9

CORRELATION ANALYSIS

9.1 INTRODUCTION

In the previous chapters we have studied the characteristics of only one variable in the form of the measures of central tendency, the measures of dispersion, skewness and kurtosis. A distribution of the values of one variable is called a *univariate distribution* and so far we have confined ourselves to the study of the characteristics of univariate distributions. In this chapter, we shall deal with problems and methods related with the determination of relationship between two variables. A distribution of paired values of two variables is called *bivariate distribution.*

In a bivariate distribution we may be interested to find if there exists any relationship between two variables such as income and expenditure, price and demand, height and weight, rainfall and crop yield etc. In statistics, we study the relationship between two variables with the help of *Correlation Analysis.* It involves various methods and techniques used for studying and measuring the nature and degree of relationship between two or more variables. According to **A.M. Tuttle, "*Correlation is an analysis of the covariance between two or more variables.*"**

9.2 CORRELATION – MEANING AND DEFINITIONS

The term *correlation* means the quantitative relationship between two variables. Two variables are said to be correlated when the value of one variable changes with the change in the value of other variable. According to **Croxton and Cowden, "*When relationship is of a quantitative nature the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation*".** The concept of correlation has been similarly defined by **W.I. King** as, "*If it is proved true that in a large number of instances two variables tend always to fluctuate in the same or in opposite directions, we consider that the fact is established and that a relationship exists. The relationship is called correlation.*" Similarly, **Prof. Boddington** states, "*Whenever some connection exists between two or more groups, classes or series of data, they are said to be correlated*".

Sir Francis Galton was one of the first few statisticians who developed the technique of establishing the concept of correlation graphically in 1896. **Prof. Karl Pearson** had introduced the formula based on mathematical treatments and established the extent of correlation.

9.3 TYPES OF CORRELATION

The main types of correlation are as follows :

1. Positive and Negative Correlation

(*a*) Positive or Direct Correlation : If the values of two variables move in same direction, the correlation is said to be *positive.* In other words, if an increase or decrease in the values of one variable is associated with an increase or decrease in the values of the other variable, the correlation between them is said to be *direct* or *positive.*

Some examples of positive correlation are :

(*i*) Price and supply of a commodity

(*ii*) Income and expenditure of a family on luxury items

(*iii*) Heights and weights

(*iv*) Temperature and sale of ice-cream during summer etc.

(*b*) Negative or Inverse Correlation : If the values of two variables move in the opposite direction, the correlation is said to be *negative*. In other words, if an increase or decrease in the value of a variable is associated with a decrease or increase in the values of the other, the correlation between them is *inverse* or *negative*.

Some examples of negative or inverse correlation are :

(*i*) Price and demand for a commodity

(*ii*) Number of workers and time required to complete the work

(*iii*) Volume and pressure of gas etc.

Note Two variables are said to have *zero correlation* if they are not related with each other. In other words, if two variables are independent of each other then there is no or zero correlation between them. For example, the height of students and marks obtained by them, price of rice and demand for coffee have zero correlation.

2. Simple and Multiple Correlation

(*a*) Simple Correlation : In *simple correlation*, the study relates to two variables only. For example, the study of correlation between income and saving, price and demand etc.

(*b*) Multiple Correlation : If there are more than two variables and one variable is related to a number of variables, the study of relationship between one variable and all other variables taken together is called *multiple correlation*. For example, the study of relationship between production of a crop(X) and rainfall(Y), use of fertilizer(Z) taken together falls under multiple correlation.

3. Partial and Total Correlation

(*a*) Partial Correlation : Under *partial correlation*, there are more than two variables and we study the relationship between any two variables keeping all other variables as constant. For example, studying the relation between yield of some crop (X) and chemical fertilizers (Y) without considering the effect of rainfall(Z) is known as partial correlation.

(*b*) Total Correlation : Total correlation refers to the study of relation between all the relevant variables under study at a time.

4. Linear and Non-linear Correlation

(*a*) Linear Correlation : Two variables are said to have *linear correlation* if corresponding to a unit change in one variable, there is a constant change in the other variable over the whole distribution.

For example, consider the following data :

X :	1	2	3	4	5
Y :	3	5	7	9	11

Here for a unit change in the value of X there is a constant change of 2 in the corresponding value of Y. So these two variables X and Y are said to have linear correlation.

(*b*) Non-linear Correlation : Two variables are said to have *non-linear correlation* if corresponding to a unit change in one variable, the other variable does not change at a constant rate.

For example, consider the following data :

X :	1	2	3	4	5
Y :	5	7	10	17	28

Here for a unit change in the value of X the value of Y does not change at a constant rate. So these two variables X and Y are said to have non-linear correlation.

Notes (*a*) Non-linear correlation is also known as *curvilinear correlation.*

(*b*) The techniques for analysis and measurement of non-linear correlation is quite complicated as compared to that for linear correlation. So generally we assume that the relationship between two variables is linear. In this chapter we shall confine our study to the linear correlation only.

5. Logical and Illogical Correlation

(*a*) Logical Correlation : If correlation between two variables is calculated mathematically and this relationship is logical too then correlation is said to be *logical.* Correlation between demand and price, income and expenditure, yield of a crop and use of fertilizer are examples of logical correlation.

(*b*) Illogical Correlation : There are some instances when we come across with some variables which have no logical relationship with each other but mathematically we can establish a relationship between them by applying usual formulae of correlation analysis. For example, consider income and height of a group of persons. These variables are not related to each other in any way but correlation between them can be determined. Such a correlation is known as *illogical correlation* or *non-sense correlation* or *spurious correlation.*

9.4 CORRELATION AND CAUSATION

The presence of high or moderate degree of correlation not always implies that there is a functional relationship between the variables but on the contrary, a functional relation between the variable always implies that there is a correlation between the variables. Following factors determine the statistical correlation between two or more variables :

(*i*) Correlation may be due to a pure chance : There can be a possibility that we get data involving variables which are not logically or functionally related to each other but still there exists a high degree of correlation between the two. Such a correlation may be due to chance only and this correlation is known as *spurious* or *non-sense* correlation.

(*ii*) Correlation may be due to mutual dependence : Mutual dependence of two variables on each other may result in high degree of correlation. But in such case it is very difficult to find out the variable as cause and another variable as effect, because both are affecting each other. For example, price and demand for a commodity are interrelated. Sometimes price determines the demand and sometimes demand determines the price of a commodity.

(*iii*) Variables may be affected by some other variables : Suppose the production of wheat and sugar in a certain year has shown a high degree of positive correlation. There exists no functional relation between the two. An increase or fall in the production of these crops may be due to some other factors such as level of rainfall, use of chemical fertilizers etc. Thus the high degree of correlation or no correlation may not be due to cause and effect of each other but due to their dependence on some other variable or variables.

Thus we conclude that correlation does not always lead to causation, but causation always leads to correlation.

9.5 USES OR SIGNIFICANCE OF CORRELATION

In real life, the study of correlation is significant in following ways :

(*i*) **It Reduces the Range of Uncertainty.** The application of the concept of correlation is wide based both in physical and social sciences. Prediction and forecasting plays an important role in policy making and planning. The study of correlation comes to our rescue in making relatively more reliable and dependable predictions.

(*ii*) **It Depicts the Average of Relationship.** While studying the movements in values of the variables we do not find, in general, any uniformity in it. Correlation analysis gives us a single value which can conclude the nature and extent of relationship between the variables.

(*iii*) **It Acts as Base for Other Statistical Measures.** Correlation analysis is closely related with regression analysis. With the help of regression analysis, we can estimate the value of one variable given the value of another variable. To sum up, we can quote **W.A. Neiswanger** that "***Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the path through which stabilizing forces may become effective***".

(*iv*) Correlation analysis is helpful for economists. With its help, they can judge about the dependence and relationship of two variables.

9.6 METHODS OF CORRELATION ANALYSIS

Following are the methods of finding correlation between two variables :

(*i*) Scatter diagram method

(*ii*) Graphic method or Correlogram

(*iii*) Karl Pearson's coefficient of correlation or Product Moment Method or Co-variance Method

(*iv*) Spearman's coefficient of correlation or Rank correlation coefficient

(*v*) Concurrent Deviation Method

Out of these methods, first two are based on diagrams and graphs and next three methods are mathematical methods.

In next sections, we shall study each of these methods.

9.7 SCATTER DIAGRAM METHOD

This is one of the easiest method of establishing correlation between two variables. This method is also called *Dot Diagram or Dotgram* or *Scattergram* . Following steps are involved in the construction of scatter diagram and measuring correlation.

(*i*) Out of the given two series, one is taken as the X-Series and the other as Y-Series.

(*ii*) X variable is measured along horizontal line known as x-axis and Y variable along the vertical line known as y-axis on a graph paper.

(*iii*) Taking the first pair of values of the variables as the co-ordinates of the first point, we plot it on the graph paper according to some suitable scale.

(*iv*) Plot all the pairs of observations on the same graph paper using the same scale.

(v) If these points establish some pattern or trend, may it be upward or downward from left to right, the two variables will be accordingly correlated. But if the scatterness of the points on the graph paper does not show any trend, the two variables are said to have no correlation. Following can be the possible shapes of the clustering of the points on the basis of which nature and extent of correlation is determined :

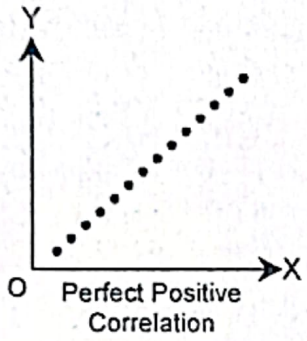


Fig. 9.1

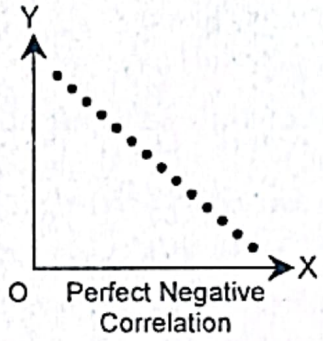


Fig. 9.2

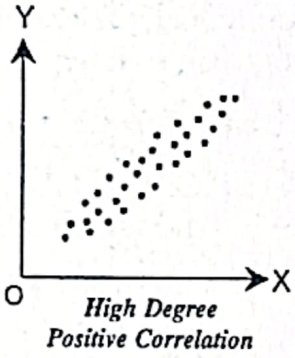


Fig. 9.3

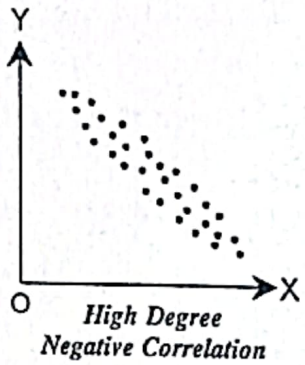


Fig. 9.4

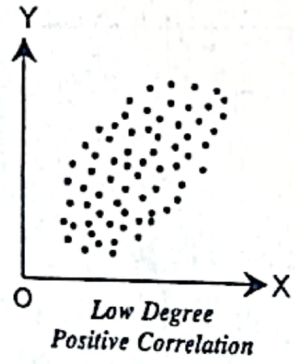


Fig. 9.5

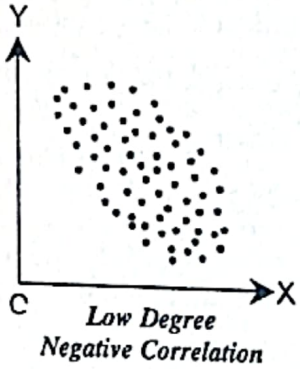


Fig. 9.6

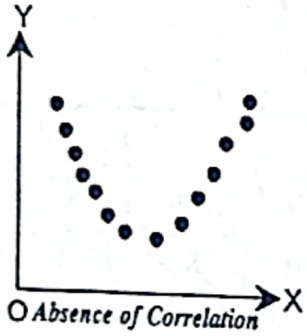


Fig. 9.7

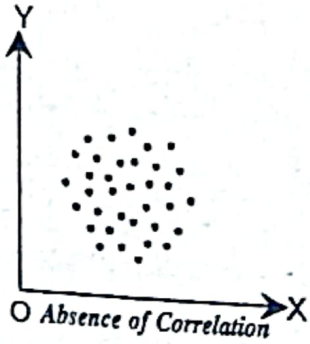


Fig. 9.8

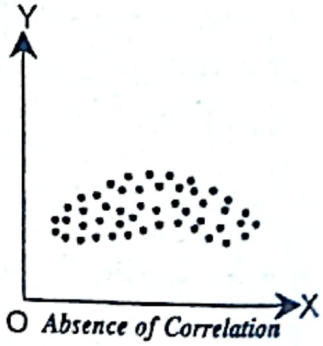


Fig. 9.9

Merits and Demerits of Scatter Diagram Method

Merits : The following are the merits of scatter diagram :

(*i*) It is the simplest method of studying and explaining correlation.

(*ii*) By just having a glance over the diagram we can establish the nature and extent of correlation whether positive or negative, perfect or high or low.

(*iii*) This method is least affected by presence of extreme values.

(*iv*) This method involves no mathematical calculations.

(*v*) The scatter diagram method is very much useful in detecting the abnormal values in the data.

(*vi*) We can trace out a line of best fit with the help of scatter diagram.

Demerits : The following are the limitations of this method :

(*i*) We cannot determine the exact numerical measure of extent or degree of correlation with this method.

(*ii*) The study of scatter diagram is limited to two variables only. If there are more than two variables then this method fails.

(*iii*) If there exists a very large number of items or number of items are too small then this method is not suitable.

