MATH 4648

Chapter 11
SIMPLE CORRELATION

11.1 Introduction

So far, we have discussed different characteristics of a single variable. For example, demand of a commodity over a period of time, income of a number of families, volume of sales by a number of salesmen, etc. The analysis with a single variable is termed as univariate analysis. In the real field two or more variables may be interrelated. There are many situations in business where we are interested to measure the relationship between two variables such as the income and expenditure of a certain class of people, price of a commodity and amount demanded, volume of sales and the experience of the salesman of a departmental store, deposit in a bank and number of clients, family income and expenditure on luxury items, the fertilizer used and production of certain crop, etc. Pairs of observations of two such variables produce a bivariate distribution. It is often required to know how stronger the relationship between two such variables is or it might be required to know the impact of change in one variable on another variable, or it may be required to forecast the value of one variable for a particular given value of another. The study of these types of relationship can be performed through two types of statistical tools, viz the correlation analysis and regression analysis. In this chapter, we will discuss different aspects of correlation analysis for two variables.

11.2 Correlation Analysis

In this section we shall consider the problem of measuring the relationship between two quantitative variables. In business we come across a large number of problems involving the use of two or more related variables. For example, there exits some relationship between family income and expenditure on luxury items, price of a commodity and amount demanded, price of a commodity and the amount supplied, advertisement expenditure of a commodity and amount sold etc. The statistical tool with the help of which these relationships between two or more than two variables can be studied is called correlation.

A. M. Tuttle gives a very simple definition of correlation. According to "A.M. Tuttle" an analysis of covariation of two more variables is usually called correlation.

Covariation or relationship between two variables is measured by covariance. We shall define covariance both for population and sample.

Population covariance. Suppose $(X_1, Y_1), (X_2, Y_2), \ldots\ldots\ldots, (X_N, Y_N)$, are N pairs of values of a a bivariate population of two variables X and Y with respective means μ_x and μ_y, then population covariance between X and Y, denoted by μ_{11} is defined as

$$\text{Cov}(X, Y) = \mu_{11} = \frac{\Sigma(X-\mu_x)(Y-\mu_y)}{N}.$$

Population covariance can also be defined for two random variables.

- **Population covariance.** Suppose X and Y are two random variables with means E[X] and E[Y] respectively, then population covariance is defined by

$$\mu_{11} = E[X - E(X)][Y - E(Y)]$$

- **Sample covariance.** Suppose (x_1, y_1), (x_2, y_2),..., (x_n, y_n) are n pairs of values of a bivariate sample of two variables x and y with respective means $\bar{x}$ and $\bar{y}$, then sample covariance between x and y is defined by

$$\text{Cov}(x, y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}.$$

Working formula for finding sample covariance is

$$\text{Cov}(x,y) = \frac{1}{n}\left\{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}\right\}.$$

Some important properties of covariance.

i) The range of covariance is $-\infty$ to ∞.
ii) It depends on the units of measurements on which the variables are measured.
iii) It gives the magnitude and direction of the statistical relationship between two variables.
iv) The value of covariance will be positive if the increase or decrease of one variable is associated with the increase or decrease of the other variable.
v) The value of covariance will be negative if the increase or decrease of one variable is associated with the decrease or increase of the other variable.
vi) The value of covariance is zero if the two variables are linearly independent.

Generally, there are three types of correlation. They are (i) Simple correlation, (ii) Partial correlation, and (iii) Multiple correlations.

In this chapter, we shall discuss only simple correlation.

11.3 Simple Correlation

If only two variables are chosen to study the correlation between them, then such correlation is referred to as simple correlation. The most widely used measure of simple relationship between two variables is called Karl Pearson product-moment correlation coefficient or simply the correlation coefficient, which is a better than covariance as a measure of relationship between two variables.

- **Simple Correlation coefficient.** Simple correlation coefficient is a quantitative measure of strength and direction of linear relationship between two numerically measured variables.

We shall define correlation coefficient both for population and sample.

Population simple correlation coefficient. Population simple correlation coefficient measures the strength of linear relationship between two variables of a bivariate population. Suppose $(X_1, Y_1), (X_2, Y_2), \ldots\ldots, (X_N, Y_N)$ are N pairs of values of two variables X and Y of a bi-variate population with means μ_x and μ_y, then the population correlation coefficient denoted by ρ is defined by Karl Pearson as

$$\rho = \frac{\Sigma(X-\mu_x)(Y-\mu_y)}{\sqrt{\Sigma(X-\mu_x)^2 \Sigma(Y-\mu_y)^2}}.$$

Simple population correlation coefficient can also be defined for two random variables of a bi-variate population.

Population correlation coefficient. Simple population correlation coefficient also measures the strength of linear relationship between two random variables. Suppose X and Y are two random variables of a bi-variate population, and then population correlation coefficient denoted by ρ is defined as

$$\rho = \frac{E[X-E(X)][Y-E(Y)]}{\sqrt{E[X-E(X)]^2 E[Y-E(Y)]^2}} = \frac{\mu_{11}}{\sigma_x \sigma_y}$$

where $\sigma_x^2 = E[X-E(X)]^2$ and $\sigma_y^2 = E[Y-E(Y)]^2$.

Actually, E(X) and E(Y) are the population means of X and Y, and σ_x^2 and σ_y^2 are population variances of the random variables X and Y which were defined in Chapter 9.

Simple sample correlation coefficient. Simple sample correlation coefficient measures the strength of linear relationship between two variables when a bivariate sample is taken from a bivariate population. Suppose $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are n pairs of sample values of two variables x and y from a bivariate population. Let $\bar{x}$ and $\bar{y}$ be the sample means of x and y. Then the sample Karl Pearson's correlation coefficient between x and y denoted by r is defined as

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}}.$$

Actually, it is defined with the help of the sample covariance and variances of x and y as

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})/n}{\sqrt{\{\Sigma(x-\bar{x})^2/n\}\{\Sigma(y-\bar{y})^2/n\}}} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\,\text{var}(y)}}.$$

The sample correlation coefficient r is used to estimate the population correlation coefficient ρ. So, sample correlation coefficient is important in correlation analysis. Usually, correlation coefficient is computed with any one of the following formulae

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y - (\Sigma y)^2\}}}.$$

These two formulae are used to compute the value of r. They are also known as the working formula for finding the value of r. The value of correlation coefficient r or ρ lies between -1 to $+1$.

We shall use sample notation r for the further correlation analysis. The value of correlation coefficient may be positive, negative or zero.

11.3.1 Examples of positive correlation coefficient. The value of correlation coefficient between two variables will be positive if the increase or decrease of one variable is associated with the increase or decrease of the other variable. Some examples of positive correlation coefficient:

i) The heights and weights of a group of persons;
ii) The income and expenditure of a certain class of people;
iii) The fertilizer used and the production of certain crop;
iv) The amount of sales and the experience of the salesman of a departmental store;
v) The deposit in a bank and the number of clients;
vi) The ages of husbands and the ages of wives etc.

11.3.2 Examples of negative correlation coefficient. The value of correlation coefficient will be negative if the increase or decrease of one variable is associated with the decrease or increase of the other variable.

11.3.3 Some examples of negative correlation coefficients are

i) The price and demand of a commodity, as the price of a product increases the demand for that product decreases. The demand of mobile set increases as the price decreases;
ii) The volume of a perfect gas increases as the pressure decreases;
iii) The price of a commodity decreases as the supply increases.

11.3.4. Some examples of independence of two variables are

i) The rainfall of Bangladesh and the production of rice in Vietnam;
ii) The demand of some commodities does not depend on the increase or decrease of the prices of the commodities. For example, salt, oil, rice etc are such commodities. They are known as perishable goods;
iii) The heights and ages of university students;
iv) The price of gasoline and the rainfall etc.

Remarks. When two variables are linearly independent, then the value of correlation coefficient is zero. But $r = 0$ does not mean that the two variables x and y are not related. For example, the

correlation coefficient between x and y is zero when $y = x^2$. Here x and y are not linearly related. Actually, this is the equation of a parabola.

11.4 Assumption Underlying Karl Pearson's Correlation Coefficient or Simple Correlation Coefficient

The simple correlation coefficient r is based on the following assumptions:

i) The relationship between the variables is linear;

ii) Both the variables are measured on interval or ratio scales;

iii) The two variables must follow bivariate normal distribution;

iv) The sample is of adequate size to assume normality.

11.5 Some Important Properties of Correlation Coefficient

i) The value of r lies between -1 to +1.

ii) It measures the magnitude and direction of statistical relationship between two variables.

iii) It is a pure number. That is, it is independent of the units of measurements of the variables.

iv) It is a symmetrical function of x and y. That is $r_{xy} = r_{yx}$.

v) $r = 0$ Indicates no linear relationship between x and y.

vi) $r = +1$ Indicates a perfect positive relationship between x and y.

vii) $r = -1$ Indicates a perfect negative relationship between x and y.

viii) The geometric mean of two regression coefficients is equal to the correlation coefficient.

ix) The correlation coefficient is independent of the shift of origin and change of scale.

