\frac{1}{12}(2^3-2)+\frac{1}{12}(3^3-3)\right\}}{8^3-8}$$

$$=1-\frac{6(81.5+0.5+2)}{504}=1-\frac{6\times 84}{504}=0$$

There is no correlation between the marks obtained in the two subjects.

Illustration 11. Ten competitors in a beauty contest are ranked by three judges in the following order :

1st Judge	:	1	6	5	10	3	2	4	9	7	8
2nd Judge	:	3	5	8	4	7	10	2	1	6	9
3rd Judge	:	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty.

(MBA, M.D. Univ., 1996; MBA, Bharthidasan Univ., 2001)

Solution : In order to find out which pair of judges has the nearest approach to common tastes in beauty, we compare rank correlation between the judgement of

(*i*) 1st judge and 2nd judge.
(*ii*) 2nd judge and 3rd judge.
(*iii*) 1st judge and 3rd judge.

Rank by 1st Judge R_1	*Rank by 2nd Judge* R_2	*Rank by 3rd Judge* R_3	$(R_1 - R_2)^2$ D^2	$(R_2 - R_3)^2$ D^2	$(R_1 - R_3)^2$ D^2
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
$N = 10$	$N = 10$	$N = 10$	$\Sigma D^2 = 200$	$\Sigma D^2 = 214$	$\Sigma D^2 = 60$

$$R(\text{I\&II}) = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 200}{10^3 - 10} = 1 - \frac{1200}{990} = -0.212$$

$$R(\text{II\&III}) = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 214}{10^3 - 10} = 1 - \frac{1284}{990} = -0.297$$

$$R(\text{I\&III}) = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{360}{990} = 0.636$$

Since coefficient of correlation is maximum in the judgment of the first and third judges, we conclude that they have the nearest approach to common tastes in beauty.

Merits and Limitations of the Rank Method

Merits. 1. This method is simpler to understand and easier to apply compared to the Karl Pearson's method. The answers obtained by this method and the Karl Pearson's method will be the same provided no value is repeated, *i.e.*, all the items are different.

2. Where the data are of a qualitative nature like honesty, efficiency, intelligence, etc., this method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and the degree of correlation established by applying the method.

3. This is the only method that can be used where we are given the ranks and not the actual data.

4. Even where actual data are given, rank method can be applied for ascertaining rough degree of correlation.

Limitations. 1. This method cannot be used for finding out correlation in a grouped frequency distribution.

2. Where the number of observations exceed 30, the calculations become quite tedious and require a lot of time. Therefore, this method should not be applied where N is exceeding 30 unless we are given the ranks and not the actual values of the variable.

When to Use Rank Correlation Coefficient

The rank method has two principal uses :

(1) The initial data are in the form of ranks.

(2) If N is fairly small (say, not greater than 25 or 30) rank method is sometimes applied to interval data as an approximation to the more time-consuming r. This requires that the interval data be transferred to rank orders for both variables. If N is much in excess of 30, the labour required in ranking the scores becomes greater than what is justified by the anticipated saving of time through the rank formula.

Illustration 12. The coefficient of rank correlation between debenture prices and share prices is found to be 0.143. If the sum of squares of the differences in rank is given to be 48, find the value of N.

Solution. $$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

where $$R = 0.143, \Sigma D^2 = 48$$

$$0.143 = 1 - \frac{6 \times 48}{N^3 - N}$$

$$\frac{288}{N^3 - N} = 0.857 \text{ or } 0.857 (N^3 - N) = 288$$

$$(N^3 - N) = \frac{288}{0.857} = 336$$

or $$N^3 - N - 336 = 0 \text{ or } N^3 - N - 343 + 7 = 0$$

$$(N - 7)(N^2 + 7N) + 48 (N - 7) = 0$$

or $$(N - 7)(N^2 + 7N + 48) = 0$$

either $$N - 7 = 0 \quad i.e., N = 7 \quad \text{or} \quad N^2 + 7N + 48 = 0$$

Since $b^2 - 4ac$ is negative, value of N belongs to the set of complex numbers. Hence $N = 7$.

Illustration 13. The coefficient of rank correlation of the marks obtained by 10 students in statistics and accountancy was found to be 0.8. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 7 instead of 9. Find the correct coefficient of rank correlation.

Solution. $$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$0.8 = 1 - \frac{6\Sigma D^2}{10^3 - 10} \text{ or } 0.8 = 1 - \frac{6\Sigma D^2}{990} \text{ or } \frac{6\Sigma D^2}{990} = 0.2$$

$$6\Sigma D^2 = 198 \quad \text{or} \quad \Sigma D^2 = 33$$

But this is not correct ΣD^2

$$\text{Correct } \Sigma D^2 = 33 - (7)^2 + (9)^2 = 65$$

$$R = 1 - \frac{6 \times 65}{990} = 1 - \frac{390}{990} = 1 - 0.394 = 0.606$$

Thus the correct value of the rank correlation coefficient is 0.606.

IV. METHOD OF LEAST SQUARES

For finding out correlation by the coefficient method of least squares we have to calculate the values of two regression coefficients—that of x on y and y on x. The correlation coefficient is the square root of the product of two regression coefficients. Symbolically,

$$r^* = \sqrt{b_{xy} \times b_{yx}}$$

Lag and Lead in Correlation

The study of lag and lead is of special significance while studying economic and business series. In the correlation of time series the investigator may find that there is a time gap before a cause-and-effect

*For details of this method refer to next chapter on Regression Analysis.

relationship is established. For example, the supply of a commodity may increase today, but it may not have an immediate effect on prices—it may take a few days or even months for prices to adjust to the increased supply. The difference in the period before a cause-and-effect relationship is established is called 'Lag'. While computing correlation this time gap must be considered; otherwise, fallacious conclusions may be drawn. The pairing of items is adjusted according to the time lag.

If the supply affects the prices, say, after 5 months, then the pairing would be done as follows :

Months	*Supply*	*Price*
Jan.	100	
Feb.	105	70
March	108	69
April	112	80
May	118	72
June	120	75
July	125	70
Aug.	104	74
Sept.	112	75
Oct.	116	78
Nov.	122	80
Dec.	127	78
		75

Taking the new pairs of values, correlation can be calculated in the same manner as discussed earlier.

Illustration 14. The following are the monthly figures of advertising expenditure and sales of a firm. It is generally found that advertising expenditure has its impact on sales generally after two months. Allowing for this time lag, calculate coefficient of correlation between expenditure on advertisement and sales.

Month	*Advertising expenditure*	*Sales (Rs.)*	*Month*	*Advertising expenditure*	*Sales (Rs.)*
Jan.	50	1,200	July	140	2,400
Feb.	60	1,500	Aug.	160	2,600
March	70	1,600	Sept.	170	2,800
April	90	2,000	Oct.	190	2,900
May	120	2,200	Nov.	200	3,100
June	150	2,500	Dec.	250	3,900

Solution. Allow for a time lag of 2 months, *i.e.*, link advertising expenditure of January with sales for March, and so on.