(*iv*) This is a rough method.

CHECKPOINTS

1. What do you understand by Correlation ?

(G.N.D.U. B.Sc. C.Sc. April 2002; P.U. B.C.A. Sept. 2008)

2. Discuss (*i*) Positive and Negative Correlation

(*ii*) Linear and Nonlinear Correlation (*iii*) Scatter diagram.

(G.N.D.U. B.C.A. April 2003, 2004)

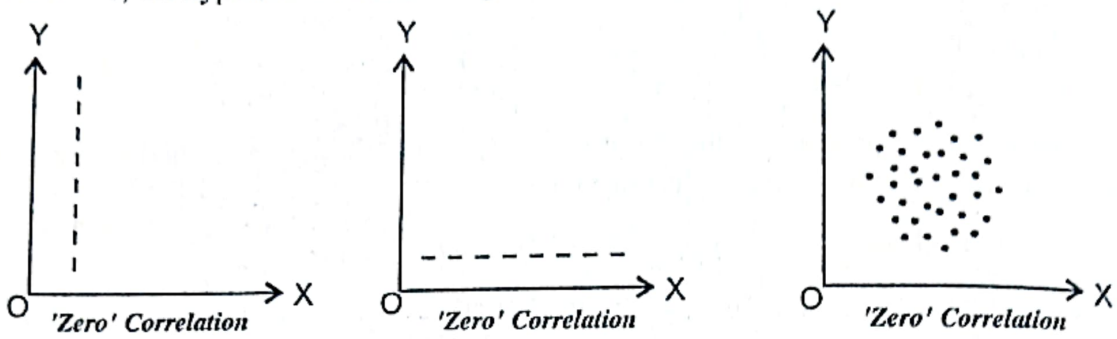
3. Discuss the significance of Correlation in statistical analysis.
4. List various methods of correlation analysis.
5. What are the merits and demerits of scatter diagram method ?

ILLUSTRATIVE EXAMPLES

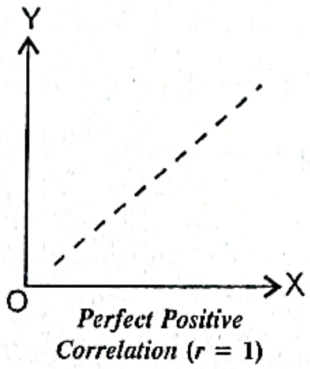
Example 1. Draw the hypothetical scatter diagrams to explain the followings :

(*i*) $r = 0$, (*ii*) $r = 1$, (*iii*) $r = -1$, (*iv*) $0 < r < 1$, (*v*) $-1 < r < 0$

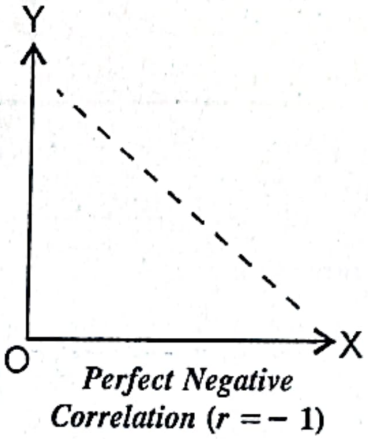
Sol. (*i*) When $r = 0$, the hypothetical scatter diagram will be as under :



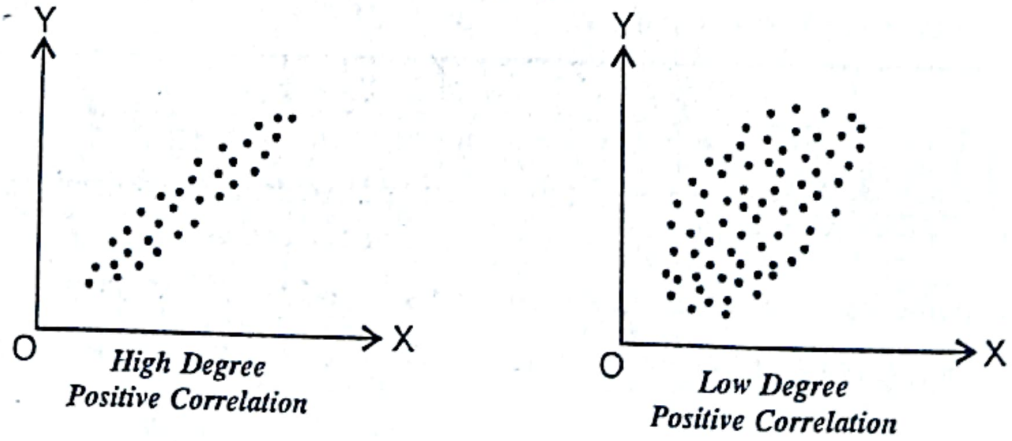
(*ii*) When $r = 1$, the hypothetical scatter diagram will be as under ;



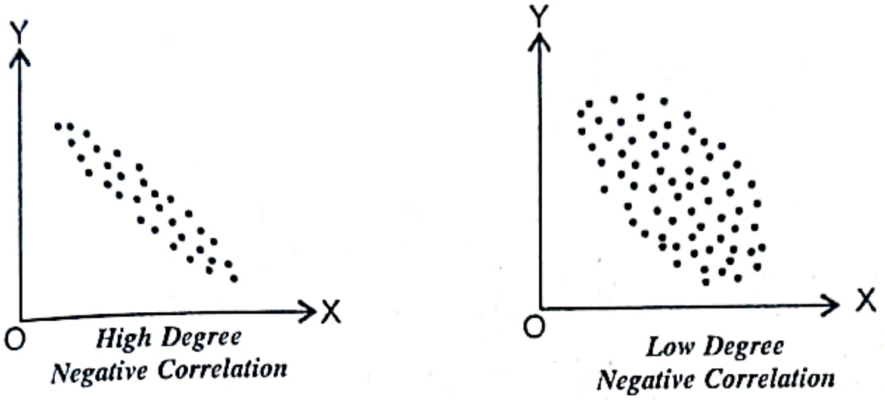
(*iii*) When $r = -1$, the hypothetical scatter diagram will be as under :



(*iv*) When $0 < r < 1$, the hypothetical diagrams will be as under :



(v) When $-1 < r < 0$ the hypothetical scatter diagrams will be as under :

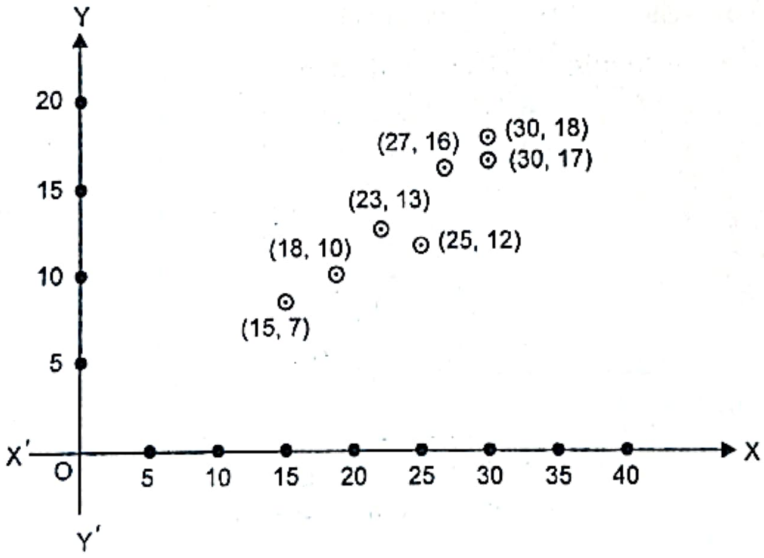


High Degree Negative Correlation *Low Degree Negative Correlation*

Example 2. Draw a scatter diagram to represent the following values of X and Y variables. Comment on the type and degree of correlation.

X :	15	18	30	27	25	23	30
Y :	7	10	17	16	12	13	18

Sol. We plot the points as given in the table as follows :



The diagram reveals a positive correlation. Because the points are too much scattered therefore there is a high degree of positive correlation.

EXERCISE 9.1

1. The following marks have been obtained by a class of 11 students in Economics and Mathematics. Establish Correlation between the two by the method of scatter diagram.

Economics :	85	80	75	70	68	65	60	58	56	55	45
Mathematics :	90	82	70	70	65	64	62	60	48	50	56

2. Given below is the data regarding the prices of potatoes and their sale on a particular shop for first 10 days of a month. Draw a scatter diagram and guess the correlation between the prices and sales of potatoes.

Day of the month	1st,	2nd,	3rd,	4th,	5th,	6th,	7th,	8th,	9th,	10th,
Prices of potatoes Rs.	60,	65,	65,	70,	75,	75,	80,	85,	90,	100
Sale of potatoes kg.	120,	125,	120,	110,	105,	'00,	100,	90,	80,	60

3. Draw a scatter diagram for the data given below and interpret it.

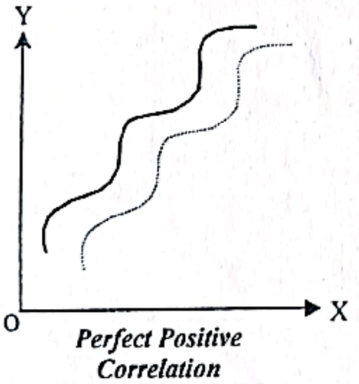
X	10	20	30	40	50	60	70	80
Y	32	20	24	36	40	28	38	44

ANSWERS

1. High degree of positive correlation
2. High degree of negative correlation
3. Low degree of positive correlation

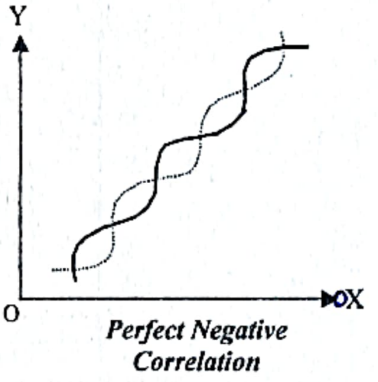
9.8 GRAPHIC METHOD OR CORRELOGRAM

Graphs can also be used to have an idea about the correlation between the two variables. If in a graph, the two curves representing two variables show similar tendency, it is an indication of positive correlation. If on the other hand, two curves move in different directions, correlation is negative. Correlation with the help of two curves can be expressed graphically as under :



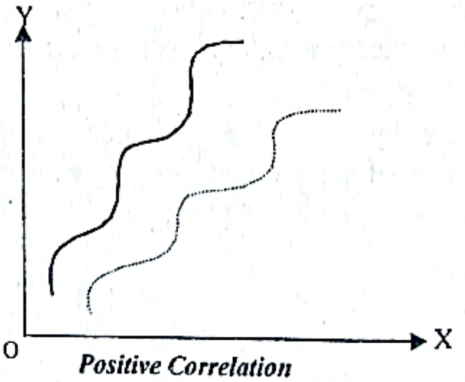
Perfect Positive Correlation

Fig. 9.10



Perfect Negative Correlation

Fig. 9.11



Positive Correlation

Fig. 9.12

Notes (*a*) This method is generally used when we are given time series data.

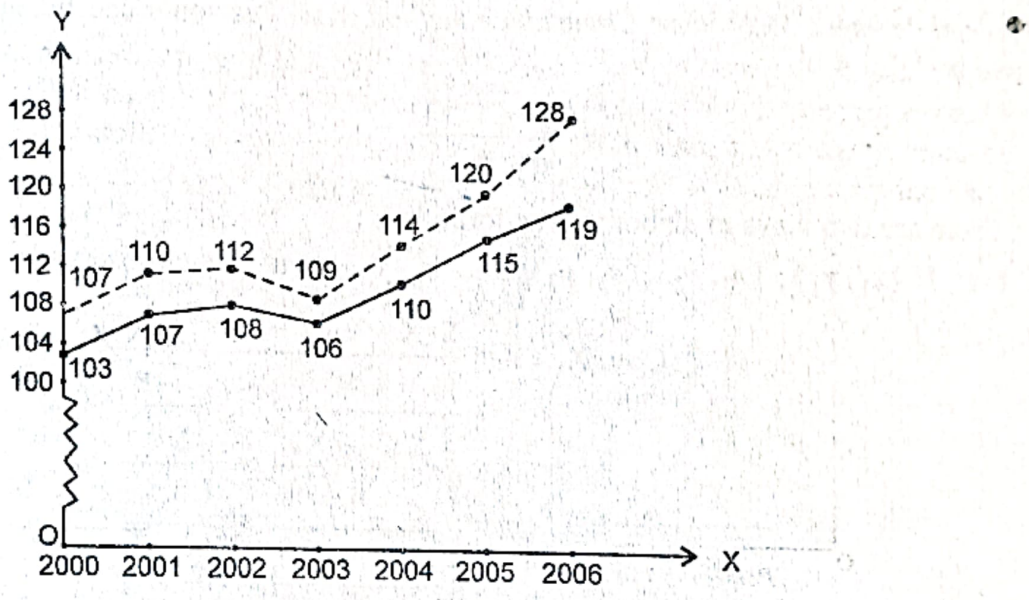
(*b*) As in scatter diagram method this method also does not provide the exact numerical measure of degree of correlation between two variables.

ILLUSTRATIVE EXAMPLES

Example 1. Represent the following by means of graph and comment upon the relationship between demand and supply :

Year :	2000	2001	2002	2003	2004	2005	2006
Demand :	103	107	108	106	110	115	119
Supply :	107	110	112	109	114	120	125

Sol. Let us take years along x-axis. Demand and supply is to be taken along y-axis on the same scale. We plot the points as given in the data and join these points by means of lines. The solid line indicates the graph of year Vs demand and the dotted line indicates the graph of year Vs supply.



