Notes

318en10

10

CORRELATION ANALYSIS

In previous lessons you have learnt how to summarize the mass of data and variations in the similar variable. Many a time, we come across situations which involve the study of association among two or more variables. For example we may find that there is some relationship between the two variables such as amount of rain fall and production of wheat; figures of accidents and number of motor cars in a city; money spent on advertising and sales. On the other hand, if we compare the figures of rainfall in India and the production of cars in Japan, we may find that there is no relationship between the two variables. If there is any relation between two variables i.e. when one variable changes the other also changes in the same or in the opposite direction, we say that the two variables are correlated.

OBJECTIVES

After completing this lesson, you will be able to:

- explain the meaning of the term correlation;
- explain the relationship between two variables;
- calculate the different measures of correlation; and
- analyze the degree and direction of the relationships.

10.1 MEANING OF CORRELATION

Correlation refers to the associations between variables. When an association exists between two variables, it means that the average value of one variable changes as there is a change in the value of the other variable. A correlation is the simplest type of association. When a correlation is weak, it means that the average value of one variable changes only slightly (only occasionally) in response to changes in the other variable. If there is no association, it means that there is no

change in the value of one variable in response to the changes in the other variable. In some cases, the correlation may be positive or it may be negative. A positive correlation means that as one variable increases the other variable increases, e.g. Height of a child and age of the child. Negative correlation implies as one variable increases the other variable decrease, e.g. value of a car and age of the car.

10.2 CORRELATION AND CAUSATION

The correlation between two variables measures the strength of the relationship between them but it doesn't indicate the cause and effect relationship between the variables. Correlation measures co-variation, not causation. Causation means changes in one variable affects/ causes the changes in other variable. In other words, just because two events or things occur together does not imply that one is the cause of the other. A positive "linear" correlation between two variables say X and Y implies that high values of X are associated with high values of Y, and that low values of X are associated with low values of Y. It does not imply that X causes Y. for example, a high degree of positive correlation may be obtained between the size of arms of children and their reasoning ability i.e. children with longer arms reason better than those with shorter arms, but there is no causal connection here. Children with longer arms reason better because they're older! In this example the common third factor 'age' is responsible for the high correlation between size of arms and reasoning ability. This refers to spurious correlation. Similarly a Researcher found a high degree of positive correlation between the number of temple goers and the number of burglaries committed in different towns. An explanation that more temple goers means more empty houses or attending temple makes people want to rob would be a logical fallacy. Instead the third factor population is causing this relationship. The highly populated area tends to have more temple goers and also case of burglaries. The following table 11.1 provides some interesting examples of influence of third variable on correlation between variables.

Table 10.1: Spurious Correlation and Influence of Third Variable.

Observed Spurious Correlation	Influence of Third Variable.
Positive Correlation between Amount of ice cream sold and deaths by drowning at the beach during summer.	**Summer Season:** Ice cream sales and drowning tend to be high during the warm months of the year.
Shoe size and reading performance for elementary school children.	**Age:** Older children have larger shoe sizes and read better.
Number of doctors in region and number of people dying from disease.	**Population density:** In highly dense areas, there are more doctors and more people die.

Number of police officers and number of crimes.	**Population density:** In highly dense areas, there are more police officers and more crimes.
Teachers' salaries and the price of vegetables.	**Time:** Both tend to increase over time.

Further, It is found that there is a positive and a high degree of correlation between the amount of oranges imported and road accidents i.e. as the amount of imported oranges increases, so do the traffic fatalities. However, it is fairly obvious just from logical thought that there is likely to be no causal relationship between the two. That is, the importing of oranges does not cause traffic fatalities. Conversely, if we stopped importing oranges, we would not expect the number of traffic fatalities to decline. It may be a sheer coincidence that a high degree of correlation is obtained between them.

10.3 TYPES OF CORRELATION

Correlation may be:

1. Positive and negative correlation
2. Linear and non-linear correlation

A) If two variables change in the same direction (i.e. if one increases the other also increases, or if one decreases, the other also decreases), then this is called a **positive correlation.** For example: Advertising and sales.

Some other examples of series of positive correlation are:

(i) Heights and weights;

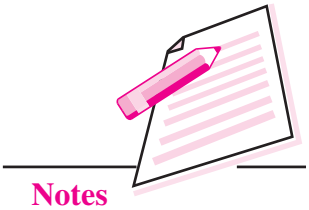
(ii) Household income and expenditure;

(iii) Price and supply of commodities;

(iv) Amount of rainfall and yield of crops.