CALCULATION OF CORRELATION COEFFICIENT

Month	*Advertising expenditure* X	$(X-\bar{X})/10$ x	x^2	*Sales* Y	$(Y-\bar{Y})/100$ y	y^2	xy
Jan.	50	−7	49	1,600	−10	100	70
Feb.	60	−6	36	2,000	−6	36	36
March	70	−5	25	2,200	−4	16	20
April	90	−3	9	2,500	−1	1	3
May	120	0	0	2,400	−2	4	0
June	150	+3	9	2,600	0	0	0
July	140	+2	4	2,800	+2	4	4
Aug.	160	+4	16	2,900	+3	9	12
Sept.	170	+5	25	3,100	+5	25	25
Oct.	190	+7	49	3,900	+13	169	91
	$\Sigma X = 1,200$	$\Sigma x = 0$	$\Sigma x^2 = 222$	$\Sigma Y = 26,000$	$\Sigma y = 0$	$\Sigma y^2 = 364$	$\Sigma xy = 261$

$$\bar{X} = \frac{1,200}{10} = 120, \quad \bar{Y} = \frac{26,000}{10} = 2,600$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{261}{\sqrt{222 \times 364}} = \frac{261}{284.27} = 0.918$$

There is a very high degree of positive correlation between advertising expenditure and sales.

MISCELLANEOUS ILLUSTRATIONS

Illustration 15. A Computer while calculating the correlation coefficient between the variables X and Y obtained following results :

$$N = 30,\ \Sigma X = 120,\ \Sigma X^2 = 600,\ \Sigma Y = 90,\ \Sigma Y^2 = 250,\ \Sigma XY = 335$$

It was, however, later discovered at the time of checking that it had copied down two pairs of observations as:

X	Y
8	10
12	7

While the correct values were

X	Y
8	12
10	8

Obtain the correct value of the correlation coefficient between X and Y.

(MBA, Vikram, Univ.; MBA, Kumaun Univ., 2007)

Solution.

Correct $\Sigma X = 120 - 8 - 12 + 8 + 10 = 120 - 2 = 118$

Correct $\Sigma Y = 90 - 10 - 7 + 12 + 8 = 93$

Correct $\Sigma X^2 = 600 - (8)^2 - (12)^2 + (8)^2 + (10)^2 = 600 - 64 - 144 + 64 + 100 = 556$

Correct $\Sigma Y^2 = 250 - (10)^2 - (7)^2 + (12)^2 + (8)^2 = 250 - 100 - 49 + 144 + 64 = 309$

Correct $\Sigma XY = 335 - (8 \times 10) - (12 \times 7) + (8 \times 12) + (10 \times 8)$

$= 335 - 80 - 84 + 96 + 80 = 347$

$$r = \frac{N\Sigma XY - \Sigma X \Sigma Y}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{(30 \times 347) - (118 \times 93)}{\sqrt{(30 \times 556) - (118)^2}\sqrt{30 \times 309 - (93)^2}}$$

$$= \frac{10410 - 10974}{\sqrt{16680 - 13924}\sqrt{9270 - 8649}} = \frac{-564}{\sqrt{2756}\sqrt{621}} = \frac{-564}{52.50 \times 24.92} = -0.43$$

Thus the correct value of correlation coefficient between X and Y is -0.43.

Illustration 16. Coefficient of correlation between two variates X and Y is 0.3. Their covariance is 9. The variance of X is 16. Find the standard deviation of Y series.

Solution. Covariance is given by $\frac{\Sigma xy}{N}$, where x and y are the deviations of X and Y series from their respective means.

Variance of X series is 16, or $\sigma_x = \sqrt{16} = 4$

Substituting the given value in the formula $r = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$, we get

$0.3 = 9 \times \frac{1}{4\sigma_y}$ or $1.2\sigma_y = 9$. Hence $\sigma_y = 9/1.2 = 7.5$.

Illustration 17. Family income and its percentage spent on food in the case of one hundred families gave the following bivariate frequency distribution. Calculate the coefficient of correlation and interpret its value.

Food Expenditure	*Monthly Family Income (Rs. '000's)*				
(in%)	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30
10–15	—	—	—	3	7
15–20	—	4	9	4	3
20–25	7	6	12	5	—
25–30	3	10	19	8	—

(MBA, Delhi Univ., 2003)

Solution. Let family income be denoted by X and food expenditure (in %) by Y.

CALCULATION OF CORRELATION COEFFICIENT

Y \ X		m.p.	5–10 7.5	10–15 12.5	15–20 17.5	20–25 22.5	25–30 27.5				
		d_x / d_y	–2	–1	0	1	2	f	fd_y	fd_y^2	fd_xd_y
10–15	m.p. 12.5	–1				3 (–3)	7 (–14)	10	–10	10	–17
15–20	17.5	0		4 (0)	9 (0)	4 (0)	3 (0)	20	0	0	0
20–25	22.5	1	7 (–14)	6 (–6)	12 (0)	5 (5)		30	30	30	–15
25–30	27.5	2	3 (–12)	10 (–20)	19 (0)	8 (16)		40	80	160	–16
		f	10	20	40	20	10	$N = 100$	$\Sigma fd_y = 100$	$\Sigma fd_y^2 = 200$	$\Sigma fd_xd_y = -48$
		fd_x	–20	–20	0	20	20	$\Sigma fd_x = 0$			
		fd_x^2	40	20	0	20	40	$\Sigma fd_x^2 = 120$			
		fd_xd_y	–26	–26	0	18	–14	$\Sigma fd_xd_y = -48$			

$$r = \frac{N\Sigma fd_xd_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{(100 \times -48) - (0 \times 100)}{\sqrt{100 \times 120 - (0)^2}\sqrt{100 \times 200 - (100)^2}}$$

$$= \frac{-4800}{\sqrt{12000}\sqrt{10000}} = \frac{-48}{\sqrt{120}\sqrt{100}} = -0.438.$$

There seems to be a low degree of negative correlation between family income and its percentage spent on food expenditure.

Illustration 18. An office contains 12 clerks. The long serving clerks feel that they should have a seniority increment based on length of service built into their salary structure. An assessment of their efficiency by their departmental manager and the personnel department produces a ranking of efficiency. This is shown below together with a ranking of their length of service. Do the data support the clerks, claim for seniority increment ?

Ranking according to length of service	:	1	2	3	4	5	6	7	8	9	10	11	12
Ranking according to efficiency	:	2	3	5	1	9	10	11	12	8	7	6	4

Solution. CALCULATION OF RANK CORRELATION

Ranking according to length of service R_1	*Ranking according to efficiency* R_2	$(R_1 - R_2)$ D	D^2
1	2	−1	1
2	3	−1	1
3	5	−2	4
4	1	+3	9
5	9	−4	16
6	10	−4	16
7	11	−4	16
8	12	−4	16
9	8	+1	1
10	7	+3	9
11	6	+5	25
12	4	+8	64
$N = 12$			$\Sigma D^2 = 178$

Rank correlation coefficient is given by :

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 178}{12^3 - 12} = 1 - \frac{1068}{1716} = 1 - 0.622 = 0.378$$

Since there is a low degree of positive correlation between length of service and efficiency, the clerks' claim does not justify for a seniority increment based on the length of service.