Now, we shall prove some important properties of correlation coefficient.

Theorem 11.5.1 The value of correlation coefficient lies between -1 to +1.

Proof. Suppose $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ are n pairs of values of a bivariate sample. The correlation coefficient between x and y is

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

Here $X = (x-\bar{x})$ and $Y = (y-\bar{y})$, and $\bar{x}$ and $\bar{y}$ are means x and y.

Let us consider the expression which is always positive

$$\Sigma\left(\frac{X}{\sqrt{\Sigma X^2}} + \frac{Y}{\sqrt{\Sigma Y^2}}\right)^2 \geq 0$$

$$\Rightarrow \Sigma\left(\frac{X^2}{\Sigma X^2} + \frac{Y^2}{\Sigma Y^2} + \frac{2XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}\right) \geq 0$$

$$\Rightarrow \frac{\Sigma X^2}{\Sigma X^2}+\frac{\Sigma Y^2}{\Sigma Y^2}+2r\geq 0$$

$$\Rightarrow 1+1+2r \geq \Rightarrow r\geq -1 \quad (11.$$

Again $$\Sigma\left(\frac{X}{\sqrt{\Sigma X^2}}-\frac{Y}{\sqrt{\Sigma Y^2}}\right)^2\geq 0$$

$$\Rightarrow \Sigma\left(\frac{X^2}{\Sigma X^2}+\frac{Y^2}{\Sigma Y^2}-\frac{2XY}{\sqrt{\Sigma X^2\Sigma Y^2}}\right)\geq 0$$

$$\Rightarrow \frac{\Sigma X^2}{\Sigma X^2}+\frac{\Sigma Y^2}{\Sigma Y^2}-2r\geq 0$$

$$\Rightarrow 1+1-2r\geq \Rightarrow r\leq 1 \quad (11.$$

From (11.5.1) and (11.5.2), we have $-1\leq r\leq 1$

This is the proof of the theorem.

Theorem 11.5.2 The value of the correlation coefficient is 1 when $y = a + bx$.

Proof. The correlation coefficient between x and y is

$$r=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}} \quad (11.$$

Here $y = a + bx$ and $\bar{y}=a+b\bar{x}$. Now, put the values of y and $\bar{y}$ in (11.5.3), we have

$$r=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}=\frac{\Sigma(x-\bar{x})(a+bx-a-b\bar{x})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y+bx-a-b\bar{x})^2}}=\frac{b\Sigma(x-\bar{x})^2}{b\Sigma((x-\bar{x}))^2}=1.$$

This proves the theorem.

Theorem 11.5.3 The value of the correlation coefficient is –1 when $y = a - bx$.

Proof. The correlation coefficient between x and y is

$$r=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}} \quad (11.$$

Here, $y = a - bx$ and $\bar{y}=a-b\bar{x}$. Now, put the values of y and $\bar{y}$ in (11.5.4), we have

$$r=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}}=\frac{\Sigma(x-\bar{x})(a-bx-a+b\bar{x})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-bx-a+b\bar{x})^2}}=\frac{-b\Sigma(x-\bar{x})^2}{b\Sigma(x-\bar{x})^2}=-1.$$

This proves the theorem.

Theorem 11.5.4 Correlation coefficient is independent of the shift of origin and change of scale

Proof. Suppose x and y are two variables. Now we shall define two new variables u and v as

$$u=\frac{x-A}{h} \quad \text{and} \quad v=\frac{y-B}{k}.$$

This means we have shifted the origin of x to A and y to B. Also we have change the scale of x by dividing it by h and y by k.

That is $x = A + hu \Rightarrow \bar{x} = A + h\bar{u}$

$$y = B + kv \Rightarrow \bar{y} = A + k\bar{v}$$

The correlation coefficient between x and y is defined by

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

Now by putting the values of x, $\bar{x}$, y, $\bar{y}$ in the above formula, we have

$$r_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\,\Sigma(y-\bar{y})^2}} = \frac{\Sigma(A+hu-A-h\bar{u})(B+kv-B-k\bar{v})}{\sqrt{[\Sigma(A+hu-A-h\bar{u})^2\,\Sigma(B+kv-B-k\bar{v})^2]}}$$

$$= \frac{hk\Sigma(u-\bar{u})(v-\bar{v})}{hk\sqrt{\Sigma(u-\bar{u})^2\,\Sigma(v-\bar{v})^2}} = r_{uv}.$$

It is one of the important properties of correlation coefficient. We shall now give some of its applications.

Remarks. (i) The property says if a constant value A and a constant value B are subtracted from x and y respectively then the correlation coefficient of the new variables u and v is the same as the correlation coefficient between the original variables x and y. This means correlation coefficient is independent of the shift of origin.

(ii) If the values of x and the values of y are very large, we can divide each value of x by a constant value h and each of y by k, then the correlation coefficient between the two new variables is the same as the original variables. This means correlation coefficient is independent of the change of scale.

Very often we shift the origins and change the scales of both the variables to get the maximum computations benefit. Now we shall cite some examples.

The above formula is very useful to find the correlation coefficient from a bivariate frequency distribution which is not given in this book.

11.6 Scatter Diagram

The nature of association between two variables may be observed by plotting pairs of observations of the variables on a graph, such a graph is known as a scatter diagram or scatter plot, since it depicts how two variables are related or the pairs of points are scattered. The scatter plot is an essential and important step in studying the relationship between two variables.

Definition. Scatter Diagram. The simplest device for showing the relationship between t variables on a graph paper in the form of dots is called scatter diagram or scatter plot.

Actually, it gives a rough idea about the relationship between two quantitative variab Suppose (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) are n pairs of sample values of two variables x and y the values of the variables x and y be plotted along the x-axis and y-axis respectively in the plane, the diagram of dots so obtained is known as scatter diagram. Actually, it portrays relationship between these two variables graphically. By looking at the scatter of the vario points on the chart, it is possible to determine the extent of association between these t variables. The wider the scatter on the chart, the less close is the relationship. On the other han the closer the points and the closer they come to falling on a line passing through them, higher the degree of relationship. If all the points fall on a line, the relationship is perfect. If t line goes up from the lower left hand corner to the upper right hand corner, i.e., if the slope of t line is positive, then the correlation between the two variables is considered to be perfect positi and the value of r is +1. Similarly, if the line starts at upper left hand corner and comes down the lower right hand corner of the diagram, i.e., if the slope is negative, and also all the points on the line, then their correlation is said to be perfect negative and the value of r is –1.

Scatter diagrams for different values of r are as follows:

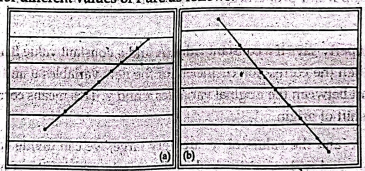


Fig 11.6.1a *Scatter Diagram showing* r = +1

Fig 11.6.1b *Scatter Diagram showing* r = –1

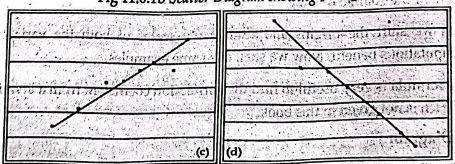


Fig 11.6.1c Scatter Diagram showing $0 < r \leq 1$

Fig 11.6.1d Scatter Diagram showing $-1 \leq r < 0$

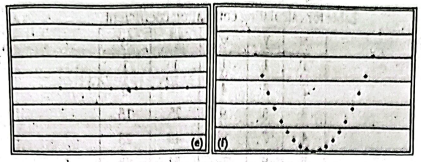


Fig 11.6.1e Scatter Diagram showing $r = 0$

Fig 11.6.1f. Scatter Diagram showing $y = x^2$

Interpreting the values of r

1. $r = +1$ Indicates a perfect positive relationship between x and y. In this case all the values of x and y fall in a straight line and the scatter diagram will be as in Fig. 116.1a The mathematical relationship between x and y is $y = a + bx$.
2. $r = -1$ Indicates a perfect negative relationship between x and y. In this case all the values of x and y fall in a straight line and the scatter diagram will be as in Fig. 11.6.1b .The mathematical relationship between x and y is $y = a - bx$.
3. $r = 0$ Means there is no linear relationship between the variables x and y. In this case the two variables are linearly independent. The scatter diagram will be as Fig. 11.6.1e and Fig.11.6.1f.
4. The closer r to +1 or -1, the closer the relationship between the variables x and y. The closer r to zero, the less close the relationship. In these cases, the scatter diagrams will be as Fig. 11.6.1c and 11.6.1d.

Now we shall cite some examples to show the different values of r.