INTEXT QUESTIONS 10.1

1. It has been noted that there is a positive correlation between the I.Q. level and the size of women's shoes. With smaller size of shoes of women corresponds to lower intelligence level and higher size of shoes of women corresponds to higher intelligence level of women. Comment on the conclusion that economic factors cause hemlines to rise and fall.



2. A researcher has a large number of data pairs (age, height) of humans beings from birth to 70 years. He computes a correlation coefficient. Would you expect it to be positive or negative? Why?

B) **If two variables change in the opposite direction** (i.e. if one increases, the other decreases and vice versa), then the correlation is called a **negative correlation.** For example: T.V. registrations and cinema attendance.

Some other examples of series of negative correlation are:

(i) Volume and pressure of perfect gas;

(ii) Current and resistance [keeping the voltage constant]

(iii) Price and demand for goods.

INTEXT QUESTIONS 10.2

1. What sort of correlation would be expected between a company's expenditure on health and safety and the number of work related accidents.
   (a) positive
   (b) negative
   (c) none
   (d) infinite

2. When "r" is negative, one variable increases in value, …..
   (a) the other increases
   (b) the other increases at a greater rate
   (c) the other variable decreases in value
   (d) there is no change in the other variable
   (e) all of the above

10.4 LINEAR AND NON-LINEAR CORRELATION

The nature of the graph gives us the idea of the linear type of correlation between two variables. If the graph is in a straight line, the correlation is called a **"linear correlation"** and if the graph is not in a straight line, the correlation is **non-linear** or **curvi-linear.**

For example, if variable x changes by a constant quantity, say 20 then y also changes by a constant quantity, say 4. The ratio between the two always remains the same (1/5 in this case). In case of a curvi-linear correlation this ratio does not remain constant.

Notes

In general two variables x and y are said to be linearly related, if there exists a relationship of the form

$$y = a + bx$$

where 'a' and 'b' are real numbers. This is nothing but a straight line when plotted on a graph sheet with different values of x and y and for constant values of a and b. Such relations generally occur in physical sciences but are rarely encountered in economic and social sciences.

The relationship between two variables is said to be non – linear if corresponding to a unit change in one variable, the other variable does not change at a constant rate but changes at a fluctuating rate. In such cases, if the data is plotted on a graph sheet we will not get a straight line curve. For example, one may have a relation of the form

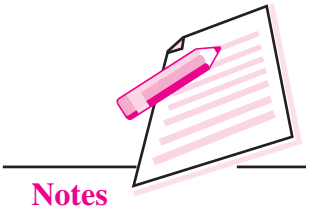
$$y = a + bx + cx^2$$

10.5 DEGREES OF CORRELATION

Through the coefficient of correlation, we can measure the degree or extent of the correlation between two variables. On the basis of the coefficient of correlation we can also determine whether the correlation is positive or negative and also its degree or extent.

1. **Perfect correlation:** If two variables change in the same direction and in the same proportion, the correlation between the two is perfect positive. According to Karl Pearson the coefficient of correlation in this case is +1. On the other hand, if the variables change in the opposite direction and in the same proportion, the correlation is perfect negative. Its coefficient of correlation is -1. In practice we rarely come across these types of correlations.
2. **Absence of correlation:** If two series of two variables exhibit no relations between them or change in one variable does not lead to a change in the other variable, then we can firmly say that there is no correlation or absurd correlation between the two variables. In such a case the coefficient of correlation is 0.
3. **Limited degrees of correlation:** If two variables are not perfectly correlated or there is a perfect absence of correlation, then we term the correlation as Limited correlation.

Thus Correlation may be positive, negative or zero but lies with the limits ± 1. i.e. the value of r is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.



- If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive correlation.
- If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative correlation
- If there is no linear correlation or a weak linear correlation, r is close to 0.

The following Table reveals the effect (or degree) of coefficient of correlation.