Illustration 19. The ranks of the same 15 students in two subjects *A* and *B* are given below, the two numbers within the brackets denoting the ranks of the same student in *A* and *B* respectively :

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

Use Spearman's formula to find the rank correlation coefficient. *(MBA, Sukhadia Univ., 2008)*

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

Rank of A R_1	*Rank of B* R_2	$(R_1 - R_2)^2$ D^2
1	10	81
2	7	25
3	2	1
4	6	4
5	4	1
6	8	4
7	3	16
8	1	49
9	11	4
10	15	25
11	9	4
12	5	49
13	14	1
14	12	4
15	13	4
		$\Sigma D^2 = 272$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 272}{15^3 - 15} = 1 - .486 = 0.514.$$

There is a moderate degree of positive correlation between the ranks in subject *A* and *B*.

Illustration 20. Calculate the coefficient of correlation from the following data :

Adv. Exp. (Rs. Lakhs)	X :	10	12	15	23	20
Sales (Rs. Crores)	Y :	14	17	23	25	21

Solution. CALCULATION OF COEFFICIENT OF CORRELATION

Adv. Exp. X	$(X-16)$ x	x^2	*Sales* Y	$(Y-20)$ y	y^2	xy
10	−6	36	14	−6	36	+36
12	−4	16	17	−3	9	+12
15	−1	1	23	+3	9	−3
23	+7	49	25	+5	25	+35
20	+4	16	21	+1	1	+4
$\Sigma X = 80$	$\Sigma x = 0$	$\Sigma x^2 = 118$	$\Sigma Y = 100$	$\Sigma y = 0$	$\Sigma y^2 = 80$	$\Sigma xy = 84$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{84}{\sqrt{118 \times 80}} = \frac{84}{97.16} = +0.865.$$

There is a high degree of positive correlation between sales and advt. expenditure.

Illustration 21. Calculate coefficient of correlation from the following data taking deviation from 48 in case of X series and 20 in case of Y series :

X :	40	42	46	48	50	56
Y :	10	12	15	23	27	30

Solution. CALCULATION OF CORRELATION COEFFICIENT

X	$(X-48)$ d_x	d_x^2	Y	$(Y-20)$ d_y	d_y^2	$d_x d_y$
40	−8	64	10	−10	100	+80
42	−6	36	12	−8	64	+48
46	−2	4	15	−5	25	+10
48	0	0	23	+3	9	0
50	+2	4	27	+7	49	+14
56	+8	64	30	+10	100	+80
	$\Sigma d_x = -6$	$\Sigma d_x^2 = 172$		$\Sigma d_y = -3$	$\Sigma d_y^2 = 347$	$\Sigma d_x d_y = 232$

$$= \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$r = \frac{6 \times 232 - (-6)(-3)}{\sqrt{6(172) - (-6)^2}\sqrt{6(347) - (-3)^2}}$$

$$= \frac{1392 - 18}{\sqrt{1032 - 36}\sqrt{2082 - 9}} = \frac{1374}{\sqrt{996}\sqrt{2073}} = \frac{1374}{31.56 \times 45.53} = 0.956$$

Illustration 22. A panel of men and a panel of women were asked by a consumer testing organisation to rank 8 brands of tea according to taste. A rank of 1 was given to the best tasting tea and a rank of 8 to the worst.

Brand	:	A	B	C	D	E	F	G	I
Panel of Women (X)	:	5	4	3	6	7	8	1	2
Panel of Men (Y)	:	4	5	6	3	8	7	2	1

Determine how closely men's and women's tastes in tea are related.

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

R_1	R_2	$(R_1 - R_2)^2$
5	4	1
4	5	1
3	6	9
6	3	9
7	8	1
8	7	1
1	2	1
2	1	1
		$\Sigma D^2 = 24$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 24}{8^3 - 8} = 1 - \frac{144}{512 - 8} = 1 - 0.286 = 0.714$$

There is a high degree of positive correlation between the tastes of men and women in tea.

Illustration 23. A company gives on-the-job training to its salesmen which is followed by a test. It is considering whether it should terminate the services of any salesman who does not do well in the test.

The following data give the test scores and sales made by nine salesmen during the last one year :

Test scores	:	14	19	24	21	26	22	15	20	19
Sales (Rs. '000):		31	36	48	37	50	45	33	41	39

Compute the coefficient of correlation between test scores and sales. Does it indicate that termination of the services of salesman with low test scores is justified? ***(MBA, Madurai-Kamaraj Univ., 2007)***

Solution. CALCULATION OF CORRELATION COEFFICIENT

Test scores X	$(X - 20)$ x	x^2	*Sales* Y	$(Y - 40)$ y	y^2	xy
14	−6	36	31	−9	81	+54
19	−1	1	36	−4	16	+4
24	+4	16	48	+8	64	+32
21	+1	1	37	−3	9	−3
26	+6	36	50	+10	100	+60
22	+2	4	45	+5	25	+10
15	−5	25	33	−7	49	+35
20	0	0	41	+1	1	0
19	−1	1	39	−1	1	+1
$\Sigma X = 180$	$\Sigma x = 0$	$\Sigma x^2 = 120$	$\Sigma Y = 360$	$\Sigma y = 0$	$\Sigma y^2 = 346$	$\Sigma xy = 193$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}, \text{ where } x = (X - \bar{X}); y = (Y - \bar{Y})$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{180}{9} = 20; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{360}{9} = 40$$

Since the actual means of X and Y are whole numbers, we should take deviation from actual means of X and Y to simplify the calculations.

Substituting the values

$$r = \frac{193}{\sqrt{120 \times 346}} = \frac{193}{203.76} = 0.947$$

There is a high degree of positive correlation between test scores and sales. It does not indicate that the termination of the services of salesman with low test scores is justified.

Illustration 24. Find the correlation coefficient between age and playing habits of the following students :

Age	:	15	16	17	18	19	20
No. of students	:	250	200	150	120	100	80
Regular players	:	200	150	90	48	30	12

Solution. Let us find the percentage of regular players and then calculate coefficient of correlation between age and percentage.

X	$(X-17)$ d_x	d_x^2	No. of Students	Regular Players Y	% of Regular Players	$(Y-50)$ d_y	d_y^2	d_xd_y
15	–2	4	250	200	80	+30	900	–60
16	–1	1	200	150	75	+25	625	–25
17	0	0	150	90	60	+10	100	0
18	+1	1	120	48	40	–10	100	–10
19	+2	4	100	30	30	–20	400	–40
20	+3	9	80	12	15	–35	1225	–105
	$\Sigma d_x = +3$	$\Sigma d_x^2 = 19$				$\Sigma d_y = 0$	$\Sigma d_y^2 = 3350$	$\Sigma d_xd_y = -240$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{6(-240) - (3)(0)}{\sqrt{6\times 19 - (3)^2}\sqrt{6\times 3350 - (0)^2}} = \frac{-1440}{\sqrt{105\times 20100}} = \frac{-1440}{1452.76} = -0.991$$

Thus there is a high degree of negative correlation between age and playing habits.