Example 11.6.1 Compute r for the following paired sets of values :

a

x	1	2	3	4	5
y	1	3	5	7	9

b

x	1	2	3	4	5
y	2	3	5	4	7

c

x	1	2	3	4	5
y	10	8	6	4	2

d

x	2	3	4	5	6
y	9	5	6	2	1

e

x	1	2	3	4	5
y	3	2	2	2	3

f

x	-2	-1	0	1	2
y	4	1	0	1	4

Solution. (a) The computing formula for finding Karl Pearson's correlation coefficient is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}}$$

Let us make a table to calculate correlation coefficient.

Table for calculating correlation coefficient

x	y	x^2	y^2	xy
1	1	1	1	1
2	3	4	9	6
3	5	9	25	15
4	7	16	49	28
5	9	25	81	45
15	25	55	165	95

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{95 - \frac{15\times 25}{5}}{\sqrt{\left\{55 - \frac{(15)^2}{5}\right\}\left\{165 - \frac{(25)^2}{5}\right\}}} = \frac{20}{\sqrt{10\times 40}} = \frac{20}{20} = 1.$$

Conclusion. Here there exists a perfect and positive relationship between x and y. In this case increment of the value of y for unit change of x is the same. All the points of the scatter diagr will fall on a straight line. The mathematical relationship between x and y is $y = a + bx$.

(b) The formula for finding correlation coefficient is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}}.$$

Table for calculating correlation coefficient

x	y	x^2	y^2	Xy
1	2	1	4	2
2	3	4	9	6
3	5	9	25	15
4	4	16	16	16
5	7	25	49	35
15	21	55	103	74

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{74 - \frac{15\times 21}{5}}{\sqrt{\left\{55 - \frac{(15)^2}{5}\right\}\left\{103 - \frac{(21)^2}{5}\right\}}} = \frac{11}{\sqrt{10\times 14.8}} = \frac{11}{12.17} = 0.90.$$

Conclusion. There exists a strong positive relationship between x and y, since the value of r is 0.9 which is close to +1.

(c) The computing formula for finding Karl Pearson's correlation coefficient is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}}.$$

Let us make a table to calculate correlation coefficient.

x	y	x^2	y^2	Xy
1	10	1	100	10
2	8	4	64	16
3	6	9	36	18
4	4	16	16	16
5	2	25	4	10
15	30	55	220	70

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{70 - \frac{15\times 30}{5}}{\sqrt{\left\{55 - \frac{(15)^2}{5}\right\}\left\{220 - \frac{(30)^2}{5}\right\}}} = \frac{-20}{\sqrt{10\times 40}} = \frac{-20}{20} = -1.$$

Conclusion. Here there exists a perfect negative relationship between x and y. The decrease of the value of y for unit change of x is the same. All the points of the scatter diagram will fall on a straight line, which starts at upper left-hand corner and comes down to the lower right-hand corner of diagram. The mathematical relationship between x and y is $y = a - bx$.

(d) The working formula for finding correlation coefficient is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}}.$$

Let us make a table to calculate correlation coefficient.

x	y	x^2	y^2	Xy
2	9	4	81	18
3	5	9	25	15
4	6	16	36	24
5	2	25	4	10
6	1	36	1	6
20	23	90	147	73

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{73 - \frac{20\times 23}{5}}{\sqrt{\left\{90 - \frac{(20)^2}{5}\right\}\left\{147 - \frac{(23)^2}{5}\right\}}} = \frac{20}{\sqrt{10\times 41.2}} = \frac{-19}{20.3} = -0.94.$$

Conclusion. There exists a strong negative relationship between x and, since the value o 0.94 which is close to –1.

(e) The computing formula for finding r is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}}$$

Calculation table for correlation coefficient

x	y	x^2	y^2	xy
1	3	1	9	3
2	2	4	4	4
3	2	9	4	6
4	2	16	4	8
5	3	25	9	15
15	12	55	30	36

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{36 - \frac{15\times 12}{5}}{\sqrt{\left\{55 - \frac{(15)^2}{5}\right\}\left\{36 - \frac{(12)^2}{5}\right\}}} = \frac{0}{\sqrt{10\times 7.2}} = 0.00.$$

Conclusion. Here there exists no linear relationship between the variables x and y. That is th variables x and y are linearly independent.

(f) The formula for finding correlation r is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}}$$

Let us make a table to calculate correlation coefficient.

x	y	x^2	y^2	xy
–2	4	4	16	–8
–1	1	1	1	–1
0	0	0	0	0
1	1	1	1	1
2	4	4	16	8
0	10	10	34	0

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{0 - \frac{0 \times 10}{5}}{\sqrt{\left\{10 - \frac{(0)^2}{5}\right\}\left\{34 - \frac{(10)^2}{5}\right\}}} = \frac{0}{\sqrt{10 \times 14}} = 0.$$

Conclusion. Here the value of r is zero. But there exists a perfect non-linear relationship between x and y. Actually, the relationship between x and y is $y = x^2$. This means simple correlation coefficient cannot measure the strength of non-linear relationship between x and y.

We can measure the non-linear relationship between x and y by correlation ratio, which is, beyond the scope of this book.

11.7 Probable Error of Correlation Coefficient

If r is the correlation coefficient in a sample of n pairs of observations, then its standard error is $S.E(r) = \frac{1-r^2}{\sqrt{n}}$.

The probable error (P.E.) of the coefficient of correlation helps in interpreting its value.

Probable of error (P.E.) of coefficient of correlation is $P.E.(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}}$.

Here r is the sample correlation coefficient and n is the number of pairs of observations.

The probable error (P.E.) of the coefficient of correlation helps in interpreting its value.

Probable error is an old measure for testing the reliability of an observed correlation coefficient. The reason for taking the factor 0.6745 is that in a normal distribution, the range $\mu \pm 0.6745\sigma$ covers 50% of the total area.

Conclusion.

1. If the value of r is less than probable error, there is no evidence of correlation between the variables i.e., the value of r is not at all significant.
2. If the value of r is more than six times the value of probable error, the existence of correlation is practically certain, i.e., the value of r is significant.
3. The population correlation coefficient ρ is expected to lie in the interval $r \pm P.E.(r)$.

Remarks. Now a day t-test is used to test the significance of an observed correlation coefficient which will be discussed in Chapter 16.

Example 11.7.1. The following data relate to advertising expenditure (in lakhs of taka) and sales (in crores of taka) of a firm:

Advertising expenditure (in lakhs of Tk.)	10	12	15	20	23
Sales (in crores of Tk.)	14	17	23	21	25

Compute the coefficient correlation between advertising expenditures and sales and comment on the value of r.

Solution.

Table for calculation of correlation coefficient

Advertising expenditure: x	Sale : y	x^2	y^2	xy
10	14	100	196	140
12	17	144	289	204
15	23	225	529	345
20	21	400	441	420
23	25	529	525	575
80	100	1396	2080	1684

Coefficient of correlation,

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma y - (\Sigma y)^2\}}} = \frac{(5\times 1684) - (80\times 100)}{\sqrt{\{(5\times 1398) - (80)^2\}\{(5\times 2080) - (100)^2\}}},$$

$$= \frac{8420 - 8000}{\sqrt{(6990 - 6400)(10400 - 100000)}} = \frac{420}{\sqrt{(590\times 400)}} = \frac{420}{\sqrt{23600}} = \frac{420}{4855} = 0.864.$$

Here probable error of r is

$$\text{P.E.}(r) = 0.6745 \times \frac{1-r^2}{\sqrt{n}} = 0.6745 \times \frac{1-(.864)^2}{\sqrt{5}} = \frac{0.254}{2.236} = 0.077.$$

Six times of the P.E.(r) = .6 × 0.077 = 0.462 which is less than the value of correlation coefficient. Hence the value of r is significant. That means, there exists a strong positive relationship between x and y.

Example 11.7.2 A researcher wants to find out if there is any relationship between the ages of husbands and the ages of wives. In other words, do old husbands have old wives and young husbands have young wives? He took a random sample of 7 couples whose respective ages are given below:

Age of husbands :x	25	27	29	32	35	37	39
Ages of wives :y	18	20	20	25	25	30	37

Compute coefficient of correlation between the ages of husbands and the ages of wives.

Solution. The computing formula for finding the correlation coefficient is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}}.$$

Computation table for finding the value of r

x	y	x^2	y^2	xy
25	18	625	324	450
27	20	729	400	540
29	20	841	400	580
32	25	1024	625	800
35	25	1225	625	875
37	30	1369	900	1110
39	37	1521	1369	1443
$\Sigma x = 224$	$\Sigma y = 175$	$\Sigma x^2 = 7334$	$\Sigma y^2 = 4643$	$\Sigma xy = 5798$

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{5798 - \frac{224 \times 175}{7}}{\sqrt{\left\{7334 - \frac{(224)^2}{7}\right\}\left\{4643 - \frac{(175)^2}{7}\right\}}}$$

$$= \frac{5798 - 5600}{\sqrt{166 \times 268}} = \frac{198}{21092} = 0.94.$$

Conclusion. Here probable error of r is

$$P.E.(r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}} = 0.6745 \times \frac{1 - (0.94)^2}{\sqrt{7}} = \frac{0.1164}{2.646} = 0.044.$$

Six times of the P.E.(r) = 6×0.044 = 0.2639 which is less than the value of correlation coefficient. Hence the value of r is significant and there exists a very strong relationship between the ages of husbands and the ages of wives.