Table 10.2: Degree and Type of Correlation

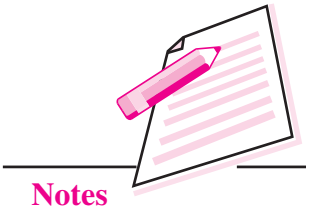
Degrees	Positive	Negative
Absence of correlation →	Zero	Zero
Perfect correlation →	+ 1	–1
High degree →	+ 0.75 to + 1	– 0.75 to –1
Moderate degree →	+ 0.25 to + 0.75	– 0.25 to – 0.75
Low degree →	0 to 0.25	0 to – 0.25

Note that r is a dimensionless quantity; that is, it does not depend on the units employed

INTEXT QUESTIONS 10.3

1. The coefficient of correlation ranges between
   (a) 0 and 1
   (b) –1 and +1
   (c) minus infinity and plus infinity
   (d) 1 and 100
2. If two variables are absolutely independent of each other the correlation between them must be,
   (a) –1
   (b) 0
   (c) +1
   (d) +0.1
3. The coefficient of correlation:
   (a) can be larger than 1
   (b) cannot be larger than 1
   (c) cannot be negative

4. If height is independent of average yearly income, what is the predicted correlation between these two variables?
   (a) 1
   (b) –1
   (c) 0
   (d) Impossible to say for sure
5. A student produces a correlation of +1.3. This is
   (a) a high positive correlation
   (b) a significant correlation
   (c) an impossible correlation
   (d) only possible if N is large
6. If A scored the top mark in the apprentices test on computing and the correlation between that test and the test on English language was +1.0 what position did A get in the test on English language.
   (a) middle
   (b) bottom
   (c) top
   (d) cannot say from the information given
7. Which correlation is the strongest +0.65 or –0.70
   (a) –0.70
   (b) +0.65
   (c) depends on N
   (d) cannot say from the information given
7. The symbol for the Karl Pearson Correlation Co-efficient is
   (a) Σ (b) σ (c) α (d) r
8. For a normal good, if price increases then the quantity demanded decreases. What type of correlation co-efficient would you expect in this situation?
   (a) 0
   (b) positive
   (c) 0.9
   (d) negative
   (e) unknowable



10.6 PROPERTIES OF CORRELATION COEFFICIENT

1. The correlation coefficient 'r' lies between -1 to +1.
2. The correlation coefficient 'r' is the pure number and is independent of the units of measurement of the variables.
3. The correlation coefficient 'r' is independent of change of origin i.e. the value of r is not affected even if each of the individual value of two variables is increased or decreased by some non-zero constant.
4. The correlation coefficient 'r' is independent of change of scale i.e. the value of r is not affected even if each of the individual value of two variables is multiplied or divided by some non-zero constant.

INTEXT QUESTIONS 10.4

1. Given a set of paired data (X, Y)
   (a) If Y is independent of X, then what value of a correlation coefficient would you expect?
   (b) If Y is linearly dependent on X, then what value of a correlation coefficient would you expect?
2. State whether the following statement is true or false: "If a positive correlation exists between height and weight, a person with above average height is expected to have above average weight".

10.7 METHODS OF DETERMINING CORRELATION

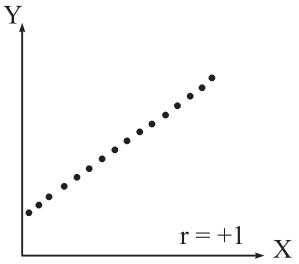
We shall consider the following most commonly used methods.

1. Scatter Plot
2. Karl Pearson's coefficient of correlation
3. Spearman's Rank-correlation coefficient.

10.7.1 Scatter Plot (Scatter diagram or dot diagram)

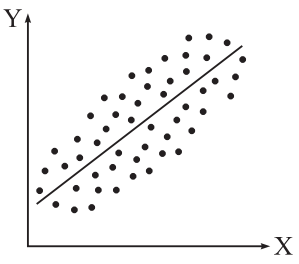
Scatter Plots (also called scatter diagrams) are used to graphically investigate the possible relationship between two variables without calculating any numerical value. In this method, the values of the two variables are plotted on a graph paper. One is taken along the horizontal (X-axis) and the other along the vertical (Y-axis). By plotting the data, we get points (dots) on the graph which are generally scattered and hence the name 'Scatter Plot'.

The manner in which these points are scattered, suggest the degree and the direction of correlation. The degree of correlation is denoted by 'r' and its direction is given by the signs positive and negative.