From the comparison of two graphs it is clear that there is a very high degree of positive correlation.

EXERCISE 9.2

1. Represent the following by means of correlation graph and comment upon the relationship between volume and value of exports of carpets from India in 1985–86.

Month	April	May	June	July	Aug.	Sept.
Volume (in thousands)	53	80	89	95	56	69
Value in Lakhs of Rs.	22	34	45	50	33	43
Month	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.
Volume (in thousands)	32	60	22	102	60	49
Value in Lakhs of Rs.	23	48	19	83	51	46

2. The following changes in the price indices of A and B Shares were recorded. Calculate correlation between them by graphic method.

Months of a year :	Jan.	Feb.	March	April	May	June
Changes in Share A :	–20	15	12	16	–18	–14
Changes in Share B :	–16	12	10	15	–19	12

ANSWERS

1. High degree of positive correlation 2. Low degree of positive correlation

9.9 KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson's measure of correlation between two series X and Y is a numerical measure of linear relationship between them and is defined as the ratio of covariance between X and Y to the product of the standard deviations of X and Y. The formula is named after famous statistician Karl Pearson and is popularly known as *Karl Pearson's coefficient of correlation* or *product moment correlation coefficient.* This mathematical method of measuring the degree of linear relationship between two variables is known as *Product Moment Method* or *Covariance method.* Karl Pearson's coefficient of correlation is generally denoted by '*r*'.

i.e.
$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \, \sigma_Y} \qquad \ldots(9.1)$$

There are two ways to elaborate this formula :

(*a*) If (x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n) are n pairs of observations of variables *X* and *Y* then

$$\text{Cov}(X, Y) = \frac{\Sigma (X - \overline{X})(Y - \overline{Y})}{n} = \frac{\Sigma xy}{n}$$

$$\sigma_X = \sqrt{\frac{\Sigma (X - \overline{X})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n}}$$

$$\sigma_Y = \sqrt{\frac{\Sigma (Y - \overline{Y})^2}{n}} = \sqrt{\frac{\Sigma y^2}{n}}$$

Where $X - \overline{X} = x$, $Y - \overline{Y} = y$

Substituting these values in (9.1), we have

$$r = \frac{\Sigma (X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma (X - \overline{X})^2} \sqrt{\Sigma (Y - \overline{Y})^2}} \qquad \ldots(9.2)$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}} \qquad \ldots(9.3)$$

(*b*)
$$\text{Cov}(X, Y) = \frac{1}{n} \Sigma (X - \overline{X})(Y - \overline{Y})$$

$$= \frac{1}{n} \Sigma (XY - X\overline{Y} - \overline{X}Y + \overline{X}\,\overline{Y})$$

$$= \frac{\Sigma XY}{n} - \left(\frac{\Sigma X}{n}\right)(\overline{Y}) - \left(\frac{\Sigma Y}{n}\right)(\overline{X}) + \frac{n\,\overline{X}\,\overline{Y}}{n}$$

$$= \frac{\Sigma XY}{n} - \overline{X}\,\overline{Y} - \overline{X}\,\overline{Y} + \overline{X}\,\overline{Y} = \frac{\Sigma XY}{n} - \overline{X}\,\overline{Y}$$

$$= \frac{\Sigma XY}{n} - \left(\frac{\Sigma X}{n}\right)\left(\frac{\Sigma Y}{n}\right) = \frac{1}{n^2}\,[n\Sigma XY - \Sigma X \cdot \Sigma Y]$$

$$\sigma_X^2 = \frac{1}{n}\Sigma (X - \overline{X})^2 = \frac{1}{n}\Sigma [X^2 + \overline{X}^2 - 2X.\overline{X}]$$

$$= \frac{\Sigma X^2}{n} + \frac{n\,\overline{X}^2}{n} - 2\,\overline{X}.\frac{\Sigma X}{n} = \frac{\Sigma X^2}{n} + \overline{X}^2 - 2\,\overline{X}^2$$

$$= \frac{\Sigma X^2}{n} - \overline{X}^2 = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2$$

or $$\sigma_X^2 = \frac{1}{n^2}[n\,\Sigma X^2 - (\Sigma X)^2]$$

or $$\sigma_X = \frac{1}{n}\cdot\sqrt{n\Sigma X^2 - (\Sigma X)^2}$$

Similarly $$\sigma_Y = \frac{1}{n}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}$$

Substituting these values in (9.1)

$$r = \frac{\frac{1}{n^2}[n\,\Sigma XY - \Sigma X.\Sigma Y]}{\frac{1}{n}\cdot\sqrt{n\,\Sigma X^2 - (\Sigma X)^2}\,.\frac{1}{n}\cdot\sqrt{n\,.\Sigma Y^2 - (\Sigma Y)^2}}$$

or $$r = \frac{n\,\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\,\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}} \qquad \ldots(9.4)$$

Note Theoretically, the formulae (9.2) (or (9.3)) and (9.4) for calculating correlation coefficient are equivalent. But in practice, if $\overline{X}$ and $\overline{Y}$ are integers then we shall use formulae (9.2) (or (9.3)) and if $\overline{X}$ and/or $\overline{Y}$ are in fractions then we shall prefer to use formula (9.4).

Properties of Karl Pearson's coefficient of correlation

Karl pearson's coefficient of correlation has the following important properties :

Property 1 : The value of coefficient of correlation lies between – 1 and + 1 *i.e.*, $-1 \le r \le 1$ or $|r| \le 1$.

Proof : Suppose that x and y are the deviations of X and Y series from their respective means $\overline{X}$ and $\overline{Y}$ and σ_X and σ_Y are their respective standard deviations.

Now, consider the quantity $\sum\left(\frac{x}{\sigma_X} \pm \frac{y}{\sigma_Y}\right)^2$

$$= \sum\left(\frac{x^2}{{\sigma_X}^2} + \frac{y^2}{{\sigma_Y}^2} \pm 2\frac{x}{\sigma_X}\frac{y}{\sigma_Y}\right) = \frac{\Sigma x^2}{{\sigma_X}^2} + \frac{\Sigma y^2}{{\sigma_Y}^2} \pm \frac{2\,\Sigma xy}{\sigma_X \sigma_Y}$$

$$= \frac{n{\sigma_X}^2}{{\sigma_X}^2} + \frac{n{\sigma_Y}^2}{{\sigma_Y}^2} \pm \frac{2rn\sigma_X\sigma_Y}{\sigma_X\sigma_Y}$$

$$\left[\because {\sigma_X}^2 = \frac{\Sigma x^2}{n} \Rightarrow \Sigma x^2 = n{\sigma_X}^2,\ {\sigma_Y}^2 = \frac{\Sigma y^2}{n} \Rightarrow \Sigma y^2 = n{\sigma_Y}^2, r = \frac{\Sigma xy}{n\sigma_X\sigma_Y} \Rightarrow \Sigma xy = rn\sigma_X\sigma_Y\right]$$

$$= n + n \pm 2rn = 2n \pm 2rn$$

Since the quantity $\sum\left(\frac{x}{\sigma_X} \pm \frac{y}{\sigma_Y}\right)^2$ involves the sum of squares of the terms so it is always non negative.

i.e. $\sum\left(\frac{x}{\sigma_X} \pm \frac{y}{\sigma_Y}\right)^2 \geq 0$

$\Rightarrow \quad 2n \pm 2rn \geq 0$

$\Rightarrow \quad 2n + 2rn \geq 0 \quad$ or $\quad 2n - 2rn \geq 0$

$\Rightarrow \quad 1 + r \geq 0 \quad$ or $\quad 1 - r \geq 0$

$\Rightarrow \quad r \geq -1 \quad$ or $\quad r \leq 1$

$\Rightarrow \quad -1 \leq r \leq 1 \Rightarrow |r| \leq 1$

Property 2. The coefficient of correlation is independent of change of origin and scale *i.e.* if X and Y are given variables and these are transformed to new variables U and V by using the transformation $U = \frac{X - A}{h}$ and $V = \frac{Y - B}{k}$ where $h\,(>0)$, $k\,(>0)$, A and B are constants then $r_{XY} = r_{UV}$

Proof : Since $U = \frac{X - A}{h}$ and $V = \frac{Y - B}{k}$

$\therefore \quad X = A + hU$ and $Y = B + kV$

Now $\quad \overline{X} = \frac{\Sigma X}{n} = \frac{\Sigma(A + hU)}{n} = \frac{\Sigma A}{n} + \frac{\Sigma hU}{n} = \frac{nA}{n} + h\frac{\Sigma U}{n}$

$= A + h\overline{U}$

Similarly $\quad \overline{Y} = B + k\overline{V}$

So $\quad X - \overline{X} = A + hU - (A + h\,\overline{U}) = h(U - \overline{U})$

and $$Y - \overline{Y} = B + kV - (B - k\overline{V}) = k(V - \overline{V})$$

Now $$r_{XY} = \frac{\Sigma (X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma (X - \overline{X})^2}\sqrt{\Sigma (Y - \overline{Y})^2}}$$

$$= \frac{\Sigma h(U - \overline{U})\, k(V - \overline{V})}{\sqrt{\Sigma h^2 (U - \overline{U})^2}\sqrt{\Sigma k^2 (V - \overline{V})^2}} = \frac{hk\,\Sigma (U - \overline{U})(V - \overline{V})}{h\sqrt{\Sigma (U - \overline{U})^2}\; k\sqrt{\Sigma (V - \overline{V})^2}}$$

$$= \frac{\Sigma (U - \overline{U})(V - \overline{V})}{\sqrt{\Sigma (U - \overline{U})^2}\sqrt{\Sigma (V - \overline{V})^2}} = r_{UV}$$

Hence $$r_{XY} = r_{UV}$$

Property 3 : If the two variables X and Y are linearly related by the equation $aX + bY + c = 0$ then the correlation coefficient between X and Y is 1 or – 1 according as a and b are of opposite or same signs.

Proof : Given $$aX + bY + c = 0 \qquad \ldots(9.5)$$

Taking summation on both sides over n values and dividing by n, we get

$$a\frac{\Sigma X}{n} + b\frac{\Sigma Y}{n} + \frac{nc}{n} = 0$$

$$\Rightarrow \quad a\overline{X} + b\overline{Y} + c = 0 \qquad \ldots(9.6)$$

Subtracting equation (9.6) from equation (9.5), we get,

$$a(X - \overline{X}) + b(Y - \overline{Y}) = 0$$

$$\Rightarrow \quad X - \overline{X} = -\frac{b}{a}(Y - \overline{Y}) \qquad \ldots(9.7)$$

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

$$= \frac{\Sigma -\frac{b}{a}(Y - \overline{Y})(Y - \overline{Y})}{\sqrt{\Sigma\left[-\frac{b}{a}(Y - \overline{Y})\right]^2}\sqrt{\Sigma (Y - \overline{Y})^2}} \qquad \text{[Using equation (9.7)]}$$

$$= \frac{-\frac{b}{a}\Sigma (Y - \overline{Y})(Y - \overline{Y})}{\sqrt{\frac{b^2}{a^2}\Sigma (Y - \overline{Y})^2}\sqrt{\Sigma (Y - \overline{Y})^2}} = \frac{-\frac{b}{a}\Sigma (Y - \overline{Y})^2}{\left|\frac{b}{a}\right|\Sigma (Y - \overline{Y})^2} = \frac{-\left(\frac{b}{a}\right)}{\left|\frac{b}{a}\right|}$$

If a and b are of opposite signs then $\frac{b}{a}$ is negative *i.e.* r is positive and if a and b are of same signs then $\frac{b}{a}$ is positive *i.e.* r is negative.

$$\therefore \quad r = \begin{cases} +1 & \text{if } a \text{ and } b \text{ are of opposite signs} \\ -1 & \text{if } a \text{ and } b \text{ are of same signs.} \end{cases}$$

Property 4 : The coefficient of correlation is the geometric mean of two regression coefficients.

(See Chapter 5 Section 5.6.3)

Property 5 : The coefficient of correlation possesses the property of symmetry *i.e.* $r_{XY} = r_{YX}$

Property 6 : The coefficient of correlation is independent of units of measurement *i.e.* r is a pure number.

Assumptions of Karl Pearson's Coefficient of Correlation

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

(*i*) There exists a Linear Relationship Between Variables : This point can be fully explained with the help of scatter diagram. If the clustering of pairs of points tend to lie around a straight line, the Karl Pearson's coefficient of correlation will give much more dependable and reliable results. In case the pair of values which, on the scatter diagram, rise for some time then remain stationary and ultimately start falling, will give erratic coefficient of correlation if calculated with the help of Karl Pearson's method.

(*ii*) There exists Functional Relationship Between the Variables : Karl Pearson's formula is based on another assumption that there is a cause and effect relationship between the variables. In the absence of such a relation the correlation will be considered as meaningless.

(*iii*) The Variables Tend to be Normally Distributed : According to Karl Pearson, "the size of the complex organs (something measurable) are determined by a great variety of independent contributing causes, for example, climate, nourishment, physical training and innumerable other causes which cannot be individually observed or their effects measured". It is further observed that, "the variations in intensity of the contributory causes are small as compared with their absolute intensity and these variations follow the normal law of distribution".

Interpretation of Karl Pearson's coefficient of correlation

Karl Pearson's coefficient of correlation determines the direction and degree of relationship between two variables. In words of W.M. Harper, "The interpretation of the coefficient of correlation depends very much on experience. The full significance of r will be grasped only after working on a number of correlation problems and seeing the kinds of data which gives rise to various values of r".