Example 11.7.3. The following table gives the prices and consumption of salts of a family for the last 5 months.

Price per kg (in taka)	5	6	7	8	9
Consumption in kg	4	4	4	4	4

Find covariance between the prices and consumption of salt and comment.

Solution.

Price: x	Consumption : y	xy
5	4	20
6	4	24
7	4	28
8	4	32
9	4	36
35	20	140

The computing formula for finding covariance is

$$\text{Covariance} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \frac{1}{n}\left\{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}\right\} = \frac{1}{5}\left\{140 - \frac{35\times 20}{5}\right\} = \frac{1}{5}(140-140) = 0.$$

Comment. Here prices of the salt and consumptions of salt are independent since the covariance between them is zero, Correlation coefficient between them will also be zero.

When the mean of the variables are whole number, then it is quite easy to find the value of the correlation coefficient by the original formula. Now we shall cite some examples.

Calculation of Correlation Coefficient using the Original Formula

The sample correlation coefficient is defined by

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}} = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} \quad ; \text{ where } X = (x-\bar{x}) \text{ and } Y = (y-\bar{y}).$$

This formula is very useful when the values of x and y are large and the means of x and y are whole number.

(i) When the means of x and y are both integer

Example 11.7.4 The following data refer to the sales and expenses of 10 firms in lakh taka for

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

Compute correlation coefficient and comment.

Solution. $\Sigma x = 580, \Sigma y = 140; n = 10.$ $\bar{x} = \frac{\Sigma x}{n} = \frac{580}{10} = 58;$ $\bar{y} = \frac{\Sigma y}{n} = \frac{140}{10} = 14.$

Here both the means of x and y are integer. So we can comfortably use the above formula for finding correlation coefficient.

Table for computation of correlation coefficient

Firm	Sales x	$X=(x-\bar{x})$ $=x-58$	X^2	Expenses y	$Y=(y-\bar{y})$ $=y-14$	Y^2	XY
1	65	7	49	16	2	4	14
2	65	7	49	15	1	1	7
3	65	7	49	15	1	1	7
4	60	2	4	14	0	0	0
5	60	2	4	13	-1	1	-2
6	50	-8	64	13	-1	1	8
7	60	2	4	16	2	4	4
8	55	-3	9	14	0	0	0
9	50	-8	64	13	-1	1	8
10	50	-8	64	11	-3	9	24
	$\Sigma x = 580$	$\Sigma X = 0$	$\Sigma X^2 = 360$	$\Sigma y = 140$	$\Sigma Y = 0$	$\Sigma Y^2 = 22$	$\Sigma XY = 70$

Correlation coefficient, $r = \frac{\Sigma XY}{\sqrt{\Sigma X \Sigma Y}} = \frac{70}{\sqrt{360 \times 22}} = \frac{70}{88.994} = 0.787.$

There is a high degree of positive correlation between the variables. That is, as the sales go up expenses also go up.

(ii) When the means of x and y are not integer.

When the values of x and y are large and the actual means are in fractions, say the actual means of x and y are 24.23 and 12.57 respectively, the calculation of coefficient of correlation by the original formula would involve too many calculations and would take a lot of time. In such cases we can only shift the origins of x and y to some constants A and B which are very near to the original means of x and y to get the maximum benefit of calculations. The formula for finding correlation coefficient is

$$r = \frac{\Sigma uv - \frac{(\Sigma u \times \Sigma v)}{n}}{\sqrt{\left\{\Sigma u^2 - \frac{(\Sigma u)^2}{n}\right\}\left\{\Sigma v^2 - \frac{(\Sigma v)^2}{n}\right\}}}$$ Here $u = x - A$; $v = y - B$

Here A is the assumed mean of x which is usually taken as a whole number and very near to the actual mean of x. Similarly, B is the assumed mean of y which is usually taken as a whole number and very near to the actual mean of y. Now we shall cite one example.

Example 11.7.5. Calculate Karl Pearson's coefficient of correlation for the following data and interpret its value

X	112	116	103	116	98	118	112	104	111	105
Y	65	69	60	68	56	72	60	53	64	62

Solution. Here, $\Sigma x = 1095$, $\Sigma y = 629$ and $n = 10$.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1095}{10} = 1095; \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{629}{10} = 629.$$

Here the means of x and y are in fractions. We can take 105 and 60 as assumed means of x and y and construct the following table :

x	u = x - 105	u^2	y	v = y - 60	v^2	uv
112	7	49	65	5	25	35
116	11	121	69	9	81	99
103	-2	4	60	0	0	0
116	11	121	68	8	64	88
98	-7	49	56	-4	16	28
118	13	169	72	12	144	156
112	7	49	60	0	0	0
104	-1	1	53	-7	49	7
111	6	36	64	4	16	24
105	0	0	62	2	4	0
$\Sigma x = 1095$	$\Sigma u = 45$	$\Sigma u^2 = 599$	$\Sigma y = 629$	$\Sigma v = 29$	$\Sigma v^2 = 399$	$\Sigma uv = 437$

$$r = \frac{\Sigma uv - \frac{(\Sigma u \times \Sigma v)}{n}}{\sqrt{\left\{\Sigma u^2 - \frac{(\Sigma u)^2}{n}\right\}\left\{\Sigma v^2 - \frac{(\Sigma v)^2}{n}\right\}}} = \frac{437 - \frac{(45)(29)}{10}}{\sqrt{\left\{599 - \frac{(45)^2}{10}\right\}\left\{399 - \frac{(29)^2}{10}\right\}}}$$

$$= \frac{437 - 1305}{\sqrt{(599 - 2025)(399 - 84.1)}} = \frac{3065}{35335} = 0.867.$$

There exists a strong positive relationship between the two variables x and y.

Example 11.7.6 The following data give the advertising expenditure and sales of a firm for ten months:

Month	Expenditure : x	Sales : y	Month	Expenditure : x	Sales : y
Jan.	50	1,600	June	150	2600
Feb.	60	2000	July	140	2800
March	70	2200	Aug	160	2900
April	90	2500	Sept.	170	3100
May	120	2400	Oct.	190	3900

Compute the correlation coefficient.

Solution. Here, $\bar{x} = \frac{1200}{10} = 120$ *and* $\bar{y} = \frac{26000}{10} = 2600$.

Both the means of x and y are integers. Moreover, the values of x are some multiple of 10 and the values of y are some multiple of 100. So we can define two new variables u and v as

$$u = \frac{x-120}{10}; \qquad v = \frac{y-2600}{100}.$$

Here, A = 120, B = 2600, h = 10 and k = 100.

Computation of correlation Coefficient

x	$u=\frac{x-120}{10}$	u^2	y	$v=\frac{y-2600}{100}$	v^2	uv
50	-7	49	1600	-10	100	70
60	-6	36	2000	-6	36	36
70	-5	25	2200	-4	16	20
90	-3	9	2500	-1	1	3
120	0	0	2400	-2	4	0
150	3	9	2600	0	0	0
140	2	4	2800	2	4	4
160	4	16	2900	3	9	12
170	5	25	3100	5	25	25
190	7	49	3900	13	169	91
$\Sigma x = 1200$	$\Sigma u = 0$	$\Sigma u^2 = 222$	$\Sigma Y = 26{,}000$	$\Sigma v = 0$	$\Sigma v^2 = 364$	$\Sigma uv = 261$

$$r = \frac{\Sigma uv}{\sqrt{u^2 \times v^2}} = \frac{261}{\sqrt{222 \times 364}} = \frac{261}{28427} = 0.918.$$

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

11.8 Rank Correlation

The British Psychologist Edward Spearman developed Spearman's rank correlation coefficient in 1904. This method is applied in a situation in which quantitative measure of certain qualitative factors such as beauty, intelligent, judgment, leadership, colour, taste, cannot be fixed but individual observations can be arranged in a definite order called rank.

Definition. Suppose 1, 2, ..., n are assigned ranks to the x observations in order of magnitude and similarly 1, 2,..., n are assigned ranks to the y observations. Then the simple correlation coefficient between the two sets of ranks is called Spearman's rank correlation coefficient. It is denoted by R.

When there are no ties among either set of observations, then the formula for computing the Spearman's rank correlation coefficient is $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$.

Here, R = rank correlation coefficients;

R_1 = Rank of observations with respect to first variable x;

R_2 = Rank of observations with respect to second variable y;

$d = R_1 - R_2$, difference in a pair of ranks;

n = number of pairs of observations being ranked.

Actually, Spearman's rank correlation coefficient is a nonparametric counterpart of the k Pearson's simple correlation coefficient r.

If there are no ties among either set of observations, the value of R will usually be very cl to the value of r based on numerically observations and is interpreted in much the same way. T value of R will range from –1 to +1. A value of +1 or –1 indicates perfect association between and y, the plus sign occurring for identical rankings and the minus sign occurring rever rankings. When R is close to zero, we would conclude that the variables are uncorrelated.