Illustration 25. Calculate Karl Pearson's coefficient of correlation from the following data and interpret its value :

Roll No.	:	1	2	3	4	5
Marks in Accountancy	:	48	35	17	23	47
Marks in Statistics	:	45	20	40	25	45

Solution. Let marks in accountancy be denoted by X and that in statistics by Y.

CALCULATION OF COEFFICIENT OF CORRELATION

X	$(X-34)$ x	x^2	Y	$(Y-35)$ y	y^2	xy
48	+14	196	45	+10	100	+140
35	+1	1	20	–15	225	–15
17	–17	289	40	+5	25	–85
23	–11	121	25	–10	100	+110
47	+13	169	45	+10	100	+130
$\Sigma X = 170$	$\Sigma x = 0$	$\Sigma x^2 = 776$	$\Sigma Y = 175$	$\Sigma y = 0$	$\Sigma y^2 = 550$	$\Sigma xy = 280$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{280}{\sqrt{776\times 550}} = \frac{280}{653.3} = +0.429$$

It is a moderate case of positive correlation between marks in accountancy and statistics.

Illustration 26. Calculate the coefficient of correlation and its probable error from the following :

S.No.	Subject	% marks in final year exams.	% marks in sessionals
1	Hindi	75	62
2	English	81	68
3	Physics	70	65

4	Chemistry	76	60
5	Maths.	77	69
6	Statistics	81	72
7	Botany	84	76
8	Zoology	75	72

(MBA, Jodhpur Univ., 2001)

Solution. Let % marks in final year exams. be denoted by X and % marks in sessionals by Y.

CALCULATION OF COEFFICIENT OF CORRELATION

X	$(X-77)$ d_x	d_x^2	Y	$(Y-68)$ d_y	d_y^2	d_xd_y
75	−2	4	62	−6	36	+12
81	+4	16	68	0	0	0
70	−7	49	65	−3	9	+21
76	−1	1	60	−8	64	+8
77	0	0	69	+1	1	0
81	+4	16	72	+4	16	+16
84	+7	49	76	+8	64	+56
75	−2	4	72	+4	16	−8
$\Sigma X = 619$	$\Sigma d_x = 3$	$\Sigma d_x^2 = 139$	$\Sigma Y = 544$	$\Sigma d_y = 0$	$\Sigma d_y^2 = 206$	$\Sigma d_xd_y = 105$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{8\times 105 - (3\times 0)}{\sqrt{8\times 139 - (3)^2}\sqrt{8\times 206}} = \frac{840}{\sqrt{1103\times 1648}} = \frac{840}{1348.237} = 0.623$$

Probable error is given by :

$$PEr = .6745\,\frac{1-r^2}{\sqrt{N}} = .6745\,\frac{1-(.623)^2}{\sqrt{8}} = \frac{.6745\times .6119}{2.8284} = 0.146$$

Illustration 27. Following figures give the rainfall in inches for the year and the production in 00's of kgs. for the Rabi crop and Kharif crop. Calculate the Karl Pearson's coefficient of correlation between rainfall and total production :

Rainfall	:	20	22	24	26	28	30	32
Rabi Production	:	15	18	20	32	40	39	40
Kharif Production	:	15	17	20	18	20	21	15

Solution. Let rainfall be denoted by X and production by Y.

CALCULATION OF CORRELATION COEFFICIENT

X	$(X-26)$ d_x	d_x^2	Y	$(Y-47)$ d_y	d_y^2	d_xd_y
20	−6	36	30	−17	289	+102
22	−4	16	35	−12	144	+48
24	−2	4	40	−7	49	+14
26	0	0	50	+3	9	0
28	+2	4	60	+13	169	+26
30	+4	16	60	+13	169	+52
32	+6	36	55	+8	64	+48
$\Sigma X = 182$	$\Sigma d_x = 0$	$\Sigma d_x^2 = 112$	$\Sigma Y = 330$	$\Sigma d_y = 1$	$\Sigma d_y^2 = 893$	$\Sigma d_xd_y = 290$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$$= \frac{(7)(290) - (0)(1)}{\sqrt{(7)(112) - (0)^2}\sqrt{(7)(893) - (1)^2}} = \frac{2030}{\sqrt{784}\sqrt{6250}} = \frac{2030}{2213.594} = 0.917$$

It is a case of very high degree of positive correlation between rainfall and Agricultural production.

Illustration 28. Calculate the coefficient of correlation between weight and height for the following bivariate frequency distribution :

Weight (pounds)	*Height (inches)* 40–44	44–48	48–52	52–56	56–60	*Total*
35–55	4	40	60	—	—	104
55–75	—	—	24	88	12	124
75–95	—	—	—	8	40	48
95–115	—	—	—	—	12	12
115–135	—	—	—	4	—	4
135–155	—	—	—	4	4	8
Total	4	40	84	104	68	300

Solution. CALCULATION OF CORRELATION COEFFICIENT

Y \ X	d_y \ d_x	40–44	44–48	48–52	52–56	56–60	f	fd_y	fd_y^2	fd_xd_y
		−2	−1	0	+1	+2				
35–55	−2	16 / 4	80 / 40	0 / 60	—	—	104	−208	416	96
55–75	−1	—	—	0 / 24	−88 / 88	−24 / 12	124	−124	124	−112
75–95	0	—	—	—	0 / 8	0 / 40	48	0	0	0
95–115	+1	—	—	—	—	24 / 12	12	12	12	24
115–135	+2	—	—	—	8 / 4	—	4	8	16	8
135–155	+3	—	—	—	12 / 4	24 / 4	8	24	72	36
	f	4	40	84	104	68	$N = 300$	$\Sigma fd_y = -288$	$\Sigma fd_y^2 = 640$	$\Sigma fd_xd_y = 52$
	fd_x	−8	−40	0	104	136	$\Sigma fd_x = 192$			
	fd_x^2	16	40	0	104	272	$\Sigma fd_x^2 = 432$			
	fd_xd_y	16	80	0	−68	24	$\Sigma fd_xd_y = 52$			

$$r = \frac{N\Sigma fd_x d_y - (\Sigma fd_x)(\Sigma fd_y)}{\sqrt{N\Sigma fd_x^2 - (\Sigma fd_x)^2}\sqrt{N\Sigma fd_y^2 - (\Sigma fd_y)^2}}$$

$$= \frac{(300) \times (52) - (192)(-288)}{\sqrt{(300)(432) - (192)^2}\sqrt{(300)(640) - (-288)^2}}$$

$$= \frac{15600 + 55296}{\sqrt{129600 - 36864}\sqrt{192000 - 82944}}$$

$$= \frac{70896}{\sqrt{92736}\sqrt{109056}} = \frac{70896}{304.526 \times 330.236}$$

$$= \frac{70896}{100565.44} = +0.705.$$

Thus, it is a case of high degree of positive correlation between height and weight.

Illustration 29. The following table gives the distribution of production and also the relatively defective items among them, according to size-groups. Is there any correlation between size and defect in quality.