The following general points should be kept in mind while interpreting an observed value of coefficient of correlation r :

(*i*) If $r = +1$ then there is a perfect positive correlation between the variables. In this case, the scatter diagram will be a straight line starting from left bottom and rising upward to the right top. (See fig. 9.1)

(*ii*) If $r = -1$ then there is a perfect negative correlation between the variables. In this case, the scatter diagram will be a straight line starting from left top and falling downward to the right bottom (See Fig. 9.2).

(*iii*) If $r = 0$ then there is no linear relationship between the variables *i.e.*, the variables are uncorrelated. In this case, the scatter diagram will not be a straight line. (See Fig. 9.7, 9.8, 9.9)

It should be noted that $r = 0$ does not imply that the variables are independent.

In other words we can say that $r = 0$ implies the absence of linear relationship between the variables but the variables may be related in some other form say quadratic, trigonometric or logarithmic form etc.

(*iv*) For other values of r between – 1 and 1 except 0, there are no set guidelines for interpretation of r. We can at the most say that if the value of r is very close to 1 (or – 1), there is a high degree of positive (or negative) correlation between the variables and if r is very close to zero then there is a very low degree of positive or negative correlation between the variables depending upon the sign of r.

The closeness of relationship between the two variables is not proportional to r. *i.e.* $r = 0.6$ does not indicate that the relationship between the variables is twice as close as that for $r = 0.3$.

So one should be very careful in interpreting the values of r other than – 1, 0 and 1.

Merits and Demerits of Karl Pearson's Method

Merits : The following are the merits of this method :

1. It is most widely used mathematical method to determine correlation.
2. It measures the degree as well as direction of correlation.
3. It is useful in further statistical analysis.

Demerits : The following are the demerits of this method :

1. As compared to other methods it is time consuming method.
2. It is based on complex calculation.
3. It is based on unrealistic assumptions.
4. It is highly affected by presence of extreme values.
5. It can not be applied to qualitative phenomena *e.g.* beauty, intelligence, honesty etc.

9.9.1 CALCULATION OF KARL PEARSON'S COEFFICIENT OF CORRELATION

The methods for calculating Karl Pearson's coefficient of correlation can be broadly divided into two categories as follows :

(*a*) Direct Methods (*b*) Short Cut Methods

(*a*) Direct Methods :

(*i*) If the arithmetic means of two series are integers *i.e.*, $\overline{X}$ and $\overline{Y}$ are integers then we shall use the formula (9.3)

i.e.
$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \quad \text{where } x = X - \overline{X} \text{ and } y = Y - \overline{Y}$$

(*ii*) If the arithmetic means of two series are fractions *i.e.* $\overline{X}$ and $\overline{Y}$ are in fractions then we shall use the formula (9.4)

i.e.
$$r = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

(*b*) Short Cut Methods :

If $\overline{X}$ and $\overline{Y}$ are in fractions and X and Y assume large values then the computation of r using direct methods become quite tedious. In such cases, we can conveniently change the origin (and scale if necessary) in X and/or Y because the coefficient of correlation r is independent of change of origin and scale. If only origin is changed *i.e.*, deviations of the variables X and Y are taken from assumed means A and B respectively then

$$\text{Cov}(X,Y) = \frac{\Sigma\, dxdy}{n} - \frac{\Sigma dx}{n}\frac{\Sigma dy}{n},$$

$$\sigma_X = \sqrt{\frac{\Sigma dx^2}{n} - \left(\frac{\Sigma dx}{n}\right)^2}, \sigma_Y = \sqrt{\frac{\Sigma dy^2}{n} - \left(\frac{\Sigma dy}{n}\right)^2}$$

$$\therefore \quad r = \frac{Cov(X,Y)}{\sigma_X\, \sigma_Y} \text{ i.e. } r = \frac{n\Sigma dx\, dy - \Sigma dx\, \Sigma dy}{\sqrt{n\Sigma dx^2 - (\Sigma dx)^2}\sqrt{n\Sigma dy^2 - (\Sigma dy)^2}} \quad ...(9.8)$$

where $dx = X - A$ and $dy = Y - B$

Further if scale is also changed *i.e.* if either in X-series or Y-series or in both, step deviations are taken then the formula for r becomes

$$r = \frac{n\Sigma d'x\, d'y - \Sigma d'x\, \Sigma d'y}{\sqrt{n\Sigma d'x^2 - (\Sigma d'x)^2}\sqrt{n\Sigma d'y^2 - (\Sigma d'y)^2}} \quad ...(9.9)$$

where $d'x = \frac{dx}{c_1}$ and $d'y = \frac{dy}{c_2}$

Algorithm

Remarks : *Here, one-dimensional array X of size n is used to hold n values $x_1, x_2, x_3 \ldots\ldots, x_n$ of X-series, one-dimensional array Y of size n is used to hold corresponding n values $y_1, y_2, y_3 \ldots\ldots, y_n$ of Y-series and the coefficient of correlation is calculated using the formula* $r = \frac{n\Sigma XY - \Sigma X\, \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$

Start

read : n

for $i = 1$ to n by 1 do

read : X_i, Y_i

endfor

set $sx = 0$

set $sy = 0$

set $sxy = 0$

set $sx2$

set $sy2 = 0$

for $i = 1$ to n by 1 do :

set $sx = sx + X_i$

set $sy = sy + Y_i$

set $sxy = sxy + X_i * Y_i$

set $sx2 = sx2 + X_i * X_i$

set $sy2 = sy2 + Y_i * Y_i$

endfor

set $r = (n * sxy - sx * sy) / (\text{sqrt}\ (n * sx2 - sx * sx) * \text{sqrt}\ (n * sy2 - sy * sy))$

write : "coefficient of correlation = ", r

exit

CHECKPOINTS

1. What is meant by Correlation coefficient ?

(G.N.D.U. B.Sc. C.Sc. April 2006, Sept. 2006)

2. The values of correlation coefficient ranges from – 1 to +1. Why ?

(G.N.D.U. B.Sc. I.T. April 2009)

3. Show that coefficient of correlation is independent of change of scale and origin.

(G.N.D.U. B.C.A. April 2002)

4. The value of coefficient of correlation lies between –1 and +1. How will you interpret the data if coefficient of correlation is –1, 0, +1 respectively for three different sets of data ?

(P.U. B.C.A. Sept. 2001)

5. What are the underlying assumptions of Karl Pearson's coefficient of correlation ?

6. What are the merits and demerits of Karl Pearson's method of calculating correlation ?

ILLUSTRATIVE EXAMPLES

Example 1. The coefficient of correlation between two variables X and Y is 0.48. The covariance is 36. The variance of X is 16. Find the standard deviation of Y.

Sol. Given $r = 0.48$, $\text{cov}\ (X, Y) = 36$, $\sigma_X^2 = 16$ or $\sigma_X = 4$

Using $r = \dfrac{\text{Cov}\ (X, Y)}{\sigma_X \sigma_Y}$ we have, $\dfrac{48}{100} = \dfrac{36}{4\,\sigma_Y}$

or

$$\sigma_Y = \frac{900}{48} = 18.75$$

Example 2. From the following data, find the value of n

$r = 0.5,\ \Sigma xy = 120,\ \sigma_Y = 8,\ \Sigma x^2 = 90$

where 'x' and 'y' are deviations from arithmetic averages of variables X and Y.

Sol. Given $r = 0.5,\ \Sigma xy = 120,\ \sigma_Y = 8,\ \Sigma x^2 = 90$

We know that $r = \dfrac{\Sigma xy}{n\sigma_X\,\sigma_Y}$

i.e. $$r = \frac{\Sigma xy}{n \times \sqrt{\dfrac{\Sigma x^2}{n}} \times \sigma_Y} \qquad \left\{\because\ \sigma_X = \sqrt{\frac{\Sigma x^2}{n}}\right\}$$

i.e. $$\frac{5}{10} = \frac{120}{\sqrt{n}\,\sqrt{90} \times 8} \quad \textit{i.e.} \quad \frac{1}{2} = \frac{15}{\sqrt{90n}}$$

Squaring both sides and cross multiplying, we get

$$90\,n = 900$$

i.e. $$n = 10$$

Example 3. From the following data, calculate the coefficient of correlation between the X and Y series.

	X-series	Y-series
Number of items	15	15
Arithmetic mean	25	18
Sum of Squares of deviations from mean	136	138

Summation of product of deviations of X and Y series from their respective arithmetic means is 122.

(P.U. B.C.A. Sept. 2005)

Sol. Given $n = 15,\ \overline{X} = 25,\ \overline{Y} = 18,\ \Sigma (X - \overline{X})^2 = 136,\ \Sigma (Y - \overline{Y})^2 = 136$ and $\Sigma (X - \overline{X})(Y - \overline{Y}) = 122$

Now, $$\text{Cov}(X, Y) = \frac{1}{n}\Sigma (X - \overline{X})(Y - \overline{Y}) = \frac{1}{15} \times 122 = 8.13$$

$$\sigma_X = \sqrt{\frac{1}{n}\Sigma (X - \overline{X})^2} = \sqrt{\frac{1}{15}(136)} = 3.01$$

and $$\sigma_Y = \sqrt{\frac{1}{n}\Sigma (Y - \overline{Y})^2} = \sqrt{\frac{1}{15}(138)} = 3.03$$

$\therefore$ Coefficient of correlation between X and Y is given by

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X\,\sigma_Y} = \frac{8.13}{(3.01)(3.03)} = 0.89$$

Example 4. Calculate coefficient of correlation between X and Y for the following data :

X :	1	3	4	5	7	8	10
Y :	2	6	8	10	14	16	20

(G.N.D.U. B.C.A. April 2005)

Sol. First, we prepare the following table :

X	Y	X^2	Y^2	XY
1	2	1	4	2
3	6	9	36	18
4	8	16	64	32
5	10	25	100	50
7	14	49	196	98
8	16	64	256	128
10	20	100	400	200
$\Sigma X = 38$	$\Sigma Y = 76$	$\Sigma X^2 = 264$	$\Sigma Y^2 = 1056$	$\Sigma XY = 528$

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

$\therefore$
$$r = \frac{7(528) - (38)(76)}{\sqrt{7(264) - (38)^2}\,\sqrt{7(1056) - (76)^2}} = \frac{808}{\sqrt{404}\,\sqrt{1616}} = 1$$

Example 5. What will be the co-efficient of correlation corresponding to X = 1, 2, 3, 4, 5 if $Y = (X - 6)^5$ is a functional relationship between X and Y. Explain why the answer differs from unity.

Sol. First, we prepare the following table :

X	$Y = (X-6)^5$	$x = X - 3$	$y = Y + 885$	x^2	y^2	xy
1	− 3125	− 2	− 2240	4	5017600	4480
2	− 1024	− 1	− 139	1	19321	139
3	− 243	0	642	0	412164	0
4	− 32	1	853	1	727609	853
5	− 1	2	884	4	781456	1768
$\Sigma X = 15$	$\Sigma Y = -4425$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 10$	$\Sigma y^2 = 6958150$	$\Sigma xy = 7240$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{15}{5} = 3, \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{-4425}{5} = -885$$

$\therefore$
$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{7240}{\sqrt{10 \times 6958150}} = \frac{7240}{8341.55} = 0.7$$

Answer differs from unity because there is no linear relationship between variables X and Y.

Example 6. Calculate Karl Pearson's coefficient of correlation for the data given below, taking 63 and 66 as assumed means of x and y respectively :

x :	60	62	64	66	68	70	72
y :	61	63	63	63	64	65	62

(G.N.D.U. B.C.A. Sept. 2007)

Sol. First, we prepare the following table :

x	y	A = 63 $dx = x - A$	B = 66 $dy = y - B$	dx^2	dy^2	$dx\,dy$
60	61	−3	−5	9	25	15
62	63	−1	−3	1	9	3
64	63	1	−3	1	9	−3
66	63	3	−3	9	9	−9
68	64	5	−2	25	4	−10
70	65	7	−1	49	1	−7
72	62	9	−4	81	16	−36
		$\Sigma dx = 21$	$\Sigma dy = -21$	$\Sigma dx^2 = 175$	$\Sigma dy^2 = 73$	$\Sigma dx\,dy = -47$

Now, Karl Pearson's coefficient of correlation is given by

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

$$\therefore \quad r = \frac{7(-47) - 21(-21)}{\sqrt{7(175) - (21)^2}\ \sqrt{7(73) - (-21)^2}}$$

$$= \frac{112}{\sqrt{784}\ \sqrt{70}}$$

$$= 0.48$$

Example 7. Find the correlation coefficient between age and playing habits from the following data :

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

What conclusion would you draw from your answer ?

(P.U. B.C.A. April 2002)

Sol. First, we prepare the following table :

Age (X)	No. of Students	Regular players	% No. of players (Y)	$dx = X - 17$	$dy = Y - 40$	$d'y = dy/5$	dx^2	$d'y^2$	$dx\ d'y$
15	250	200	80	– 2	40	8	4	61	– 16
16	200	150	75	– 1	35	7	1	49	– 07
17	150	90	60	0	20	4	0	16	00
18	120	48	40	1	00	0	1	00	00
19	100	30	30	2	– 10	– 2	4	4	– 04
20	80	12	15	3	– 25	– 5	9	25	– 15
				$\Sigma dx = 3$		$\Sigma d'y = 12$	$\Sigma dx^2 = 19$	$\Sigma dy^2 = 158$	$\Sigma dx\ d'y = -42$

The coefficient of correlation is given by

$$r = \frac{n \cdot \Sigma dx\ d'y - \Sigma dx \cdot \Sigma d'y}{\sqrt{n \cdot \Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{n \cdot \Sigma d'y^2 - (\Sigma d'y)^2}}$$

$$\therefore \quad r = \frac{6(-42) - (3)(12)}{\sqrt{6(19) - 9}\ \sqrt{6(158) - 144}} = \frac{-288}{(10.25)\ (28.35)} = -0.99$$

This shows that there exists a very high degree of negative correlation between age and playing habits which means that as age increases playing habits decrease.