Remarks. We always have, $\Sigma d_i = \Sigma(R_1 - R_2) = \Sigma R_1 - \Sigma R_2 = 0$.

This serves as a check on the calculation.

Properties of rank correlation coefficient.

1. Like simple correlation coefficient, rank correlation coefficient lies between –1 to +1.
2. R = 1, when the ranks of x completely agree with the ranks of y, i.e. $(R_1, R_2) = (1, 1), (2, 2), \ldots, (n, n)$.
3. R= –1, when there is complete disagreement in the ranks, in this case $(R_1, R_2) = (1, n), (2, n-1), \ldots, (n, 1)$.
4. This is the only method for finding relationship between two qualitative variables li beauty, honesty, intelligence, efficiency and so on.
5. This is the only method for finding relationship between two variables when ranks a available.
6. If there exist no ties, then simple correlation coefficient and rank correlation coefficient diffe slightly.
7. It is a distribution free method which does not need any assumption about the population b test.

Limitations. This method cannot be used for finding correlation in a grouped frequen distribution.

For finding rank correlation coefficient, we may have two types of data (i) Actua observations are given, and (ii) Actual ranks are given.

Now we shall cite some examples to show how the different values of R changes.

Example 11.8.1 Suppose that two managers I and II are to rank 5 employees A, B, C, D and E the basis of their performance in a test and the results are recorded as follows :

(a)

Examiner/Employees	A	B	C	D	E
I	1	2	3	4	5
II	1	2	3	4	5

(b)

Examiner/Employees	A	B	C	D	E
I	1	2	3	4	5
II	2	4	1	5	4

(c)

Examiner/Employees	A	B	C	D	E
I	1	2	3	4	5
II	5	4	3	2	1

(d)

Examiner/Employees	B	C	D	A	E
I	2	3	4	1	5
II	3	4	1	5	2

(e)

Examiner/Employees	B	C	D	A	E
I	2	3	4	1	5
II	3	2	1	4	5

Solution. (a) The formula for computing rank correlation coefficient is: $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

Table for calculation

Ranking by manager I : R_1	Ranking by manager II : R_2	$d^2=(R_1-R_2)^2$
1	1	0
2	2	0
3	3	0
4	4	0
5	5	0
		$\Sigma d^2 = 0$

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{0}{5\times 24} = 1$$

Conclusion. There is a perfect positive ranking between the managers. That is, there is a full agreement between the two managers regarding the ranking of the employees.

(b) The formula for computing rank correlation coefficient is: $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

Ranking by manager I : R_1	Ranking by manager II : R_2	$d^2=(R_1-R_2)^2$
1	2	1
2	3	1
3	1	4
4	5	1
5	4	1
		$\Sigma d^2 = 8$

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 8}{5\times 24} = 1 - \frac{48}{120} = 1 - 0.4 = 0.6.$$

Conclusion. There is a positive rank correlation coefficient between the rankings of two managers.

(c) The formula for computing rank correlation coefficient is: $R = 1 - \frac{6\sum d^2}{n(n^2-1)}$.

Ranking by manager I : R_1	Ranking by manager II : R_2	$d^2 = (R_1 - R_2)^2$
1	5	16
2	4	4
3	3	0
4	2	4
5	1	16
		$\Sigma d^2 = 40$

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 40}{5\times 24} = 1 - \frac{240}{120} = 1 - 2 = -1.$$

Conclusion. There is a perfect negative relationship between the rankings of the two managers. That is, there is a full disagreement between rankings of the two managers.

(d) The formula for computing rank correlation coefficient is: $R = 1 - \frac{6\sum d^2}{n(n^2-1)}$.

Ranking by manager I : R_1	Ranking by manager II : R_2	$d^2 = (R_1 - R_2)^2$
1	5	16
2	3	1
3	4	1
4	1	9
5	2	9
		$\Sigma d^2 = 36$

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times 36}{5\times 24} = 1 - \frac{216}{120} = 1 - 1.8 = -0.8.$$

Conclusion. There is a strong negative rank correlation coefficient between the rankings of two managers.

(e) The formula for computing rank correlation coefficient is: $R = 1 - \frac{6\sum d^2}{n(n^2-1)}$.

Ranking by manager I : R_1	Ranking by manager II : R_2	$D^2 = (R_1 - R_2)^2$
2	3	1
3	2	1
4	1	9
1	4	9
5	5	0
		$\Sigma d^2 = 20$

$$R = 1-\frac{6\sum d^2}{n(n^2-1)} = 1-\frac{6\times 20}{5\times 24} = 1-\frac{120}{120} = 1-1 = 0.$$

Comment. In this case, the ranking of the two managers are independent.

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

Example 11.8.2. A Social Scientist wants to see whether there is any association between the intelligence and beauty among the female students. To verify this he randomly selected 6 female students from a class. The scores on intelligence and beauty are found as follows:

Student	A	B	C	D	E	F
Scores on intelligence	80	75	90	70	65	60
Scores on beauty	65	70	60	75	85	80

Compute rank correlation coefficient and comment.

Solution. The formula for computing R is: $R = 1-\frac{6\sum d^2}{n(n^2-1)}$.

Here ranks of the scores are not given. Let us start ranking from the highest value for both the variables as shown in the table given below:

Student	Scores on intelligence : x	Scores on beauty : y	Ranks on intelligence : R_1	Ranks on beauty : R_2	$d^2 = (R_1-R_2)^2$
A	80	65	2	5	9
B	75	70	3	4	1
C	90	60	1	6	25
D	70	75	4	3	1
E	65	85	5	1	16
F	60	80	6	2	16
Total					$\Sigma d^2 = 68$

$$R = 1-\frac{6\sum d^2}{n(n^2-1)} = 1-\frac{6\times 68}{6\times 35} = 1-\frac{68}{35} = 1-1.94 = -0.94.$$

Conclusion. There exists a very strong negative relationship between the intelligence and beauty.

Example 11.8.3. Two managers are asked to rank a group of employees in order of potential for eventually becoming top managers. The scores of the two managers are as follows:

Employee	A	B	C	D	E
Scores by Manager I	80	72	70	75	65
Scores by Manager II	75	69	71	78	67

Solution. The formula for computing R is: $R = 1-\frac{6\sum d^2}{n(n^2-1)}$.

Here ranks of the scores are not given. Let us start ranking from the highest value for both the variables as shown in the table given below :

Employees	Scores on manager I : x	Scores on Manager II : y	Ranks by manager I : R_1	Ranks by manager II : R_2	$d^2 = (R_1 - R_2)^2$
A	80	75	1	2	1
B	72	69	3	4	1
C	70	71	4	3	1
D	75	78	2	1	1
E	65	67	5	5	0
		Total			$\Sigma d^2 = 4$

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 4}{5\times 24} = 1 - \frac{1}{5} = 1 - 0.2 = 0.8.$$

Conclusion. There exists a strong positive relationship between the rankings of the two managers.

Example 11.8.4 Two house wives, Rahima and Maksuda were asked to express their preference for different kinds of detergents, gave the following replies :

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

Compute rank correlation and compare how far their preferences go together.

Solution. Here we will compute rank correlation coefficient.

Table for computing rank correlation coefficient

Detergent	R_1	R_2	d^2
A	10	10	0
B	9	9	0
C	5	6	1
D	6	5	1
E	8	7	1
F	7	8	1
G	3	3	0
H	1	2	1
I	2	1	1
J	4	4	0
n = 10			$\Sigma d^2 = 6$

Rank correlation coefficient, $R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 6}{10\times 99} = 1 - \frac{1}{990} = 1 - 0.036 = 0.964.$

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

11.8.1 Rank correlation coefficient for repeated ranks or ties observations. If there is more than one observation either x or y are same, then the Spearman's formula for calculating the rank correlation coefficient breaks down. In this case, common ranks are given to the repeated observations. This common rank is the average of the ranks, which these observations would have assumed, and the observation will get the rank next to the ranks already assumed. As a result a correction term is added in the rank correlation formula. In the formula, we add the factor $\frac{m(m^2-1)}{12}$ to Σd^2, where m is the number of times an observation is repeated. This correction factor is to be added for each repeated value. Now, we shall cite an example.

When ranks are repeated the following formula is used for finding rank correlation coefficient

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\ldots\right\}}{n(n^2-1)}$$

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}\Sigma(m_i^3 - m_i)\right\}}{n(n^2-1)}$$

Where m_i refers to the number of times ith value repeated.

Example 11.8.5. The following data refer to the marks obtained by 8 students in mathematics and statistics:

Marks in Mathematics	20	80	40	12	28	20	15	60
Marks in Statistics	30	60	20	30	50	30	40	20

Compute rank correlation coefficient and comment.

Solution. Let the marks obtained by mathematics be X and the marks obtained by Statistics be Y. Here let us ranking from the lowest values for both the variables.