Size-groups	15–16	16–17	17–18	18–19	19–20	20–21
No. of items	200	270	340	360	400	300
No. of defective items	150	162	170	180	180	120

(MBA, IGNOU, 2006)

Solution : Let us find the percentage of defective items and then find correlation between size and defect in quality.

Size-groups	*m.p.* m	$(m-17.5)$ d_x	d_x^2	*No. of items*	*No. of def. items*	*% of def.*	$(Y-50)/5$ d_y	d_y^2	$d_x d_y$
15–16	15.5	−2	4	200	150	75	+5	25	−10
16–17	16.5	−1	1	270	162	60	+2	4	−2
17–18	17.5	0	0	340	170	50	0	0	0
18–19	18.5	+1	1	360	180	50	0	0	0
19–20	19.5	+2	4	400	180	45	−1	1	−2
20–21	20.5	+3	9	300	120	40	−2	4	−6
		$\Sigma d_x = 3$	$\Sigma d_x^2 = 19$			$\Sigma d_y = 4$	$\Sigma d_y^2 = 34$		$\Sigma d_x d_y = -20$

$$r = \frac{N\Sigma dxdy - (\Sigma dx)(\Sigma dy)}{\sqrt{N\Sigma dx^2 - (\Sigma dx)^2}\sqrt{N\Sigma dy^2 - (\Sigma dy)^2}}$$

Substituting the values :

$$r = \frac{6(-20) - (3)(4)}{\sqrt{6(19) - (3)^2}\sqrt{6(34) - (4)^2}}$$

$$= \frac{-120 - 12}{\sqrt{105}\sqrt{188}} = \frac{-132}{10.25 \times 13.71} = -0.94$$

There is a very high degree of negative correlation between size and defect in quality.

Illustration 30. Calculate the rank correlation coefficient for the following data giving ranks awarded by two judges to 10 participants in a musical contest :

Rank by Judge I	:	3	5	4	8	9	7	1	2	6	10
Rank by Judge II	:	4	6	3	9	10	7	2	1	5	8

(MBA, Madurai -Kamaraj Univ., 2005)

Solution : CALCULATION OF RANK CORRELATION COEFFICIENT

Participants	*Rank by Judge I* (R_1)	*Rank by Judge II* (R_2)	$(R_1 - R_2)^2$ D^2
A	3	4	1
B	5	6	1
C	4	3	1
D	8	9	1
E	9	10	1
F	7	7	0

G	1	2	1
H	2	1	1
I	6	5	1
J	10	8	4
			$\Sigma D^2 = 12$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 12}{10^3 - 10} = 1 - \frac{72}{990} = 1 - 0.073 = 0.927$$

Illustration 31. Find the rank correlation from the following data :

Candidate	1	2	3	4	5	6	7
Marks awarded by Judge I	86	59	64	74	48	70	94
Marks awarded by Judge II	90	45	72	64	59	60	80

(MBA, Bharthidasan Univ., 2003)

Solution : Since ranks are not given, we first assign ranks and then calculate the rank correlation coefficient.

Candidate	*Marks by Judge I*	R_1	*Marks by Judge II*	R_2	$(R_1-R_2)^2$ D^2
1	86	6	90	7	1
2	59	2	45	1	1
3	64	3	72	5	4
4	74	5	64	4	1
5	48	1	59	2	1
6	70	4	60	3	1
7	94	7	80	6	1
$N=7$					$\Sigma D^2 = 10$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 10}{7^3 - 7} = 1 - \frac{60}{336} = 1 - 0.179 = 0.821$$

Illustration 32. Calculate the correlation coefficient between price and sales from the following data :

Price (Rs.)	100	90	85	92	90	84	88	90
Sales ('00)	5	6	7	6	7	8	8	7

(MBA, Madras Univ., 2003)

Solution : CALCULATION OF CORRELATION COEFFICIENT

Price (Rs.) X	$(X-90)$ d_x	d_x^2	*Sales* Y	$(Y-7)$ d_y	d_y^2	$d_x d_y$
100	+10	100	5	−2	4	−20
90	0	0	6	−1	1	0
85	−5	25	7	0	0	0
92	+2	4	6	−1	1	−2
90	0	0	7	0	0	0
84	−6	36	8	+1	1	−6
88	−2	4	8	+1	1	−2
90	0	0	7	0	0	0
$\Sigma X = 719$	$\Sigma d_x = -1$	$\Sigma d_x^2 = 169$	$\Sigma Y = 54$	$\Sigma d_y = -2$	$\Sigma d_y^2 = 8$	$\Sigma d_x d_y = -30$

$$r = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{\sqrt{N\Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\Sigma d_y^2 - (\Sigma d_y)^2}}$$

$N = 8$, $\Sigma d_x d_y = -30$, $\Sigma d_x = -1$, $\Sigma d_y = -2$, $\Sigma d_x^2 = 169$, $\Sigma d_y^2 = 8$

$$r = \frac{8(-30) - (-1)(-2)}{\sqrt{8(169) - (-1)^2}\sqrt{8(8) - (-2)^2}} = \frac{-240 - 2}{\sqrt{1352 - 1}\sqrt{64 - 4}}$$

$$= \frac{-242}{\sqrt{1351 \times 60}} = \frac{-242}{284.71} = -0.85$$

There is a high degree of negative correlation between price and sales.

Illustration 33. Newspapers in India are complaining that rising level of unemployment is affecting the level of crime in the country. To study this claim, a research team studied a random sample of 12 states in the country. For each state, they measured the level of unemployment rate and the crime rate in the state. Then they did a ranking X = level of unemployment, Y = crime rate, the results are shown in the following table. Higher X ranks more unemployment, and higher Y ranks means higher crime rate. Test the claim of Newspapers.

States	:	1	2	3	4	5	6	7	8	9	10	11	12
Level of unemployment (X)	:	5	8	3	2	6	1	10	12	7	4	9	11
Crime Rate (Y)	:	8	6	9	12	7	10	2	1	5	11	4	3

(MBA, Delhi Univ., 2009)

Solution : For testing the claim of the newspapers, we calculate the rank correlation coefficient.

CALCULATION OF RANK CORRELATION COEFFICIENT

States	R_x	R_y	$(R_x - R_y)^2$ D^2
1	5	8	9
2	8	6	4
3	3	9	36
4	2	12	100
5	6	7	1
6	1	10	81
7	10	2	64
8	12	1	121
9	7	5	4
10	4	11	49
11	9	4	25
12	11	3	64
$N = 12$			$\Sigma D^2 = 558$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$\Sigma D^2 = 558, \quad N = 12$

$$R = 1 - \frac{6 \times 558}{12^3 - 12} = 1 - \frac{3348}{1728 - 12} = 1 - 1.951 = -0.951$$

There is a high degree of negative correlation between level of unemployment and crime rate.