Example 8. The following table gives the distribution of items 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	15 – 16	16 – 17	17 – 18	18 – 19	19 – 20	20 – 21
No. of items	400	540	680	720	800	600
No. of defective items	300	324	340	360	360	240

(G.N.D.U. B.C.A. April 2006)

Sol. First, we prepare the following table :

Size	No. of Items	No. of def. Items	M.V. (X)	$dx =$ X – 17·5	% No. of def. Items (Y)	$dy =$ Y – 50	$d'y = dy/5$	dx^2	$d'y^2$	$dx\ d'y$
15 – 16	400	300	15·5	– 2	75	25	5	4	25	– 10
16 – 17	540	324	16·5	– 1	60	10	2	1	04	– 02
17 – 18	680	340	17·5	0	50	0	0	0	00	00
18 – 19	720	360	18·5	1	50	0	0	1	00	00
19 – 20	800	360	19·5	2	45	– 5	– 1	4	01	– 02
20 – 21	600	240	20·5	3	40	– 10	– 2	9	04	– 06
				$\Sigma dx = 3$			$\Sigma d'y = 4$	$\Sigma dx^2 = 19$	$\Sigma d'y^2 = 34$	$\Sigma dx d'y = -20$

The coefficient of correlation is given by

$$r = \frac{n \cdot \Sigma dx\, d'y - \Sigma dx \cdot \Sigma d'y}{\sqrt{n \cdot \Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{n \cdot \Sigma d'y^2 - (\Sigma d'v)^2}}$$

$$\therefore \quad r = \frac{6(-20) - 3 \times 4}{\sqrt{6(19) - (3)^2}\ \sqrt{6(34) - (4)^2}} = \frac{-132}{(10.25)\ (13.71)} = -0.94$$

Clearly there is a high degree of negative correlation between size and defect in quality.

EXERCISE 9.3

1. Coefficient of correlation between two variables X and Y is 0.8 and their co-variance is 29. If standard deviation of Y is 5.196, find the S.D. of X.
2. Coefficient of correlation between X and Y variables is 0.25, the co-variance is 7.5 and variance of X is 9. Find the standard deviation of Y.
3. If co-variance between X and Y variables is 6.1 and the variance of X and Y are respectively 5.4 and 9.6, find the coefficient of correlation.
4. For variates X and Y having 10 pairs of observations, the following calculations are made :

   $\Sigma X = 250, \Sigma Y = 300, \Sigma XY = 7900, \Sigma X^2 = 6500, \Sigma Y^2 = 10000$

   Find coefficient of correlation between X and Y.

   (P.U. B.C.A. Sept. 2008)

5. The following results are obtained between two series from their respective means. Compute the co-efficient of correlation.

	X – Series	Y – Series
Number of Items	7	7
Arithmetic mean	4	8
Sum of square of deviations from Arithmetic Mean	28	76

   Summation of product of deviations of X and Y series from their respective arithmetic means = 46.

6. Given the following data :

	X-Series	Y-Series
Assumed Mean	41	32
Sum of deviations from assumed mean	– 170	– 20
Sum of the squares of deviations from assumed mean	8180	2290

   Sum of the product of deviations from the assumed means = 3480

   Number of pairs of observations = 10

   Calculate Karl Pearson's Coefficient of correlation.

7. Given, total of product of deviations of X and Y series = 3044,

Number of pairs of observations = 10, Total of deviations of X series = – 170,

Total of deviations of Y series = –20, Total of square of deviation of X and Y series are 8288 and 2264 respectively.

Find out coefficient of correlation when arbitrary means of X and Y series are 82 and 68 respectively.

8. Calculate the coefficient of correlation for the following data :

X :	2	4	5	6	8	11
Y :	18	12	10	8	7	5

(G.N.D.U. B.Sc. I.T. April 2009)

9. If X and Y are connected by the relation $Y = 15 - X^2$ what will be the coefficient of correlation between X and Y for X = – 3, – 2, – 1, 0, 1, 2, 3. Will it give a similar value of the coefficient of correlation for all integral values of X ?

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

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

(G.N.D.U. B.C.A. April 2009)

11. Ten students got the following percentage of marks in Principles of Economics and Statistics.

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

Calculate the coefficient of correlation.

12. Find the coefficient of correlation from the following data and interpret the result :

X	300	350	400	450	500	550	600	650	700
Y	800	900	1000	1100	1200	1300	1400	1500	1600

13. The deviations from their means of two series, X and Y, are respectively given below :

x :	– 4	– 3	– 2	– 1	0	1	2	3	4
y :	3	– 3	– 4	0	4	1	2	– 2	– 1

Calculate Karl Pearson's coefficient of correlation and interpret the result.

14. Calculate Karl Pearson's Co-efficient of Correlation from the following data using 44 and 26 respectively as the origins of X and Y

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

15. Calculate Karl Pearson's coefficient of correlation of the following data relating to price and demand :

(43, 105), (54, 98), (85, 53), (91, 49), (59, 84), (95, 40), (68, 73), (79, 59), (73, 63), (77, 52)

What do you conclude ? (P.U. B.C.A. Sept. 2002)

16. Calculate Karl Pearson's correlation coefficient for the following data :

X :	8	12	15	20	24	27	32
Y :	30	34	36	44	56	64	72

(P.U. B.C.A. April 2007)

17. Obtain the correlation coefficient for the following data :

x :	30	40	70	14	25	66	03	50	27	30
y :	50	40	30	62	48	30	12	46	51	31

(G.N.D.U. B.Sc. C.Sc. Sept. 2007)

18. Calculate the coefficient of correlation between X and Y :

X :	150	153	154	155	157	160	163	164
Y :	65	66	67	70	68	53	70	63

(G.N.D.U. B.Sc. C.Sc. April 2007)

19. From the data given below, calculate the Karl Pearson's coefficient of correlation :

X	28	41	40	38	35	33	40	32	36	39
Y	23	34	33	34	30	26	28	31	36	38

(P.U. B.C.A. April 2006)

20. Calculate Karl Pearson's coefficient from the following data :

X :	23	27	28	28	29	30	31	33
Y :	18	20	20	27	21	29	27	29

(G.N.D.U. B.Sc. C.Sc. Sept. 2006)

21. Calculate Karl Pearson's co-efficient of correlation for the following data :

X :	1	2	3	4	5	6
Y :	20	35	60	100	120	135

22. Calculate co-efficient of correlation for the following data :

X :	10	17	28	22	27	30
Y :	25	35	45	55	65	75

23. The following are the results of some examination :

Age of candidates	13 – 14	14 – 15	15 – 16	16 – 17	17 – 18
Candidates appeared	200	300	100	50	150
Successful candidates	124	180	65	34	99
Age of candidates	18 – 19	19 – 20	20 – 21	21 – 22	22 – 23
Candidates appeared	400	250	150	25	75
Successful candidates	252	145	81	12	33

Calculate coefficient of correlation between age and successful candidates.

ANSWERS

1. 6.98	2. 10	3. 0.85	4. 0.8
5. 0.997	6. 0.91	7. 0.78	8. – 0.92
9. 0, No	10. 0.95	11. 0.78	
12. 1, Perfectively Positive Correlation		13. 0, No Correlation	14. – 0.73
15. – 0.98	16. 0.98	17. – 0.86	18. – 0.21
19. 0.67	20. 0.80	21. 0.99	22. 0.865
23. – 0.77			

9.9.2 CORRECTING INCORRECT COEFFICIENT OF CORRELATION

The following examples illustrate the procedure of correcting incorrect coefficient of correlation.

ILLUSTRATIVE EXAMPLES

Example 1. A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results.

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

It was, however, discovered at the time of checking that two pairs of observations were not correctly copied. They were taken as (6, 14) and (8, 6) while the correct values were (8, 12) and (6, 8). Prove that the correct value of the correlation coefficient should be 2/3.

(P.U. B.C.A. April 2003)

Sol. Using the given information, we calculate the correct values as follows :

Incorrect $\Sigma X = 125$

$\therefore$ Correct $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Incorrect $\Sigma Y = 100$

$\therefore$ Correct $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Incorrect $\Sigma X^2 = 650$

$\therefore$ Correct $\Sigma X^2 = 650 - (6)^2 + (8)^2 + (8)^2 + (6)^2 = 650$

Incorrect $\Sigma y^2 = 460$

$\therefore$ Correct $\Sigma Y^2 = 460 - (14)^2 - (6)^2 + (12)^2 + (8)^2 = 436$

Incorrect $\Sigma XY = 508$

$\therefore$ Correct $\Sigma XY = 508 - (6)(14) - (8)(6) + (8)(12) + (6)(8) = 520$

The correct value of the coefficient of correlation is given by

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

$$= \frac{(25)(520)-(125)(100)}{\sqrt{(25)(650)-(125)^2}\times\sqrt{(25)(436)-(100)^2}} = \frac{500}{\sqrt{625\times 900}} = \frac{500}{(25)(30)} = \frac{2}{3}$$

Example 2. In two sets of variables X and Y with 50 observations each, the following data were observed

$\overline{X} = 10, \sigma_X = 3, \ \overline{Y} = 6, \ \sigma_Y = 2, r_{XY} = 0.3.$

However on subsequent verification, it was found that one value of X = 10 and one value of Y = 6 were inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of the correlation co-efficient affected ?

Sol. Using the given information, we calculate the correct values as follows :

Using $\overline{X} = \frac{\Sigma X}{n}$

We have $10 = \frac{\Sigma X}{50}$

So Incorrect $\Sigma X = 500$

$\therefore$ Correct $\Sigma X = 500 - 10 = 490$

Now $\sigma_X^2 = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2$

$\therefore \quad 9 = \frac{\Sigma X^2}{50} - (10)^2$

So Incorrect $\Sigma X^2 = 5450$

$\therefore$ Correct $\Sigma X^2 = 5450 - (10)^2 = 5350$

Using $\overline{Y} = \frac{\Sigma Y}{n}$

We have $6 = \frac{\Sigma Y}{50}$

Incorrect $\Sigma Y = 300$

Correct $\Sigma Y = 300 - 6 = 294$

Now $\sigma_Y^2 = \frac{\Sigma Y^2}{n} - \left(\frac{\Sigma Y}{n}\right)^2$

$\therefore \quad 4 = \frac{\Sigma Y^2}{50} - (6)^2$

So Incorrect $\Sigma Y^2 = 2000$

$\therefore$ Correct $\Sigma Y^2 = 2000 - (6)^2 = 1964$

Now $r = \frac{n\cdot\Sigma XY - \Sigma X\cdot\Sigma Y}{\sqrt{n\cdot\Sigma X^2 - (\Sigma X)^2}\,\sqrt{n\cdot\Sigma Y^2 - (\Sigma Y)^2}}$

Substituting the incorrect values, we have

$$\frac{3}{10} = \frac{50\Sigma XY - (500)(300)}{\sqrt{50(5450)-(500)^2}\,\sqrt{50(2000)-(300)^2}} = \frac{50\Sigma XY - 150000}{\sqrt{22500}\,\sqrt{10000}} = \frac{50\Sigma XY - 150000}{150\times 100}$$

$\therefore \quad \frac{3}{10} = \frac{\Sigma XY - 3000}{300}$ *i.e.*, $3 = \frac{\Sigma XY - 3000}{30}$

So Incorrect $\Sigma XY = 3090$

$\therefore$ Correct $\Sigma XY = 3090 - (10)(6) = 3030$

$\therefore$ The correct value of coefficient of correlation is given by

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

$$= \frac{49(3030) - (490)(294)}{\sqrt{49(5350) - (490)^2}\sqrt{49(1964) - (294)^2}} = \frac{4410}{\sqrt{22050}\sqrt{9800}} = \frac{4410}{(148.49)(98.99)} = 0.3$$

which is the same as the old value of r.

EXERCISE 9.4

1. Given $n = 10$, $\Sigma X = 25$, $\Sigma Y = 5$, $\Sigma X^2 = 650$, $\Sigma Y^2 = 290$, $\Sigma XY = 335$

   On verification it was found that the pair ($X = 10$, $Y = 4$) was copied wrongly, the correct values being ($X = 12$, $Y = 9$). Find the correct value of correlation.

2. A computer while calculating correlation coefficient between two variables X and Y from 30 pairs of observations obtained the following results :

   $n = 30$ $\quad \Sigma X = 120$ $\quad \Sigma Y = 90$

   $\Sigma XY = 356$ $\quad \Sigma X^2 = 600$ $\quad \Sigma Y^2 = 250$

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

X	Y
8	10
12	7

   while the correct values are

X	Y
8	12
10	8

   Obtain the correct value of correlation coefficient.

3. While calculating the Co-efficient of Correlation between X and Y, following results are obtained :

   $n = 30$, $\Sigma X = 125$, $\Sigma Y = 100$, $\Sigma X^2 = 740$, $\Sigma Y^2 = 439$, $\Sigma XY = 475$.

   Later on it was discovered that two points of values were copied down as (8, 10) and (8, 6), while the correct values were (8, 11) and (6, 8). Find the correct coefficient of correlation.