Table for computation of rank correlation

X	R_x	Y	R_y	d^2
20	3.5	30	4	0.25
80	8	60	8	0.00
40	6	20	2	16.00
12	1	30	4	9.00
28	5	50	7	4.00
20	3.5	30	4	0.25
15	2	40	6	16.00
60	7	10	1	36.00

Here, $\Sigma d^2 = 81.5$. In series X, 20 have come two times and in series Y, 30 have come three times. The necessary adjustment for this repeated rank has to be made. Hence adjusted formula for rank correlation coefficient is

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\ldots\right\}}{n(n^2-1)}$$

Here, $\Sigma d^2 = 81.5$, $m_1 = 2$, $m_2 = 3$ and $n = 8$.

$$R = 1 - \frac{6\left\{81.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right\}}{8(8^2-1)} = 1 - \frac{6 \times 84}{504} = 1 - \frac{504}{504} = 0.$$

Comment. Marks obtained by mathematics and statistics are independent.

Example 11.8.6. The data refer to the marks obtained by a student in ten subjects by two examiners.

Examiner I	68	64	75	50	64	80	75	40	55	64
Examiner II	62	58	68	45	81	60	68	48	50	70

Find rank correlation coefficient.

Solution. Suppose the marks by examiner I is x and the examiner II is y. Let us start ranking by the highest value for both the variables.

Table for computing rank correlation coefficient R

Marks of Examiner I : x	Marks of Examiner II : y	Ranks of marks by Examiner I : R_1	Ranks of marks by Examiner II : R_2	$d = R_1 - R_2$	d^2
68	62	4	5	-1	1
64	54	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				$\Sigma d = 0$	$\Sigma d^2 =$

In the observations for x we see that the value 75 occurs 2 times. The common rank give these values is 2.5, which is the average of 2 and 3, the ranks that these values would have t if they were different. The next value 68 then gets the next rank, which is 4. Again, we see value 64 occurs thrice. The common rank given to it is 6, which is the average of 5, 6 a Similarly, in the observations for y we see that the value 68 occurs twice and its common ra 3.5, which is the average of 3 and 4. As a result of these common rankings, the formula for R to be corrected. To Σd^2 we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the numb times a value occurs. Correction for x is to be applied twice, once for the value-75 that oc

...ice ($m_1 = 2$) and then for the value 64, which occurs thrice ($m_2 = 3$). The total correction for x

$$\frac{2(2^2-1)}{12}+\frac{3(3^2-1)}{12}=\frac{6}{12}+\frac{24}{12}=\frac{5}{2}=2.5.$$

Similarly, the correction for y is: $\frac{2(2^2-1)}{12}=\frac{1}{2}=0.5.$

Thus, $R=1-\frac{6[\sum d^2+2.5+0.5]}{n(n^2-1)}=1-\frac{6(72+3)}{10\times 99}=1-\frac{450}{990}=1-0.455=0.545$

8.2. Advantages of rank correlation coefficient over simple correlation coefficient.

Rank correlation coefficient can be safely used in case of linear and curvilinear relationship between two variables x and y. But simple correlation coefficient measures only the strength of linear relationship between the variables.

No assumption of normality is required for testing the significance of R whereas the assumption of normality is required to test the significance of the sample correlation coefficient.

Rank correlation coefficient can be used for finding the association between two qualitative as well as two quantitative variables, whereas simple correlation coefficient is used for finding linear relationship between two quantitative variables only.

Rank correlation coefficient is easy to understand and apply as compared with simple correlation coefficient.

If there are no ties among the either set of observations or no many ties exist, the rank correlation coefficient is slightly over than the simple correlation coefficient.

...mple 11.8.7. The scores of eight students in Statistics and English in an examination are given ...low.

Students Number	1	2	3	4	5	6	7	8
Marks in Statistics	52	60	50	54	55	58	48	70
Marks in English	48	51	68	55	60	53	47	62

...Compute (i) Simple correlation coefficient and (ii) Rank correlation coefficient.

...lution. *(i) Let the marks in statistics be denoted by x and English by y.*

Here $\sum x = 447, \sum y = 444$ and $n = 8$. $\bar{x}=\frac{\sum x}{n}=\frac{447}{8}=55.88$ *and* $\bar{y}=\frac{\sum y}{n}=\frac{444}{8}=55.5.$

Here both the means of x and y are fractions. Let us take the assumed mean for x as 55 ...for y as 55 which are very close to their actual means.

Computation table for correlation coefficient

X	$d_x = x-55$	d_x^2	y	$d_y = y-55$	d_y^2	$d_x d_y$
52	-3	9	48	-7	49	21
60	5	25	51	-4	16	-20
50	-5	25	68	13	169	-65
54	-1	1	55	0	0	0
55	0	0	60	5	25	0
58	3	9	53	-2	4	-6
48	-7	49	47	-8	64	56
70	15	225	62	7	49	105
$\Sigma x = 447$	$\Sigma d_x = 7$	$\Sigma d_x^2 = 343$	$\Sigma y = 444$	$\Sigma d_y = 4$	$\Sigma d_y^2 = 376$	$\Sigma d_x d_y = 91$

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x \times \Sigma d_y)}{n}}{\sqrt{\left\{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{n}\right\}\left\{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{n}\right\}}} = \frac{91 - \frac{7\times 4}{8}}{\sqrt{\left\{343 - \frac{(7)^2}{8}\right\}\left\{376 - \frac{(4)^2}{8}\right\}}}$$

$$= \frac{91-3.5}{\sqrt{(343-6.125)(376-2)}} = \frac{87.5}{\sqrt{336875\times 374}} = \frac{87.5}{355495} = 0.25.$$

(ii)

Calculation of rank correlation coefficient

X	R_x	y	R_y	d^2
52	6	48	7	1
60	2	51	6	16
50	7	68	1	36
54	5	55	4	1
55	4	60	3	1
58	3	53	5	4
48	8	47	8	0
70	1	62	2	1
				$\Sigma d^2 = 60$

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6\times 60}{8\times 63} = 1 - \frac{360}{504} = 1 - 0.714 = 0.29$$

Comment. Here it is seen that rank correlation coefficient is slightly more than the sim correlation coefficient.

Limitations. This method cannot be used for finding out correlation in a bivariate freque distribution. Usually for $n > 30$, this formula should not be used unless the ranks are gi since in the contrary case the calculations are quite time-consuming.

Example 11.8.8 Two judges have ranked 12 students in order of their merits as follows:

Students	A	B	C	D	E	F	G	H	I	J	K	L
Rank by 1st Judge	5	2	4	1	8	9	10	6	3	11	7	12
Rank by 2nd judge	6	9	7	10	1	2	4	12	3	5	11	8

Calculate rank correlation coefficient to find out whether the judges are in agreement with each other or not.

Solution.

Calculation table

Student	R_1	R_2	d^2
A	5	6	1
B	2	9	49
C	4	7	9
D	1	10	81
E	8	3	25
F	9	2	49
G	10	4	36
H	6	12	36
I	3	3	0
J	11	5	36
K	7	11	16
L	12	8	16
			$\Sigma d^2 = 354$

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 354}{12(12^2-1)} = 1 - \frac{2124}{1716} = 1 - 1.237 = -0.237$$

There is a negative rank correlation coefficient that indicates that the judges are not in agreement with each other.

Example 11.8.9 Two bank officers examined eleven loan applications and ranked them,

Applicants	A	B	C	D	E	F	G	H	I	J	K
Officer-I	1	7	4	2	3	6	5	9	10	8	11
Officer-II	1	6	5	2	3	4	7	11	8	10	9

Compute rank correlation coefficient and comment.

Solution.

Applicants	R_1	R_2	d^2
A	1	1	0
B	7	6	1
C	4	5	1
D	2	2	0
E	3	3	0
F	6	4	4
G	5	7	4
H	9	11	4
I	10	8	4
J	8	10	4
K	11	9	4

$$R = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 26}{11(11^2-1)} = 1 - \frac{156}{1320} = 1 - 0.118 = 0.882$$

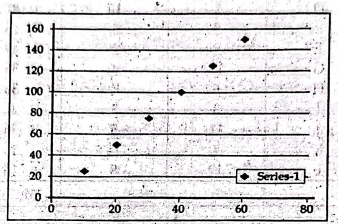
Comment. There is a close agreement about the loan applicants by the two officers since value of the rank correlation coefficient is 0.88.

Example 11.8.10 Given the following pairs of values of the variables X and Y

X	10	20	30	40	50	60
Y	25	50	75	100	125	150

Draw a scatter diagram. Is there any correlation between X and Y. *N.U.BBS.*

Solution. Now we plot the different values of X in the X-axis and the different values of Y in Y-axis.



It is obvious that the scatter diagram is a straight line, since all the points fall in a upw straight line. This shows that there exists a perfect and positive relationship between X and Y. this case the value of the correlation coefficient is 1.