Illustration 34. Compute Spearman's rank correlation for the following observations :

Candidate	:	1	2	3	4	5	6	7	8
Judge X	:	20	22	28	23	30	30	23	24
Judge Y	:	28	24	24	25	26	27	32	30

(MBA, GGSIP Univ., 2009)

Solution : CALCULATION OF SPEARMAN'S RANK CORRELATION

Candidate	*Judge X*	R_1	*Judge Y*	R_2	$(R_1 - R_2)^2$ D^2
1	20	1	28	6	25.00
2	22	2	24	1.5	0.25
3	28	6	24	1.5	20.25
4	23	3.5	25	3	0.25
5	30	7.5	26	4	12.25
6	30	7.5	27	5	6.25
7	23	3.5	32	8	20.25
8	24	5	30	7	4.00
$N = 8$					$\Sigma D^2 = 88.50$

$$R = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_1) + \ldots\right]}{N^3 - N}$$

$$R = 1 - \frac{6\left[88.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]}{N^3 - N}$$

$$= 1 - \frac{6[88.5 + 0.5 + 0.5 + 0.5]}{504} = 1 - \frac{540}{504} = 1 - 1.071 = -0.071$$

PROBLEMS

I-A : Answer the following questions, each question carries **one** mark:

(*i*) What are the properties of correlation coefficient ?

(*ii*) What are the limitations of correlation analysis ? (*MBA, Madurai-Kamaraj Univ., 2005*)

(*iii*) State the formula for coefficient of correlation in terms of regression coefficients.

(*iv*) What is meant by correlation ?

(*v*) What are the limits of coefficient of correlation ?

(*vi*) What is the use of scatter diagram ?

(*vii*) What is 'Rank correlation' ? (*MBA, Madurai-Kamaraj Univ., 2003*)

(*viii*) Write down the formula for rank correlation coefficient.

(*ix*) Interpret the following value of r : $r = 0$, $r = -1$, $r = +1$, $r = 0.25$.

(*x*) How can 'r' be determined through regression coefficients ?

I-B : Answer the following questions, Each question carries **four** marks:

(*i*) The coefficient of correlation between the variables x and y is 0.64, their covariance is 16. The variance of x is 9. Find the standard deviation of y.

(*ii*) Briefly explain the various types of correlation. (*M.Com., M.K. Univ., 2002*)

(*iii*) What do you understand by correlation ? Describe the uses of the study of correlation. (*M.A. Eco., M.K. Univ., 2003*)

(*iv*) Define correlation between two variables. How is the value of 'r' interpreted ? (*MBA, Madras Univ., 2003*)

(*v*) Does correlation always signify a cause and effect relationship between variables ?

2. (*a*) Explain the meaning and significance of the term correlation.

(*b*) What is correlation ? Clearly explain with suitable illustration its role in taking some business problem. (*MBA, Delhi Univ., 2002*)

3. Define the coefficient of correlation. What is it intended to measure ? How would you interpret the sign and magnitude of a calculated r ? Consider in particular the values of $r = 0$, $r = +1$ and $r = -1$.

4. What is a scatter diagram? How does it help in studying the correlation between two variables, in respect of both its direction and degree ? (*MBA, Delhi Univ., 2007*)

5. (*a*) What is Spearman's rank correlation coefficient? Bring out its usefulness. How does the coefficient differ from Karl Pearson's coefficient of correlation ?

(*b*) Explain briefly the different methods of measuring correlation.

6. (*a*) Does correlation always signify a cause and effect relationship between the variables?

(*b*) Does a high positive correlation between the increase in cigarette smoking and the increase in lung cancer prove that one causes the other ?

7. (*a*) Define correlation coefficient 'r' and give its limits. What interpretation would you give if told that the correlation between the number of truck accidents per year and the age of the driver is (–) 0.60 if only drivers with at least one accident are considered ?

(*b*) What is a scatter diagram ? How do you interpret a scatter diagram ?

8. (*a*) What is correlation? Does it always signify cause and effect relationship?

(*b*) What is coefficient of Rank correlation? Bring out its usefulness. How does this coefficient differ from coefficient of correlation ?

9. (*a*) Prove that the correlation coefficient is unaffected by the change of origin and scale.

(*b*) How is Scatter Diagram helpful in the study of correlation ?

10. (*a*) Explain how covariance of X and Y is related to the coefficient of simple correlation between X and Y.

(*b*) What is meant by correlation ? Distinguish between positive, negative and zero correlation. (*MBA, Delhi Univ., 2005; MBA, UP Tech. Univ., 2006*)

(*c*) Explain critically any two methods of measuring correlation.

11. Find Karl Pearson's coefficient of correlation from the following index numbers and interpret it :

Wages	:	100	101	103	102	104	99	97	98	96	96
Cost of living	:	98	99	99	97	95	02	95	94	90	91

[$r = 0.85$]

12. Find Karl Pearson's coefficient of correlation between capital employed and profit obtained from the following data :

Capital employed (*Rs. crore*)	*Profits obtained* (*Rs. crore*)	*Capital employed* (*Rs. crore*)	*Profits obtained* (*Rs. crore*)
10	2	60	15
20	4	70	14
30	8	80	20
40	5	90	22
50	10	100	50

[$r = 0.85$]

13. Using the following data :

(*a*) Calculate the coefficient of correlation.

(*b*) Estimate the percentage of the group with lung cancer in a country where 15 per cent of the group smoke heavily :

Country	*% of group smoking heavily*	*% of group with lung cancer*
A	10	5
B	20	15
C	20	20
D	30	25
E	30	20

[$r = 0.91$]

14. From the following data, calculate coefficient of correlation between the percentage yield on securities and wholesale price indices for certain years :

Year	:	2004	2005	2006	2007	2008	2009	2010
% Yield on securities	:	5.0	5.1	5.2	4.9	4.8	5.3	5.4
Index No. of wholesale prices	:	140	138	126	132	140	135	132

What inference do you draw from the result ?

[$r = -0.16$]

15. Find the correlation by Karl Pearson's method between the two kinds of assessment of postgraduate student's performance (marks out of 100) :

Roll No. of students	:	1	2	3	4	5	6	7	8	9	10
Internal assessment	:	45	62	67	32	12	38	47	67	42	85
External assessment	:	39	48	65	32	20	35	45	77	30	62

[$r = 0.88$]

16. Two housewives, Mrs. Neena and Mrs. Meena, asked to express their preferences for different kinds of detergents, gave the following replies :

Detergent	:	A	B	C	D	E	F	G	H	I	J
Neena	:	1	2	4	3	7	8	6	5	9	10
Meena	:	1	4	2	3	5	7	6	8	9	10

To what extent the preferences of these two ladies go together ?

[$R = +0.89$]

17. An office contains 10 clerks. The longer-serving clerks feel that they should have a seniority increment based on length of service built into their salary structure. An assessment of their efficiency by their departmental manager and the personnel department produces a ranking of efficiency. This is shown below together with a ranking of their length of service. Do the data support the clerk's claim for seniority increment ?

Ranking according to length of service	:	1	2	3	4	5	6	7	8	9	10
Ranking according to efficiency	:	2	5	3	10	6	4	8	9	7	1

[R = +0.164]

18. The following table gives the frequency, according to age groups, of marks obtained by 68 students in a general knowledge test. Measure the degree of relationship between age and general knowledge.