4. In two sets of variables of X and Y with 20 observations each, the following data were observed :

   $\bar{X} = 15$ $\quad$ Standard deviation of $X = 4$

   $\bar{Y} = 18$ $\quad$ Standard deviation of $Y = 5$

   Coefficient of correlation between X and Y is 0.7. However, on subsequent verification it was found that one value of X (= 12) and one value Y (= 25) were inaccurate where as the correct values were $X = 21$ and $Y = 15$. How is the original value of correlation coefficient affected ?

5. Co-efficient of correlation between X and Y for 20 items is 0·3 ; mean of X is 15 and that of Y is 20, standard deviations are 4 and 5 respectively. At the time of calculation one item 27 was wrongly taken as 17 in case of X and 35 instead of 30 in Y – series. Find the correct Co-efficient of correlation. (G.N.D.U. B.C.A. April 2007)

ANSWERS

1. 0.81	2. 0.05	3. 0.36	4. 0.73	5. 0.52

9.9.3 KARL PEARSON'S COEFFICIENT OF CORRELATION IN GROUPED SERIES

In grouped or bivariate frequency series Karl Pearson's method can also be used to calculate correlation coefficient. The following methods are used to calculate coefficient of correlation :

(*i*) Direct Method : ***When deviations are taken from actual mean***

$$r_{XY} = \frac{\Sigma f(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma f(X-\overline{X})^2}\sqrt{\Sigma f(Y-\overline{Y})^2}} \quad ...(9.10)$$

(*ii*) Short-cut Method : ***When deviations are taken from assumed mean***

$$r_{XY} = \frac{N\Sigma fdx\,dy - \Sigma fdx\Sigma fdy}{\sqrt{N\Sigma fdx^2 - (\Sigma fdx)^2}\sqrt{N\Sigma fdy^2 - (\Sigma fdy)^2}} \quad ...(9.11)$$

where $N = \Sigma f$

When some common factor can be taken from dx or dy or both, we can use the following formula to calculate correlation coefficient :

$$r_{XY} = \frac{N\Sigma fd'x\,d'y - \Sigma fd'x\Sigma fd'y}{\sqrt{N\Sigma fd'x^2 - (\Sigma fd'x)^2}\sqrt{N\Sigma fd'y^2 - (\Sigma fd'y)^2}} \quad ...(9.12)$$

Note After taking common factor from dx or dy or both we do not multiply with common factor in formula because coefficient of correlation is independent of change of scale.

Generally we make use of short-cut method in practical problems for the sake of simplicity of calculations. The different steps involved in short-cut method are as follows :

(*i*) If data are given in class intervals take the mid values of the class intervals as X and Y series.

(*ii*) Take deviations in both series from assumed means. These deviations are represented by dx and dy. Step deviations may be used and represented as $d'x$ and $d'y$.

(*iii*) The frequencies of X series are multiplied by the corresponding values of $d'x$ and the product so obtained are added to find $\Sigma f\,d'x$. Similarly the frequencies in Y series are operated to find the values of $\Sigma f\,d'y$.

(*iv*) Take square of $d'x$. Multiply the frequency of X series by corresponding value of $d'x^2$ and take summation to get $\Sigma f\,d'x^2$. Similarly find out $\Sigma f\,d'y^2$.

(*v*) Now multiply cell frequencies with corresponding value of $d'x$ and $d'y$. Put this value $f\,d'x\,d'y$ in upper right hand corner of each cell.

(*vi*) Add all the values of $f\,d'x\,d'y$ to get $\Sigma f\,d'x\,d'y$.

(*vii*) All the above values are put in the formula to find the Karl Pearson's coefficient of correlation.

ILLUSTRATIVE EXAMPLES

Example 1. Find the coefficient of correlation between ages of 100 mothers and daughters :

Age of mothers in years (X)	Age of daughters in years (Y)					Total
	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	
15 – 25	6	3	–	–	–	9
25 – 35	3	16	10	–	–	29
35 – 45	–	10	15	7	–	32
45 – 55	–	–	7	10	4	21
55 – 65	–	–	–	4	5	9
Total	9	29	32	21	9	100

(G.N.D.U. B.Sc. I.T. April 2007)

Sol. First, we prepare the following table :

			$d'y = dy/5$	–2	–1	0	1	2				
			$dy = m' - 17.5$	–10	–5	0	5	10				
			M.V. (m')	7.5	12.5	17.5	22.5	27.5				
$d'x = dx/10$	$dx = m - 40$	M.V. (m)	Age of Mothers (X) ↓ / Age of Daughters (Y) →	5–10	10 – 15	15 – 20	20 – 25	25 – 30	Total (f)	$fd'x$	$fd'x^2$	$fd'x\,d'y$
–2	–20	20	15 – 25	24 (6)	6 (3)	0 (–)	0 (–)	0 (–)	9	–18	36	30
–1	–10	30	25–35	6 (3)	16 (16)	0 (10)	0 (–)	0 (–)	29	–29	29	22
0	0	40	35 – 45	0 (–)	0 (10)	0 (15)	0 (7)	0 (–)	32	0	0	0
1	10	50	45 – 55	0 (–)	0 (–)	0 (7)	10 (10)	8 (4)	21	21	21	18
2	20	60	55 – 65	0 (–)	0 (–)	0 (–)	8 (4)	20 (5)	9	18	36	28
			Total (f)	9	29	32	21	9	N = 100	$\Sigma fd'x = -8$	$\Sigma fd'x^2 = 122$	$\Sigma fd'x\,d'y = 98$
			$fd'y$	–18	–29	0	21	18	$\Sigma fd'y = 8$			
			$fd'y^2$	36	29	0	21	36	$\Sigma fd'y^2 = 122$			
			$fd'x\,d'y$	30	22	0	18	28	$\Sigma fd'x\,d'y = 98$			

The coefficient of correlation is given by

$$r = \frac{N\Sigma fd'x\,d'y - \Sigma fd'x\,\Sigma fd'y}{\sqrt{N\Sigma fd'x^2 - (\Sigma fd'x)^2}\,\sqrt{N\Sigma fd'y^2 - (\Sigma fd'y)^2}}$$

$$= \frac{100(98) - (-8)(-8)}{\sqrt{100(22) - (-8)^2}\,\sqrt{100(122) - (-8)^2}} = \frac{9736}{\sqrt{12136}\,\sqrt{12136}} = 0.80$$

Example 2. The following table gives the number of candidates obtaining marks in Economics and Statistics :

Marks in Economics (X)	Marks in Statistics (Y)				
	20 – 30	30 – 40	40 – 50	50 – 60	Total
20 – 30	3	1	1	–	5
30 – 40	2	6	1	2	11
40 – 50	1	2	2	1	6
50 – 60	–	1	1	1	3
Total	6	10	5	4	25

Calculate coefficient of correlation between marks in Statistics and marks in Economics.

Sol. First, we prepare the following table :

			$d'y = dy/10$	−2	−1	0	1				
			$dy = m' - 45$	−20	−10	0	10				
			M.V. (m')	25	35	45	55				
$d'x = dx/10$	$dx = m-45$	M.V. (m)	Eco. ↓ Stat →	20–30	30–40	40–50	50–60	Total	$fd'x$	$fd'x^2$	$fd'x\,d'y$
−2	−20	25	20–30	12 / 3	2 / 1	0 / 1	0 / –	5	−10	20	14
−1	−10	35	30–40	4 / 2	6 / 6	0 / 1	−2 / 2	11	−11	11	8
0	0	45	40–50	0 / 1	0 / 2	0 / 2	0 / 1	6	0	0	0
1	10	55	50–60	0 / –	−1 / 1	0 / 1	1 / 1	3	3	3	0
			Total	6	10	5	4	N = 25	$\Sigma fd'x = -18$	$\Sigma fd'x^2 = 34$	$\Sigma fd'x\,d'y = 22$
			$fd'y$	−12	−10	0	4	$\Sigma fd'y = -18$			
			$fd'y^2$	24	10	0	4	$\Sigma fd'y^2 = 38$			
			$fd'x\,d'y$	16	7	0	−1	$\Sigma fd'x\,d'y = 22$			

The coefficient of correlation is given by

$$r = \frac{N\Sigma fd'x\,d'y - (\Sigma fd'x)(\Sigma fd'y)}{\sqrt{N\Sigma fd'x^2 - (\Sigma fd'x)^2}\,\sqrt{N\Sigma fd'y^2 - (\Sigma fd'y)^2}}$$

$$= \frac{25(22)-(-18)(-18)}{\sqrt{25(34)-(-18)^2}\sqrt{25(38)-(-18)^2}} = \frac{226}{\sqrt{526}\sqrt{626}} = \frac{226}{(22.93)\ (25.02)} = 0.39$$

Example 3. Compute the coefficient of correlation between dividends and price of security as given below :

Security Price (in Rs.) (Y)	Annual dividends (in Rs.) (X)					
	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	–	–	–	–

Sol. First, we prepare the following table :

$d'x = dx/2$	– 3	– 2	– 1	0	1	2
$dx = m - 13$	– 6	– 4	– 2	0	2	4
M.V. (m)	7	9	11	13	15	17

$d'y = dy/10$	$dy = m' - 105$	M.V. (m')	Y \ X →	6–8	8–10	10–12	12–14	14–16	16–18	Total	$fd'y$	$fd'y^2$	$fd'x\,d'y$
3	+ 30	135	130–140	0 –	0 –	– 3 1	0 3	12 4	12 2	10	30	90	21
2	+ 20	125	120–130	0 –	– 4 1	– 6 3	0 3	6 3	4 1	11	22	44	0
1	+ 10	115	110–120	0 –	– 2 1	– 2 2	0 3	2 2	0 –	8	8	8	– 2
0	0	105	100–110	0 –	0 2	0 3	0 2	0 –	0 –	7	0	0	0
– 1	– 10	95	90–100	6 2	4 2	1 1	0 1	0 –	0 –	6	– 6	6	11
– 2	– 20	85	80–90	18 3	4 1	2 1	0 –	0 –	0 –	5	– 10	20	24
– 3	– 30	75	70–80	18 2	6 1	0 –	0 –	0 –	0 –	3	– 9	27	24
			Total	7	8	11	12	9	3	N = 50	$\Sigma fd'y = 35$	$\Sigma fd'y^2 = 195$	$\Sigma fd'x\,d'y = 78$
			$fd'x$	– 21	– 16	– 11	0	9	6	$\Sigma fd'x = -33$			
			$fd'x^2$	63	32	11	0	9	12	$\Sigma fd'x^2 = 127$			
			$fd'x\,d'y$	42	8	– 8	0	20	16	$\Sigma fd'x\,d'y = 78$			

$$r = \frac{N\Sigma fd'xd'y - (\Sigma fd'x)(\Sigma fd'y)}{\sqrt{N\Sigma fd'x^2 - (\Sigma fd'x)^2}\sqrt{N\Sigma fd'y^2 - (\Sigma fd'y)^2}}$$

$$= \frac{50(78) - (-33)\,35}{\sqrt{50(127) - (-33)^2}\sqrt{50\,(195) - (35)^2}} = \frac{5055}{\sqrt{5261}\sqrt{8525}} = \frac{5055}{(72.53)(92.33)} = 0.75$$

EXERCISE 9.5

1. Find coefficient of correlation for the following data :

Income (in Rs).	Saving (in Rs).			
	50	100	150	200
400	10	4	–	–
600	8	12	24	6
800	–	9	7	2
1000	–	–	10	5
1200	–	–	9	4

2. Determine Karl Pearson's coefficient of correlation in the following distribution :

Age of Husbands (in years) (X)	Age of Wives (in years) (Y)					
	16 – 23	23 – 30	30 – 37	37 – 44	44 – 51	Total
18 – 25	9	3	–	–	–	12
25 – 32	–	20	10	4	–	34
32 – 39	–	–	12	5	3	20
39 – 46	–	–	8	7	5	20
46 – 53	–	–	–	10	4	14
Total	9	23	30	26	12	100

ANSWERS

1. 0.55 2. 0.78

9.9.4 PROBABLE ERROR

Probable error denoted by P.E. (r) is used to measure the significance or reliability and dependability of the value of r, the Karl Pearson's coefficient of correlation. According to **Wheldon, "*Probable error defines the limits above and below the size of coefficient determined within which there is an equal chance that the coefficient of correlation similarly calculated from other sample will fall*"**. Thus, it can be summed up that if we find out two limits, one upper and other lower, by adding and subtracting respectively the value of probable error from the value of 'r', the values of coefficients of correlation calculated for the other samples from the same universe (or population) will too fall within the same limits.

The probable error is calculated by the following formula :

$$\text{P.E.}(r) = 0.6745 \frac{1-r^2}{\sqrt{n}} \quad \ldots(9.13)$$

where 'r' is Karl Pearson's correlation coefficient and 'n' is the number of pair of observations of 'X' and 'Y'.

Uses of Probable Error

The probable error has two main useful functions :

1. Determination of Limits : The concept of probable error is used to determine the upper and lower limits with in which the correlation coefficient of a randomly related sample from the same population will lie. The limits are calculated as :

r + P.E. (r) → Upper limit

r – P.E.(r) → Lower limit

The limits defined by ($r \pm$ P.E.(r)) confirm that there is 50% chance that the coefficient of correlation calculated for the other randomly selected samples, taken from the same universe will fall in the same limits.

2. Testing the Significance of Correlation Coefficient : The concept of probable error is regarded as a measure of testing the significance of Karl Pearson's coefficient of correlation. The following rules are followed for the interpretation of the significance of r based on P.E. (r).