Calculation of correlation coefficient

X	Y	U = (X − 40)/10	V = (Y − 100)/25	U^2	V^2	UV
10	25	−3	−3	9	9	9
20	50	−2	−2	4	4	4
30	75	−1	−1	1	1	1
40	100	0	0	0	0	0
50	125	1	1	1	1	1
60	150	2	2	4	4	4
		−3	−3	19	19	19

The formula for finding the correlation coefficient r is

$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\sqrt{\left\{\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right\}\left\{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right\}}} = \frac{\Sigma UV - \frac{\Sigma U \Sigma V}{n}}{\sqrt{\left\{\Sigma U^2 - \frac{(\Sigma U)^2}{n}\right\}\left\{\Sigma V^2 - \frac{(\Sigma V)^2}{n}\right\}}}$$

$$= \frac{19-\frac{(-3)(-3)}{6}}{\sqrt{\left(19-\frac{9}{6}\right)\left(19-\frac{9}{6}\right)}} = \frac{19-3/2}{\sqrt{(19-3/2)(19-3/2)}} = \frac{17.5}{17.5} = 1.$$

Comment. The value of the correlation coefficient fully agrees with the scatter diagram.

Example 11.8.11. Ten competitors in a music contest are ranked by three Judges in the following order:

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

Use the rank correlation coefficient to discuss which pair of Judges have nearest approach to common testes in music. *N.U. '200*

Solution. The possible rank correlation coefficients are three. They are between first and second Judges, first and third Judges and the second and third Judges.

The formula for finding rank correlation coefficient is: $R = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)}$.

Calculation of rank correlation between first and second Judges

Ranked by Judge-1 : X	Ranked by Judge-2 : Y	d = X - Y	d^2
4	4	0	0
6	8	-2	4
9	7	2	4
5	5	0	0
1	6	-5	25
3	9	-6	36
10	10	0	0
8	3	5	25
7	2	5	25
2	1	1	1
			119

Rank correlation coefficient between Judge-1 and Judge-2 is

$$R = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6\times 119}{10\times 99} = 1 - \frac{714}{990} = 1 - 0.72 = 0.28.$$

Calculation of rank correlation between first and third Judges.

Ranked by Judge-1 : X	Ranked by Judge-3 : Y	d = X – Y	d^2
4	7	–3	9
6	8	–2	4
9	6	3	9
5	1	4	16
1	5	–4	16
3	10	–7	49
10	9	1	1
8	2	6	36
7	3	4	16
2	4	–2	4
			160

The rank correlation coefficient between first Judge and third Judge is

$$R = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6 \times 160}{10 \times 99} = 1 - \frac{960}{990} = 1 - 0.97 = 0.03.$$

Calculation of rank correlation between second Judge and third Judge

Ranked by Judge-2 : X	Ranked by Judge-3 : Y	d = X – Y	d^2
4	7	–3	9
8	8	0	0
7	6	1	1
5	1	4	16
6	5	1	1
9	10	–1	1
10	9	1	1
3	2	1	1
2	3	–1	1
1	4	–3	9
			40

Rank correlation coefficient between Judge-2 and Judge-3 is

$$R = 1 - \frac{6\Sigma d_1^2}{n(n^2-1)} = 1 - \frac{6 \times 4}{10 \times 99} = 1 - \frac{4}{990} = 1 - 0.024 = 0.976.$$

Comment. Second and third Judges have the nearest ranks in judging the tastes in music, so the rank correlation coefficient between the two Judges is very high compare to other two pairs of ranks.

Example 11.8.12. Find the correlation coefficient between age and playing habits of following students and comment.

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

N.U.BBS-2008

Solution. First, we have to find the percentage of regular players and then we have to calculate the correlation coefficient between the age and percentage of regular players.

Age : X	No. of Students	Regular Players	% of Playing Habits : Y	U = X-17	U^2	V=(Y-60)/5	V^2	UV
15	250	200	80	-2	4	4	16	-8
16	200	150	75	-1	1	3	9	-3
17	150	90	60	0	0	0	0	0
18	120	48	40	1	1	-4	16	-4
19	100	30	30	2	4	-6	36	-12
20	80	12	15	3	9	-9	81	-27
				3	19	-12	158	-54

The correlation coefficient between age and playing habits is

$$r=\frac{\Sigma XY-\frac{\Sigma X\Sigma Y}{n}}{\sqrt{\left\{\Sigma X^2-\frac{(\Sigma X)^2}{n}\right\}\left\{\Sigma Y^2-\frac{(\Sigma Y)^2}{n}\right\}}}=\frac{\Sigma UV-\frac{\Sigma U\Sigma V}{n}}{\sqrt{\left\{\Sigma U^2-\frac{(\Sigma U)^2}{n}\right\}\left\{\Sigma V^2-\frac{(\Sigma V)^2}{n}\right\}}}$$

$$=\frac{54-\frac{(3)(-12)}{6}}{\sqrt{\left(19-\frac{9}{6}\right)\left(158-\frac{144}{6}\right)}}=\frac{-54+6}{\sqrt{(19-1.5)(158-24)}}=\frac{-48}{\sqrt{17.5\times134}}=\frac{-48}{48.425}=-0.99.$$

The value of the correlation coefficient $r=-0.99$ shows that there exists a very strong negative relationship between the age and the playing habits.

Remarks. The beauty of finding the correlation coefficient is that we can shift the origin for one variable and at the same time we can shift the origin and change the scale for another variable if necessary, for computation suitability. Usually, change of scale is useful if we have some common multiplier for all the values of a variable. For example, in the above example it is 5 for the variable Y.

Example 11.8.13 Following figures give the rainfall in inches for the last 7-years and the production in mounds for the Rabi crop and kharif crop. Calculate the Karl Pearson's coefficient of correlation between Rainfall and total production

Rainfall in inches	20	22	24	26	28	30	32
Rabi production	16	18	20	32	80	38	35
Kharif Production	16	17	20	16	20	22	20

N.U.BBS - 2007

Solution. First, we have to find the total production and then we have to find the correlation coefficient between the rainfall and the total production. Let rainfall is denoted by X and the total production by Y.

Table for calculation of correlation coefficient

X	U=(X−26)/2	U^2	Rabi production	Kharif Production	Total Production : Y	V=Y−48	V^2	UV
20	−3	9	16	16	32	−16	256	48
22	−2	4	18	17	35	−13	169	26
24	−1	1	20	20	40	−8	64	8
26	0	0	32	16	48	0	0	0
28	1	1	40	20	60	12	144	12
30	2	4	38	22	60	12	144	24
32	3	9	35	20	55	7	49	21
	0	28				6	826	139

The correlation coefficient between Rainfall and Production is

$$r=\frac{\Sigma XY-\frac{\Sigma X\Sigma Y}{n}}{\sqrt{\left\{\Sigma X^2-\frac{(\Sigma X)^2}{n}\right\}\left\{\Sigma Y^2-\frac{(\Sigma Y)^2}{n}\right\}}}=\frac{\Sigma UV-\frac{\Sigma U\Sigma V}{n}}{\sqrt{\left\{\Sigma U^2-\frac{(\Sigma U)^2}{n}\right\}\left\{\Sigma V^2-\frac{(\Sigma V)^2}{n}\right\}}}$$

$$=\frac{139-\frac{(0)(6)}{7}}{\sqrt{\left(28-\frac{0}{7}\right)\left(826-\frac{36}{7}\right)}}=\frac{139}{\sqrt{(28)(826-5.143)}}=\frac{139}{\sqrt{28\times 820857}}=\frac{139}{15160}=0.92$$

The value of the correlation coefficient $r=0.92$ shows that there exists a strong and positive relationship between the rainfall and the production of the crops.

Example 11.8.14. A sample of 12 students in a particular university revealed the following figures for number of hours studied daily and the marks obtained in an examination.

Hours studied	Marks obtained	Hours studied	Marks obtained
9	45	5	55
6	60	8	80
5	65	7	85
6	75	5	45.
4	40	10	85
8	70	4	30

i) What is the relationship between marks obtained in the examination and the number of hours studied daily?

ii) Does more study bring more marks? *C.U. Acct. 2010*

Solution. (i) Let X be the number of hours studied daily and Y be the marks obtained in the examination. The simple correlation coefficient between X and Y is the required relationship.

Calculation of correlation coefficient

X	Y	V=(Y-55)/5	X^2	V^2	XV
9	45	-2	81	4	-18
6	60	1	36	1	6
5	65	2	25	4	10
6	75	4	36	16	24
4	40	-3	16	9	-12
8	70	3	64	9	24
5	55	0	25	0	0
8	80	5	64	25	40
7	85	6	49	36	42
5	45	-2	25	4	-10
10	85	6	100	36	60
4	30	-5	16	25	-20
77		15	537	169	146

The correlation coefficient between X and Y is the same as the correlation coefficient between and V. The correlation coefficient between X and V is

$$r=\frac{\Sigma XV-\frac{(\Sigma X)(\Sigma V)}{n}}{\sqrt{\left(\Sigma X^2-\frac{(\Sigma X)^2}{n}\right)\left(\Sigma V^2-\frac{(\Sigma V)^2}{n}\right)}}=\frac{146-\frac{77\times15}{12}}{\sqrt{\left\{537-\frac{(77)^2}{12}\right\}\left\{169-\frac{(15)^2}{12}\right\}}}$$

$$=\frac{146-96.25}{\sqrt{(537-494.08)(169-18.75)}}=\frac{49.75}{\sqrt{(42.92\times150.25)}}=\frac{49.75}{80.30}=0.62.$$

The value of $r = 0.62$ shows that there exists a moderate positive relationship between number of hours studied daily with the marks obtained in the examination.