	Age in Years			
Test Marks	21	22	23	24
200 – 250	4	4	2	1
250 – 300	3	5	4	2
300 – 350	2	6	8	5
350 – 400	1	5	6	10

[$r = 0.415$]

19. Find coefficient of correlation between output and cost per scooter from the following data :

Output of scooter (in '000s) :	3.5	4.0	5.2	6.3	6.8	7.4	8.5	9.0
Cost per scooter (in '000 Rs.) :	12.0	11.8	11.2	10.6	10.3	9.8	9.3	9.2

[r=0.0996]

20. Find the coefficient of correlation between price and sales from the following data :

Price (Rs.) :	103	98	85	92	90	84	88	90	93	95
Sales (Units) :	500	610	700	630	670	800	800	750	700	680

[r = 0.85]

21. Calculate correlation coefficient from the following two-way table, with X representing the average salary of families selected at random in a given area and Y representing the average expenditure on entertainment (movies, magazines, etc.) :

Expenditure on entertainment (in 00's Rs.)	*Average salary (in 00's Rs.)* 100 – 150	150 – 200	200 – 250	250 – 300	300 – 350
0 – 10	5	4	5	2	4
10 – 20	2	7	3	7	1
20 – 30	–	6	–	4	5
30 – 40	8	–	4	–	8
40 – 50	–	7	3	5	10

[r = 0.205]

(MBA, Delhi Univ., 2003)

22. A psychologist wanted to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so that the students in a pair have approximately equal scores on an intelligence test. In each pair, one student was taught by method A and the other by method B and examined after the course. The marks obtained by them are tabulated below :

Pair :	1	2	3	4	5	6	7	8	9	10	11
A :	24	29	19	14	30	19	27	30	20	28	11
B :	37	35	16	26	23	27	19	20	16	11	21

(*i*) Find the correlation coefficient between the two sets of scores.

(*ii*) Find the rank correlation coefficient.

(MBA, HPU, 2004)

[(*ii*) –0.175]

23. The mileage (Y) that can be obtained from a certain gasoline depends on the amount (X) of certain chemical in the gasoline. The value of ten observations, where X and Y are measured in appropriate units are shown in the table below :

Amount (X)	*Mileage* (Y)	*Amount* (X)	*Mileage* (Y)
0.10	10.98	0.60	14.63
0.20	11.14	0.70	15.66
0.30	13.17	0.80	13.71
0.40	13.34	0.90	15.43
0.50	14.39	1.00	18.36

Find the coefficient of correlation between X and Y and represent the data by a graph.

24. Calculate the coefficient of correlation between age and sum assured from the data given below and comment on the value :

	Sum assured (in lakh Rs.)				
Age	5	10	15	20	Total
20 – 30	2	3	4	6	15
30 – 40	–	2	3	5	10
40 – 50	–	2	2	3	7
50 – 60	5	8	3	2	18
Total	7	15	12	16	50

[r = 0.3442]

(MBA, Delhi Univ., 2002)

25. Compute the coefficient of correlation between dividends and prices of securities as given below :

Security Prices (in Rs.)	*Annual Dividends (in hundred Rs.)*					
	6 – 8	8 – 10	10 – 12	12 – 14	14 – 16	16 – 18
130 – 140	—	—	1	3	4	2
120 – 130	—	1	3	3	3	1
110 – 120	—	1	2	3	2	—
100 – 110	—	2	3	2	—	—
90 – 100	2	2	1	1	—	—
80 – 90	3	1	1	—	—	—
70 – 80	2	1	—	—	—	—

[$r = +0.71$]

26. The top executives of Sonal Electrical rank managerial candidates on the basis of what they know about each candidate. In order to determine if there is any consistency in the ranking obtained in this manner, two vice-presidents were asked to rank the same ten candidates. Compute the coefficient of rank correlation from the following two sets of ranks :

Candidate	A	B	C	D	E	F	G	H	I	J
V-P 1 :	3	1	8	1	4	9	5	7	10	6
V-P 2 :	2	5	9	1	6	10	3	4	8	7

($R = +0.746$]

27. Seven methods of imparting business education were ranked by the MBA students of two universities as follows :

Method of teaching	:	I	II	III	IV	V	VI	VII
Rank by Students of Univ. A	:	2	1	5	3	4	7	6
Rank by Students of Univ. B	:	1	3	2	4	7	5	6

Calculate rank correlation coefficient and comment on its value.

[$R = +0.5$] *(MBA, South Gujarat Univ., 2002; MBA, Delhi Univ, 2005)*

28. (*a*) Coefficient of correlation between X and Y for 20 items is 0.3, mean of X is 15 and that of Y=20, standard deviations are 4 and 5 respectively. At the time of calculation one item 26 was wrongly taken as 17 in case of X series and 35 instead of 30 in case of Y series. Find the correct value of correlation coefficient.

[Correct value of correlation coefficient is 0.504.]

(*b*) In order to find the correlation coefficient between two variables 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 verification it was found that the pair ($X = 11, Y = 4$) was copied wrongly, the correct value being ($X = 10, Y = 14$). Find the correct value of correlation coefficient.

[$r = 0.78$]

29. A Statistician while calculating the correlation coefficient between two variates X and Y from 25 pairs of observations obtained the following results :

$n = 25, \Sigma X = 125, \Sigma X^2 = 650, \Sigma Y = 100, \Sigma Y^2 = 460, \Sigma XY = 508.$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	:	6	8
Y	:	14	6

While the correct values were

X	:	8	16
Y	:	12	8

Obtain the correct value of the correlation coefficient.

(M.Com., Madras Univ., 2009)

[$r = 0.67$]

30. The following data relate to the prices and supplies of a commodity during a period of eight years :

Price (Rs./kg) :	10	12	18	16	15	19	18	17
Supply (100 kg) :	30	35	45	44	42	48	47	46

Calculate the coefficient of correlation between the two series.

(MBA, Punjab Univ., 2002)

[$r = 0.98$]

31. Calculate the coefficient of correlation between family income and its percentage spent on food for the following data :

Family Income (in Rs.)	*Food Expenditure (in percentage)*				
	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
12000 – 13000	2	3	1	4	—
13000 – 14000	3	4	2	1	5
14000 – 15000	4	1	5	12	8
15000 – 16000	1	2	3	—	4
16000 – 17000	5	6	—	3	1

[$r = 0.1048$]

32. Calculate the coefficient of correlation and probable error of r between the values of X and Y given below :

X	:	78	98	96	69	59	79	68	61
Y	:	125	137	156	112	107	136	123	108

[$r = 0.955$, P.E.r. = 0.021]

(*M.Com., Sukhadia Univ., 2000*)

33. Find the coefficient of correlation for the following bivariate frequency distribution :

Marks in Physics	*Marks in Mathematics*						
	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99	Total
90 – 99				2	4	4	10
80 – 89			1	4	6	5	16
70 – 79			5	10	8	1	24
60 – 69	1	4	9	5	2		21
50 – 59	3	6	6	2			17
40 – 49	3	5	4				12
Total	7	15	25	23	20	10	100

[$r = +0.769$]

(*M. Com., M.D. Univ., 2003*)

34. The bivariate frequency distribution based on monthly salary and age of 100 employees working in some large-scale commercial organisation is as under :

Age (Years)	*Monthly Salary (in 000's Rs.)*			
	8 – 10	10 – 12	12 – 14	14 – 16
20 and less than 30	16	6	—	—
30 and less than 40	4	10	4	4
40 and less than 50	—	4	18	12
50 and less than 60	—	—	10	12

Compute Karl Pearson's coefficient of correlation between age and monthly salary of employees and comment on its value.