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

(*ii*) If r is more than six times the probable error *i.e.* $r > 6$ P.E.(r), the value of r is certainly significant.

(*iii*) If P.E. (r) $< r <$ 6 P.E. (r), nothing can be said about the significance of correlation.

Conditions for the Use of Probable Error

(*i*) The value of n must be sufficiently large because P.E. gives unreliable results for smaller values of n.

(*ii*) The universe from which the sample for the calculation of the value of 'r' is taken must have normal distribution.

(*iii*) The sample should be selected by the random sampling method.

9.9.5 STANDARD ERROR

Now we shall defined another quantity namely *standard error* from the coefficient of correlation which is determined by the formula

$$\text{S.E.}(r) = \frac{1-r^2}{\sqrt{n}} \quad \ldots(9.14)$$

Since $\text{P.E.}(r) = 0.6745 \frac{1-r^2}{\sqrt{n}}$ we can conclude that

$$\textbf{P.E. } (r) = 0.6745 \text{ S.E.} \qquad \text{...(9.15)}$$

9.9.6 COEFFICIENT OF DETERMINATION

Let X and Y be two variables out of which Y is dependent variable and X is independent variable. Now every change in X will produce a corresponding change in Y but every change in Y may or may not be due to change in X. Thus, the changes in the variable Y can be divided into two categories.

(*i*) *Change in Y variable which occur corresponding to change in X series.* Such changes in Y series are known as *Explained variance or Accountable changes.*

(*ii*) *Changes in Y variables which are not due to the changes in X series but due to some other factors.* Such changes in Y series are known as *Unexplained variance or Unaccountable Changes.*

Total Variance will be the sum of explained variance and unexplained variance. It is necessary to determine the proportion of change in Y that has occurred due to change in X. This is determined with the help of *Coefficient of determination.* It is defined as

$$\textbf{Coefficient of Determination } (r^2) = \frac{\textbf{Explained Variance}}{\textbf{Total Variance}} \qquad \text{...(9.16)}$$

If r is 0.6 then $r^2 = 0.36$ which implies that 36% of the variations in the dependent variable *i.e.*, in Y are explained *i.e.*, due to variations in the independent variable X.

Further percentage of variance in the dependent variable *i.e.* in Y which is not due to X variable is determined by '*Coefficient of non-determination*' and is represented by

$$K^2 = 1 - r^2. \qquad \text{...(9.17)}$$

The square root of the coefficient of non-determination *i.e.* K is called '*Coefficient of alienation*'.

Thus, if $r = 0.6$ then we have the following results depending upon the value of r.

(*i*) Coefficient of Determination $= r^2 = (0.6)^2 = 0.36 = 36\%$

(*ii*) Coefficient of Non-determination $= K^2 = 1 - r^2 = 1 - 0.36 = 0.64 = 64\%$

(*iii*) Coefficient of alienation $= K = \sqrt{K^2} = \sqrt{0.64} = 0.8.$

CHECKPOINTS

1. Write a short note on Probable error. (G.N.D.U. B.C.A. April 2003)
2. What are the uses of Probable error ?
3. What do you understand by Standard error ?
4. Write a short note on coefficient of determination.

ILLUSTRATIVE EXAMPLES

Example 1. To study the correlation between the ages of husbands and wives, a sample of 100 is taken from the universe. The sample study gives the coefficient of correlation between two variables as 0.9. Within what limits does it hold good for the universe ?

Sol. Here $r = 0.9$, $n = 100$

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

$$= (0.6745)\frac{1-(0.9)^2}{\sqrt{100}} = (0.6745)\frac{1-0.81}{10}$$

$$= (0.6745)(0.019) = 0.0128$$

$\therefore$ The lower limit $= r - \text{P.E.} = 0.9 - 0.0128 = 0.8872$

and the upper limit $= r + \text{P.E.} = 0.9 + 0.0128 = 0.9128$

Example 2. A student calculates the value of r as 0.7 when the value of n is 5 and concludes that r is highly significant. Is he correct ?

Sol. Here $\text{P.E.} = \dfrac{0.6745(1-r^2)}{\sqrt{n}} = \dfrac{0.6745\{1-(0.7)\}^2}{\sqrt{5}} = \dfrac{0.6745(0.51)}{2.24} = \dfrac{0.34}{2.24} = 0.15$

Further, $\dfrac{r}{\text{P.E.}} = \dfrac{0.7}{0.15} = 4.7$

Since r is less than six times the probable error, it is not significant.

Example 3. If $r = 0.8$ and Probable Error is 0.06, find n.

Sol. We know

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

$$\Rightarrow \quad \frac{6}{100} = \frac{0.6745}{1} \times \frac{1-0.64}{\sqrt{n}} = \frac{(0.6745)(0.36)}{\sqrt{n}}$$

$$\Rightarrow \quad \sqrt{n} = 4.047$$

$$\Rightarrow \quad n = (4.047)^2 = 16.3782 \quad \Rightarrow \quad n = 16 \text{ as n cannot be in fraction.}$$

Example 4. What will be the coefficient of correlation if X and Y are related by $Y = X^2$, for $X = 1, 2, 3, 4, 5$. What can you say about the variation of Y due to changes of X ?

Sol.

X	$Y = X^2$	$\bar{X} = 3$ $x = X - \bar{X}$	$\bar{Y} = 11$ $y = Y - \bar{Y}$	x^2	y^2	xy
1	1	– 2	– 10	4	100	20
2	4	– 1	– 7	1	49	7
3	9	0	– 2	0	4	0
4	16	1	5	1	25	5
5	25	2	14	4	196	28
$\Sigma X = 15$	$\Sigma Y = 55$	0	0	$\Sigma x^2 = 10$	$\Sigma y^2 = 374$	$\Sigma xy = 60$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{15}{5} = 3, \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{55}{5} = 11$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = \frac{60}{\sqrt{10 \times 374}} = = \frac{60}{61.16} = 0.98$$

Now coefficient of determination $= r^2 = (0.98)^2 = 0.96$

This means that 96 % variation in Y series are due to X, *i.e.*, explained variance = 96%.

We can further say that only 4 % variations are unexplained.

Example 5. Is it true that a correlation coefficient of $r = 0.8$ indicates a relationship twice as close as $r = 0.4$?

Sol. Coefficient of determination *i.e.* r^2 in the first case $= (0.8)^2 = 0.64$

Coefficient of determination *i.e.* r^2 in the second case $= (0.4)^2 = 0.16$

Thus in the first case 64% of the variation is explained and in the second case 16% of the variation is explained. Hence it is not true that a correlation coefficient of $r = 0.8$ indicates a relationship twice as close as $r = 0.4$.

Example 6. "A correlation coefficient of – 0.5 means that 50% of the data are explained". Comment.

Sol. If r is the correlation coefficient then r^2 indicates the percentage variation of the data which is explained. Hence if $r = -0.5$, r^2 will be 0.25 and hence only 25% of the data will be explained. Consequently the given statement is wrong.

EXERCISE 9.6

1. A student calculates the value of 'r' as 0.7 when the number of items in the sample is 25. Find the limits within which r lies for another sample from the same universe.
2. In a correlation analysis, the values of Karl Pearson's coefficient of correlation and its Probable Error were found to be 0.90 and 0.04 respectively. Find the value of n.

3. For what value of n, the coefficient of correlation equal to 0.6 will be significant ?
4. Show by calculation which 'r' is more significant
   (*i*) $r = 0.8$ and P.E. = 0.09 (*ii*) $r = 0.7$ and P.E. = 0.05
5. What will be coefficient of correlation of X and Y which are related by $Y = X^2$ for $X = -3, -2, -1, 0, 1, 2, 3$, what can you say about the variation of Y due to change in X.
6. Is it true that a correlation coefficient 0.6 indicates relationship twice as close as 0.3 ?
7. If 50% of the data is explained does it mean that $r = 0.5$?
8. If the standard error of r for 25 pairs of observations is 0.05, find the probable error.

ANSWERS

1. 0.631 to 0.769	2. 10	3. 19	4. Case (*ii*)
5. 0	6. No	7. No	8. 0.033725

9.10 SPEARMAN'S COEFFICIENT OF CORRELATION (RANK CORRELATION)

This method was developed by a British Psychologist **Prof. Charles Edward Spearman** in 1904. Rank correlation coefficient is used for measuring the relationship between two qualitative variables such as honesty, beauty, tastes etc., which can not be measured quantitatively. This method is known as '*Spearman's method of coefficient of correlation*' or popularly known as '*Rank Correlation method*'. Unlike Karl Pearson's method of coefficient of correlation, Rank correlation method is not based on the assumption of normality of data. The data are quite irregular in such cases. As such variables are assigned grades i.e. ranks depending upon the size of the items in the ascending or descending order of magnitude.

Rank correlation coefficient is denoted by r_s or ρ (rho). Mathematically,

$$r_s \text{ or } \rho = 1 - \frac{6\Sigma D^2}{n(n^2-1)} \qquad \ldots(9.18)$$

or

$$r_s \text{ or } \rho = 1 - \frac{6\Sigma D^2}{n^3 - n} \qquad \ldots(9.19)$$

where D is the difference in ranks of the variables and n is the number of pair of observations.

Characteristics of Spearman's Rank Correlation Coefficient

The following are the main features of rank correlation coefficient :

1. Like Karl Pearson's coefficient of correlation, Spearman's rank correlation coefficient also lies between -1 to $+1$ *i.e.*

$$-1 \leq \rho \leq +1$$

2. Sum of difference of ranks between two variables is zero *i.e.* $\Sigma D = 0$.
3. It is not based on the assumption of normal distribution of population.
4. When $\Sigma D^2 = 0$, $\rho = 1$. It means each $D = 0$ and both series have similar ranks.

Merits and Demerits of Rank Correlation Method

Merits : The following are the merits of rank correlation method :

(*i*) It is easy to understand and simple to calculate.

(*ii*) When measurements are given in the quantitative form, even then this method can be used by assigning ranks to different values.

(*iii*) In the study of relationship of series having qualitative characteristics such as beauty, honesty, tastes, promotions etc., rank correlation method is the only substitute.

(*iv*) Spearman's method is not affected by the presence of extreme values because it is not based on actual values of the data.

Demerits : The following are the demerits of rank correlation method :

(*i*) Spearman's coefficient of correlation cannot be applied in the case of grouped frequency distribution.

(*ii*) It is an approximate measure of correlation because it is not based on the actual values in the data.

(*iii*) If the sample is large the method cannot be used conveniently. The procedure of ranking the items become tedious in such cases.

(*iv*) It has no further application in any statistical operation.

(*v*) We cannot determine combined coefficient of correlation based upon the values of a few samples.

9.10.1 COMPARISON OF SPEARMAN'S AND KARL PEARSON'S COEFFICIENTS OF CORRELATION

1. In the case of Rank correlation method coefficient of correlation is perfectly positive *i.e.* 1 if the two series have equal corresponding ranks so that each $D = 0$ and as such $\Sigma D^2 = 0$ where as in Karl Pearson's method coefficient of correlation is perfectly positive *i.e.* 1 if X and Y series change uniformly i.e. X and Y are linearly related.

2. Coefficients of correlation can be interpreted similarly in both cases. In Spearman's method $-1 \leq \rho \leq 1$ and in Karl Pearson's method $-1 \leq r \leq 1$

3. Karl Pearson assumes that the universe from which the sample is drawn is normal. Any diversion from this assumption will give abnormal results. In case of Spearman's method no such assumption is needed.

4. Spearman's formula is easy to understand as compared to Karl Pearson's formula. The values of ρ and r will generally differ. The difference arises because of the fact that Karl Pearson's formula is based upon actual values where as Spearman's formula is based upon ranks of the values which can remain ineffective even if the actual values in the series are changed.

5. When the data is given in the qualitative form Spearman's formula is the only suitable method. Karl Person's method cannot be applied in such cases.

6. Spearman's coefficient of correlation cannot the applied in the case of a bivariate frequency distribution. Karl Pearson's formula is the only effective formula which can measure correlation in grouped series.

7. Combined coefficient of correlation can be determined by Karl Pearson's method of coefficient of correlation for various sub-groups if their coefficients of correlation along with their number of items are given. This is not possible in case of Spearman's method.

9.10.2 CALCULATION OF RANK CORRELATION COEFFICIENT

The following situations may arise for the calculation of correlation coefficient :

1. When ranks are given.
2. When ranks are not given.
3. When items or ranks repeat in the data.

1.When Ranks are Given : In this situation Spearman's correlation coefficient is calculated by following formula (9.17) or (9.18) :

$$r_s \text{ or } \rho = 1 - \frac{6\Sigma D^2}{n^3 - n}$$

The following steps are involved in this method :

Step 1. Compute difference of ranks of two variables and denote it by D.

Step 2. Calculate ΣD^2.

Step 3. Put the values in above formula to obtain rank correlation coefficient.

2. When Ranks are Not Given : If the data are not ranked, we have to rank the values of both variables X and Y according to their magnitude. We may rank the values either by assigning 1 to the smallest item then 2, 3, 4, ... in ascending order or by assigning rank 1 to the highest item then 2, 3, 4, ... in descending order. When ranks are assigned, we apply the formula (9.17) (or (9.18)) to compute rank correlation coefficient.