(ii) Yes, the value of $r = 0.62$ shows that more study will bring more marks.

Example 11.8.15. Calculate the rank correlation coefficient from the following data :

Y	80	78	75	75	68	67	60	59	68	70
X	12	13	15	18	13	17	15	12	13	11

C.U. Acct. 2010

Solution. We shall first give the ranks to the X and Y series. Ranks are assigned by taking highest value 1 and so on and wherever there are ties for ranks, adjustments are done.

Table for calculation

X	Y	Ranks for X : x	Ranks for Y : y	d = x − y	d^2
80	12	1	8.5	−7.5	56.25
78	13	2	6	−4	16.00
75	15	3.5	3.5	0	00.00
75	18	3.5	1	2.5	06.25
68	13	6.5	6	0.5	00.25
67	17	8	2	6	36.00
60	15	9	3.5	5.5	30.25
59	12	10	8.5	1.5	02.25
68	13	6.5	6	0.5	00.25
70	11	5	10	−5	25.00
					172.5

In series X, 75 and 68 have come two times each and in series Y, 15 have come two times, have come three times and 12 have come two times. The necessary adjustment for this repeated rank has to be made. Hence the adjusted formula for rank correlation coefficient is

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3) + \frac{1}{12}(m_4^3 - m_4) + \frac{1}{12}(m_5^3 - m_5)\right\}}{n(n^2-1)}$$

Here, $\Sigma d^2 = 1725$, $m_1 = 2$, $m_2 = 2$, $m_3 = 2$, $m_4 = 3$ and $m_5 = 2$. Hence,

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3) + \frac{1}{12}(m_4^3 - m_4) + \frac{1}{12}(m_5^3 - m_5)\right\}}{n(n^2-1)}$$

$$= 1 - \frac{6\left\{1725 + \frac{1}{12}(8-2) + \frac{1}{12}(8-2) + \frac{1}{12}(8-2) + \frac{1}{12}(27-3) + \frac{1}{12}(8-2)\right\}}{10 \times 99}$$

$$= 1 - \frac{6\left(1725 + \frac{6}{12} + \frac{6}{12} + \frac{6}{12} + \frac{24}{12} + \frac{6}{12}\right)}{990}$$

$$= 1 - \frac{6\left(1725 - \frac{48}{12}\right)}{990} = 1 - \frac{6(1725 + 4.00)}{990} = 1 - \frac{1059}{990} = 1 - 1.07 = -0.07.$$

The rank correlation coefficient between X and Y is negative but very week, since its value is 0.07.

Example 11.8.16. Sales and earning of 8 companies are given below :

Company	A	B	C	D	E	F	G	H
sales (in million Tk.)	89.2	18.5	38.5	71.7	56.8	11.9	28.6	30.5
Profits (in million Tk.)	6.9	2.5	9.0	12.5	10	2.5	3.5	12.0

i) Portray the above data through a scatter diagram.

ii) Compute correlation coefficient. Interpret the finding. *C.U. Acct. 2012*

Solution. (i) Let X be the sales and Y be the profits of the company. Now we plot the values of X on the X-axis and the values of Y in the Y-axis. Then the scatter diagram of the bivariate data set is shown in the Fig. 11.8.1.

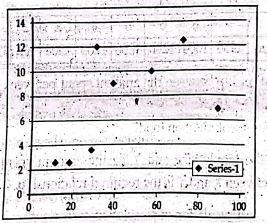


Fig. 11.8.1.

(ii) The simple correlation coefficient between X and Y is the required relationship.

Calculation of correlation coefficient

X	Y	U=(X–56.8)/10	U^2	Y^2	UY
89.2	6.9	3.24	10.50	47.61	22.36
18.5	2.5	–3.83	16.67	6.25	–9.58
38.5	9.0	–1.83	3.35	81.00	–16.47
71.7	12.5	1.49	2.22	156.25	18.63
56.8	10	0.00	0.00	100.00	0.00
11.9	2.5	–5.55	30.80	6.25	–13.88
28.6	3.5	–2.82	7.95	12.25	–9.87
30.5	12.0	–2.63	6.92	144.00	–31.56
	58.9	–11.93	78.41	553.61	–40.37

The correlation coefficient between X and Y is the same as the correlation coefficient between U and Y. The correlation coefficient between U and Y is .

$$r = \frac{\Sigma UV - \frac{(\Sigma U)(\Sigma V)}{n}}{\sqrt{\left\{\Sigma U^2 - \frac{(\Sigma U)^2}{n}\right\}\left\{\Sigma V^2 - \frac{(\Sigma V)^2}{n}\right\}}} = \frac{-40.37 - \frac{(-11.93)(58.9)}{8}}{\sqrt{\left\{78.41 - \frac{(-11.93)^2}{8}\right\}\left\{553.61 - \frac{(58.9)^2}{8}\right\}}}$$

$$= \frac{-40.37 + 87.83}{\sqrt{(78.41 - 17.79)(553.61 - 433.65)}} = \frac{47.46}{\sqrt{(60.62 \times 119.96)}} = \frac{47.46}{85.28} = 0.56$$

The value of $r = 0.56$ shows that there exists a moderate positive relationship between sales and profits.

Example 11.8.17. A sample of 12 families in a particular area revealed the following figures family size and the amount spent on food per week.

Family size.	3	6	5	6	4	3	2	8	4	5	3	1
Amount spent on food	1450	1600	1280	1300	1200	1300	980	3250	1400	1000	620	18

i) What is the relationship between the amount spent per week on food and the size of family?

ii) Do larger families spent more on food? C.U. Acct.)

Solution. (i) Let X be the family size and Y be the amount spent on per week. The sa correlation coefficient between X and Y is the required relationship.

Calculation of correlation coefficient

X	Y	V = (Y-1600)/100	V^2	X^2	XV
3	1450	-1.5	2.25	9	-4.5
6	1600	0.0	0.00	36	0.0
5	1280	-3.2	10.24	25	-16
6	1300	-3.0	9.00	36	-18
4	1200	-4.0	16.00	16	-16
3	1300	-3.0	9.00	9	-12
2	980	-6.2	38.44	4	-12.4
8	3250	16.5	272.25	64	132.0
4	1400	-2.0	4.00	16	-8.0
5	1000	-6.0	36.00	25	-30
3	620	-9.8	96.04	9	-29.4
4	1800	2.0	4.00	16	8.0
53		-22.2	497.22	265	-6.3

The correlation coefficient between X and Y is the same as the correlation coefficient between and V. The correlation coefficient between X and V is

$$r = \frac{\Sigma XV - \frac{(\Sigma X)(\Sigma V)}{n}}{\sqrt{\left(\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right)\left(\Sigma V^2 - \frac{(\Sigma V)^2}{n}\right)}} = \frac{-6.3 - \frac{(53)(-22.2)}{12}}{\sqrt{\left\{265 - \frac{(53)^2}{12}\right\}\left\{497.22 - \frac{(-22.2)^2}{12}\right\}}}$$

$$= \frac{-6.3 + 98.05}{\sqrt{(265 - 234.08)(497.22 - 41.07)}} = \frac{91.75}{\sqrt{(30.92 \times 456.15)}} = \frac{91.75}{118.76} = 0.77.$$

The value of $r = 0.77$ shows that there exists a good positive relationship between the family ze and amount spent per week on food.

The value of $r = 0.77$ tells that larger families spent more on food.

Group-A : Short Questions and Answers

What is Karl Pearson's correlation coefficient?

Ans. Karl Pearson's correlation coefficient or simple correlation coefficient measures the strength of linear relationship between two variables.

What is the range of simple correlation coefficient?

Ans. The range of simple correlation coefficient is -1 to +1.

What is simple coefficient of determination?

Ans. The square of the simple correlation coefficient is called simple coefficient determination. It measures the proportion of dependent variable explained by the independent variable.

What is scatter diagram?

Ans. The simplest device for showing the relationship between two variables on a graph paper in the form of dots is called scatter diagram or scatter plot.

What is rank correlation coefficient?

Ans. The simple correlation coefficient between two sets of ranks is called rank correlation coefficient.

What is the range of rank correlation coefficient?

Ans. The range of rank correlation coefficient is -1 to +1.

What is correlation analysis?

Ans. A technique to determine the degree to which variables are linearly related is called correlation analysis.