[$r = +0.763$]

35. A survey regarding income and savings provided the following data :

Income (Rs.)	*Saving (Rs.)*			
	1000	2000	3000	4000
8000	8	4	—	—
12000	—	12	24	6
16000	—	9	7	2
20000	—	—	10	5
24000	—	—	9	4

Compute Karl Pearson's coefficient of correlation and interpret its value.

[$r = +0.522$]

(*MBA Delhi Univ., 2006*)

36. Calculate the coefficient of correlation from the following data and interpret the value.

Advertising expenditure (Rs. lakhs)	:	10	12	13	23	27	30
Sales turnover (Rs. crores)	:	40	42	46	48	50	56

[$r = +0.956$]

(*MBA, Delhi Univ., 2002*)

37. You are given the following data of marks obtained by 11 students in statistics in two tests, one before and the other after special coaching :

First Test (Before coaching)	:	23	20	19	21	18	20	18	17	23	16	19
Second Test (After coaching)	:	24	19	22	18	20	22	20	20	23	20	17

Do the marks indicate that the special coaching has benefited the students ?

[$r = +0.477$]

(*M.Com., Delhi Univ., 2000*)

38. The scores of students in an examination in Mathematics and Statistics are given below :

Student No.	:	1	2	3	4	5	6	7	8
Marks in Mathematics	:	70	48	58	55	54	50	60	52
Marks in Statistics	:	62	47	53	60	55	68	51	48

Find : (*i*) Correlation coefficient, and

(*ii*) Rank correlation coefficient and compare the two values.

[(*i*) $r = 0.246$, (*ii*) $r = 0.286$]

39. The following data show the marks of 10 students in Mathematics and Statistics in an examination :

Marks in Mathematics :	45	70	65	30	90	40	50	75	85	60
Marks in Statistics :	35	90	70	40	95	40	60	80	80	50

Find Karl Pearson's coefficient of correlation and its probable error. *(MBA, Vikram Univ., 2007)*

40. A researcher collected the following information for two variables x and y :

No. of Pairs = 20, $r = 0.5$, $\bar{x} = 15$, $\bar{y} = 20$, $\sigma_x = 4$, $\sigma_y = 5$

Later it was found that one pair of value has been wrongly taken as $\begin{array}{|c|c|}\hline x & y \\ \hline 16 & 30 \\ \hline \end{array}$ whereas the correct values were $\begin{array}{|c|c|}\hline x & y \\ \hline 26 & 35 \\ \hline \end{array}$. Find the correct value of r.

[$r = 0.559$] *(MBA, MD Univ., 2002)*

41. Calculate the Karl Pearson's Coefficient of Correlation between age and playing habits from the data given below. Comment on the value :

Age	:	20	21	22	23	24	25
No. of students	:	500	400	300	240	200	160
Regular players	:	400	300	180	96	60	24

[$r = -0.991$] *(MBA, Delhi Univ., 2006)*

42. The following bivariate frequency distribution relates to the age and salary of 100 computer operators working in an organisation. Find the coefficient of correlation and interpret its value.

	Salary (Rs.)			
Age (Yrs)	*15000 – 16000*	*16000 – 17000*	*17000 – 18000*	*18000 – 19000*
20 – 30	4	6	5	2
30 – 40	2	5	8	5
40 – 50	8	12	20	2
50 – 60	—	8	12	1

[$r = 0.057$] *(MBA, Delhi Univ., 2006)*

43. Compute the rank correlation coefficient from the following data :

Series X :	115	109	112	87	98	98	120	100	98	118
Series Y :	75	73	85	70	76	65	82	73	68	60

[$R = 0.33$] *(MBA, KU, 2008)*

44. From the following data calculate coefficient of correlation between age and playing habit. How do you interpret the result?

Age group	*No. of Employees*	*No. of regular players*
20 – 30	50	20
30 – 40	120	60
40 – 50	80	24
50 – 60	40	4
60 – 70	20	1

(MBA, Guru Jambheshwar Univ., 2005)

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

X :	65	66	67	67	68	69	70
Y :	67	68	65	68	72	72	69

[$r = +0.603$] *(MBA, Madurai-Kamaraj Univ., 2004)*

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

X :	100	200	300	400	500	600	700
Y :	30	50	60	80	100	110	130

[$r = 0.997$]

47. Find the coefficient of correlation for the following :

A :	5	10	5	11	12	4	3	2	7	1
B :	1	6	2	8	5	1	4	6	5	2

(MBA, Madurai-Kamaraj Univ., 2003)

48. Two designs A and B gave the following output in 9 trails of each. Which is a better design ? Why ?

A :	16	16	53	15	31	17	14	30	20
B :	18	27	23	21	22	26	39	17	28

49. Calculate Pearson's coefficient of correlation from the following taking 100 and 50 as the assumed average of X and Y respectively :

X :	104	111	104	114	118	117	105	108	106	100	104	105
Y :	57	55	47	45	45	50	64	63	66	62	69	61

(MBA, Bharathidasan Univ., 2001)

50. Find the correlation coefficient from the following data :

X :	53	25	19	37	42	10	15
Y :	9	6	5	7	7	4	5

[$r = 0.348$]

51. The marking of 10 trainees in two skills, programming and analysis are as follows. What is the coefficient of rank correlation?

Programming :	3	5	8	4	7	10	2	1	6	9
Analysis :	6	4	9	8	1	2	3	10	5	7

[$r = -0.297$]

(MBA, Bharathidasan Univ., 2006)

52. The GE Capital is in the business of making bids on investments offered by various firms that desire additional financing. The company has collected the following data on yearly investments and interest rates :

Year :	*1999*	*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*
Yearly Investments (Thousand of Rs.) :	1080	948	920	1119	1695	2150	2170	2230	1880	1425
Average Interest Rate (%) :	4.8	5.1	5.9	5.1	4.8	3.8	3.7	4.5	4.9	6.2

Is the relationship between these variables significant? If the average interest rate is 6% five years from now, can yearly investment be forecast?

(MBA, Delhi Univ., 2009)

53. A consulting firm is preparing a study on consumer behaviour. The company collected the following data in thousand dollars to determine whether there is a relationship between consumer income and consumption levels :

Consumer No. :	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Income :	24.3	12.5	31.2	28.0	35.1	10.5	23.2	10.0	8.5	15.9	14.7	15
Consumption :	16.2	8.5	15	17	24.2	11.2	15	7.1	3.5	11.5	10.7	9.2

(*a*) Calculate correlation coefficient for the above data.

(*b*) Compute and interpect the regression model. Tell about the relationship between consumption and income? What consumption would the model predict for someone who earns $27500?

(MBA, Delhi Univ., 2009)

*****