3. When Items or Ranks Repeat : When two or more items in a series have equal value they are assigned equal ranks or average ranks. The next item is assigned a rank which ought to have been assigned in the absence of such tie. In such situation, an adjustment is made in the formula and we add a correction factor $\frac{1}{12}(m^3 - m)$ to ΣD^2, where m refers to the number of times an item is repeated. The correction factor is to be added for each and every repeated item. In this case the formula becomes :

$$r_s \text{ or } \rho = 1 - \frac{6\left[\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + ...\right]}{n^3 - n} \quad ...(9.20)$$

CHECKPOINTS

1. What do you understand by Rank Correlation method ?
2. Discuss the characteristics of rank correlation coefficient.
3. What are the merits and demerits of rank correlation method ?
4. Compare Spearman's and Karl Pearson's coefficient of correlation.

ILLUSTRATIVE EXAMPLES

Example 1. In a bivariate sample, the sum of the squares of the differences between ranks of observed values of two variables is 231 and the correlation coefficient between them is – 0.4. Find the number of pairs.

Sol. Given $\Sigma D^2 = 231, \quad \rho = -0.4$

Since $\rho = 1 - \dfrac{6\Sigma D^2}{n^3 - n}$

$\therefore \quad -\dfrac{4}{10} = 1 - \dfrac{6 \times 231}{(n-1)\, n\, (n+1)}$

$\Rightarrow \quad \dfrac{1386}{(n-1)\, n\, (n+1)} = \dfrac{14}{10} \qquad \Rightarrow \quad \dfrac{99}{(n-1)\, n\, (n+1)} = \dfrac{1}{10}$

$\Rightarrow \quad (n-1)\, n\, (n+1) = 9 \times 10 \times 11$

Comparing both sides, we get, $n = 10$

Example 2. The coefficient of rank correlation between X and Y was found to be 0.5 for 10 observations. It was later discovered that one wrong pair of observations whose difference of rank was 6, was taken by mistake. It was later weeded out. Find the correct rank correlation of 9 observations.

Sol. Given Incorrect $\rho = 0.5, \quad n = 10$

Now $\rho = 1 - \dfrac{6\Sigma D^2}{n^3 - n}$

$\therefore \quad 0.5 = 1 - \dfrac{6\Sigma D^2}{990}$

$\therefore \quad 6\,\Sigma D^2 = 990\,(0.5)$

$\therefore$ Incorrect $\Sigma D^2 = 82.5$

Correct $\Sigma D^2 = 82.5 - 36 = 46.5$

Now taking $n = 9$ because the incorrect pair has been left out

$\therefore$ Correct $\rho = 1 - \dfrac{6(46.5)}{9(80)} = 1 - 0.3875 = 0.6125$

Example 3. Twelve entries in painting competition were ranked by two judges as shown below :

Entry	A	B	C	D	E	F	G	H	I	J	K	L
Judge I	5	2	3	4	1	6	8	7	10	9	12	11
Judge II	4	5	2	1	6	7	10	9	11	12	3	8

Find the coefficient of rank correlation.

(G.N.D.U. B.C.A. April 2002)

Sol. First, we construct the following table :

R_I	R_{II}	$D = \lvert R_I - R_{II} \rvert$	D^2
5	4	1	1
2	5	3	9
3	2	1	1
4	1	3	9
1	6	5	25
6	7	1	1
8	10	2	4
7	9	2	4
10	11	1	1
9	12	3	9
12	3	9	81
11	8	3	9
			$\Sigma D^2 = 154$

Now, coefficient of rank correlation $\rho = 1 - \dfrac{6\Sigma D^2}{n^3 - n} = 1 - \dfrac{6(154)}{(12)^3 - 12} = 0.46.$

Example 4. Ten competitors in a beauty contest are ranked by three judges in the following order :

First Judge	:	1	6	5	10	3	2	4	9	7	8
Second Judge	:	3	5	8	4	7	10	2	1	6	9
Third Judge	:	6	4	9	8	1	2	3	10	5	7

Use the method of rank correlation to determine which pair of judges have the nearest approach to common likings in beauty.

(P.U. B.C.A. April 2004, G.N.D.U. B.C.A. Sept. 2006)

Sol.

R_1	R_2	R_3	$D_1 = \lvert R_2 - R_3 \rvert$	$D_2 = \lvert R_1 - R_3 \rvert$	$D_3 = \lvert R_1 - R_2 \rvert$	D_1^2	D_2^2	D_3^2
1	3	6	3	5	2	9	25	4
6	5	4	1	2	1	1	4	1
5	8	9	1	4	3	1	16	9
10	4	8	4	2	6	16	4	36
3	7	1	6	2	4	36	4	16
2	10	2	8	0	8	64	0	64
4	2	3	1	1	2	1	1	4
9	1	10	9	1	8	81	1	64
7	6	5	1	2	1	1	4	1
8	9	7	2	1	1	4	1	1
						$\Sigma D_1^2 = 214$	$\Sigma D_2^2 = 60$	$\Sigma D_3^2 = 200$

Now, Rank correlation between 2nd and 3rd judge

$$= 1 - \frac{6\Sigma D_1^2}{n^3 - n} = 1 - \frac{6 \times 214}{990} = 1 - \frac{214}{165} = \frac{-49}{165} = -0.30$$

Rank correlation between 1st and 3rd judge $= 1 - \frac{6\Sigma D_2^2}{n^3 - n} = 1 - \frac{6 \times 60}{990} = 1 - \frac{12}{33} = \frac{21}{33} = 0.64$

Rank correlation between 1st and 2nd judge $= 1 - \frac{6\Sigma D_3^2}{n^3 - n} = 1 - \frac{6 \times 200}{990} = -0.21$

Since coefficient of correlation is highest between 1st and 3rd judge, therefore 1st and 3rd judges have the nearest taste for beauty.

EXERCISE 9.7

1. If $n = 10$ and $\Sigma D^2 = 280$, what is the Co-efficient of Rank Correlation ?
2. The rank correlation co-efficient between the marks obtained by 10 students in Mathematics and Economics are found to be 0.5. Find the sum of squares of difference of ranks.
3. The sum of the squares of the differences in the ranks of n pairs of observations is known to be 126 and the coefficient of rank correlation is – 0.5. Find n.
4. The co-efficient of rank correlation between the debenture prices and shares prices are found to be 0.143. If the sum of squares of the difference in ranks is given to be 48 find the value of n.

5. The coefficient of rank correlation of a beauty contest involving 12 participants was calculated as 0.6. However, it was later on discovered that the difference in ranks of a participant is read as 8 instead of 3. Calculate correct correlation coefficient.

6. The rank coefficient of correlation between two variables X and Y is 0.4 for 7 pairs of observations. It was later discovered that the differences in ranks between two variables for one particular observation was wrongly taken as 4 instead of 3. Find the corrected rank correlation.

7. Two Judges in the beauty competition rank 12 entries as follows :

X :	1	2	3	4	5	6	7	8	9	10	11	12
Y :	12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between the Judges ?

8. The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics :

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

Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Physics. (P.U. B.C.A. Sept. 2004)

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

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

Use Rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

10. Ten entries are submitted for a competition. Three judges study each entry and then list them in the rank order as under :

Entry No. :	1	2	3	4	5	6	7	8	9	10
Judge A :	9	3	7	5	1	6	2	4	10	8
Judge B :	9	1	10	4	3	8	5	2	7	6
Judge C :	6	3	8	7	2	4	1	5	9	10

Calculate the appropriate rank correlation to help you answer the followings :

(*i*) Which pair of Judges agree the most ?

(*ii*) Which pair of Judges disagree the most ?

11. Calculate Spearman's coefficient of correlation between marks assigned to ten students by judges X and Y in a certain competitive test as shown below :

Sr. No. :	1	2	3	4	5	6	7	8	9	10
Marks by judge X :	52	53	42	60	45	41	37	38	25	27
Marks by judge Y :	65	68	43	38	77	48	35	30	25	50

12. Calculate Rank correlation from the following marks given out of 200 by two judges X and Y in a music competition to 8 participants :

Sr. No.	:	1	2	3	4	5	6	7	8
Marks awarded by X	:	74	98	110	70	65	85	88	59
Marks awarded by Y	:	121	133	170	102	90	152	160	85

Also give the comments on the computed coefficient.

13. Calculate rank correlation from the following scores :

Physics :	50	40	30	45	60
Chemistry :	60	50	20	30	50

(G.N.D.U. B.Sc. C.Sc. April 2004)

14. The following table shows the marks of 12 students in Mathematics and Geography. Find rank correlation coefficient.

Students	:	A	B	C	D	E	F	G	H	I	J	K	L
Marks in Mathematics :		65	40	35	75	65	80	35	20	85	65	55	33
Marks in Geography :		30	55	68	28	76	25	80	85	20	35	45	65

15. Find Coefficient of Rank Correlation :

X :	20	25	30	15	35	55	25	65
Y :	35	30	45	30	20	10	30	50

16. In an experiment the two variables X and Y were found as follows :

Step No.	1	2	3	4	5	6	7	8	9	10
Value X :	23	27	28	28	29	30	31	33	35	36
Value Y :	18	20	22	27	21	29	27	29	28	29

Find coefficient of correlation by Spearman's method.

17. Find out Spearman's rank coefficient of correlation from the following data :

X :	8	− 10	− 4	0	− 6	10	8	9	− 6	− 1
Y :	3	5	0	1	1	− 4	− 5	− 8	5	1

18. Calculate rank correlation coefficient for the following data :

X :	12	15	18	20	16	15	18	22	15	21	18	15
Y :	10	18	19	12	15	19	17	19	16	14	13	17

ANSWERS

1. – 0.70
2. 82.5
3. 8
4. 7
5. 0.792
6. 0.525
7. – 0.45
8. 0.8
9. 2nd and 3rd
10. (*i*) C and A
    (*ii*) B and C
11. 0.54
12. 0.93, high degree of correlation
13. 0.65
14. – 0.85
15. 0
16. 0.83
17. – 0.71
18. 0.065

MISCELLANEOUS EXERCISE

1. You are given the following information relating to a frequency distribution comprising of 10 observations

$\overline{X} = 5.5,\ \overline{Y} = 4.0,\ \Sigma X^2 = 385,\ \Sigma Y^2 = 192,\ \Sigma (X + Y)^2 = 947$

Calculate the value of 'r'.

2. From the data given below calculate coefficient of correlation and interpret it.

	X-Series	Y-Series
Number of items	8	8
Mean	68	69
Sum of Squares of deviations from Mean	36	44
Sum of the product of deviations x and y from Means	24	24

3. From the following data, calculate the coefficient of correlation between X-series and Y-series :

		X – Series	Y – Series
Mean	:	80	120
Assumed Mean	:	65	110
Standard Deviation	:	12.5	14.2

Sum of product of corresponding deviations of X and Y series from their assumed means $(\Sigma\, dx\, dy) = 2080$ and Number of pairs = 10.

4. Calculate the correlation coefficient from the following results :

$n = 10$ $\quad \Sigma X = 350$ $\quad \Sigma Y = 310$

$\Sigma (X - 35)^2 = 162$ $\quad \Sigma (Y - 31)^2 = 222$ $\quad \Sigma (X - 35)(Y - 31) = 92$

5. Calculate Karl Pearson's coefficient of correlation between X and Y for the following information $n = 12, \Sigma X = 120, \Sigma Y = 130, \Sigma (X - 8)^2 = 150, \Sigma(Y - 10)^2 = 200$ and $\Sigma(X - 8)(Y - 10) = 50$

6. From the following table calculate the co-efficient of correlation by Karl Pearson's method

X :	6	2	10	4	8
Y :	9	11	?	8	7

Arithmetic means of X and Y series being 6 and 8 respectively.

7. The following table gives X (no. of fishing boats) and Y (no. of accidents in that area) for comparative years :

X :	705	755	650	505	455	700	400	445	600	655
Y :	14	19	16	12	10	18	9	8	9	8

Calculate coefficient of correlation and make your comments on findings.

(G.N.D.U. B.Sc. C.Sc. April 2006)

8. Compute Karl Pearson's coefficient of correlation for the following data and calculate its probable error :

Marks in English :	77	54	27	52	14	35	90	25	56
Marks in Maths :	35	58	60	40	50	40	35	56	34

(G.N.D.U. B.C.A. April 2007)

9. From the following information relating to age of candidates and their examination results, calculate '*r*' and probable error.

Age of Candidate :	13 – 14	14 – 15	15 – 16	16 – 17	17 – 18	18 – 19	19 – 20	20 – 21	21 – 22
% of failure:	32.9	40.6	43.4	34.2	36.6	39.2	48.9	47.1	54.5

10. With the following data in six cities calculate the coefficient of correlation by Pearson's method between the density of population and death rate.

Cities	Area in Sq. miles	Population (000)	No. of Deaths
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

11. Calculate Karl Pearson's Coefficient of correlation for the following series

Price (Rs.)	110 – 111	111 – 112	112 – 113	113 – 114	114 – 115
Demand (kg.)	600	640	640	680	700
Price (Rs.)	115 – 116	116 – 117	117 – 118	118 – 119	
Demand (kg.)	780	830	900	1000	

Also calculate the probable error of correlation coefficient. From your result can you assert that the demand is correlated with price ?

12. Calculate Karl Pearson's Coefficient of correlation from the following data and find out its standard error :

X	105	104	102	101	100	99	98	96	93	92
Y	101	103	100	98	95	96	104	92	97	94

ANSWERS

1. – 0.68
2. 0.603, Not significant
3. 0.33
4. 0.48
5. 0.214
6. – 0.92
7. 0.72, moderate degree of positive correlation
8. – 0.66, 0.127
9. 0.77, 0.915
10. 0.98
11. 0.965, 0.0155, Yes
12. 0.596, 0.204