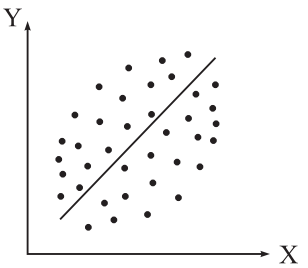
Perfect Positive Correlation

(a)



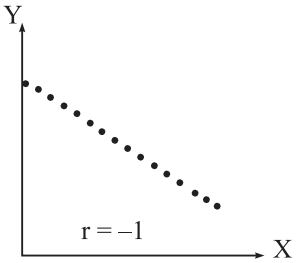
High Degree of Positive Correlation

(b)



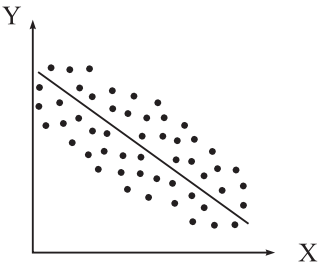
Low Degree of Positive Correlation

(c)



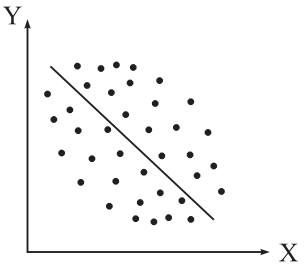
Perfect Negative Correlation

(d)



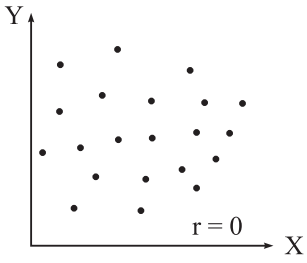
High Degree of Negative Correlation

(e)



Low Degree of Negative Correlation

(f)



No Correlation

(g)

(i) If all points lie on a rising straight line, the correlation is perfectly positive and r = +1 (see fig. a)

(ii) If all points lie on a falling straight line the correlation is perfectly negative and r = -1 (see fig. d)

(iii) If the points lie in narrow strip, rising upwards, the correlation is high degree of positive (see fig. b)

(iv) If the points lie in a narrow strip, falling downwards, the correlation is high degree of negative (see fig. e)

(v) If the points are spread widely over a broad strip, rising upwards, the correlation is low degree positive (see fig. c)

(vi) If the points are spread widely over a broad strip, falling downward, the correlation is low degree negative (see fig. f)

(vii) If the points are spread (scattered) without any specific pattern, the correlation is absent. i.e. r = 0. (see fig. g)

Though this method is simple and is a rough idea about the existence and the degree of correlation, it is not reliable. As it is not a mathematical method, it cannot measure the degree of correlation.

10.7.2 Karl Pearson's coefficient of correlation

It gives the precise numerical expression for the measure of correlation. It is denoted by 'r'. The value of 'r' gives the magnitude of correlation and its sign denotes its direction. The mathematical formula for computing r is:

$$r = \frac{\sum xy}{N\sigma_X\sigma_y} \qquad ...(1)$$

where $x = (X - \bar{X})$, $y = (Y - \bar{Y})$, σ_X = s.d. of X

σ_y = s.d. of Y

and N = number of paris of observations

Since $\sigma_X = \sqrt{\frac{\sum x^2}{N}}$ and $\sigma_y = \sqrt{\frac{\sum y^2}{N}}$

So equation 1 can be rewritten as:

$$r = \frac{\sum xy}{\sqrt{\sum x^2} \times \sqrt{\sum y^2}}$$

By using actual mean

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

By assumed mean method

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx \cdot \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\times\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}} \quad ...(3)$$

By direct method

$$r = \frac{N\Sigma XY - [\Sigma X][\Sigma Y]}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\times\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} \quad ...(4)$$

Now covariance of X and Y is defined as

$$cov(X,Y) = \frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{N}$$

$$\therefore \quad r = \frac{cov(X,Y)}{\sigma_X\sigma_Y}$$

Where N is the number of pairs of data.

$$d_x = X - A_X$$
$$d_y = Y - A_Y$$

INTEXT QUESTIONS 10.5

1. Positive values of covariance indicate
   (a) a positive variance of the X values
   (b) a positive variance of the Y values
   (c) the standard deviation is positive
   (d) positive relation between two variables

Example 1: Calculate the coefficient of correlation between the expenditure on advertising and sales of the company from the following data.

Advertising Expenditure (in 000 ₹):	165	166	167	168	167	169	170	172
Sales (in Lakh ₹)	167	168	165	172	168	172	169	171

Solution: N = 8 (pairs of observations)

Table 10.3: Calculation of coefficient of correlation

Advertising Expenditure (in 000 ₹) : X_i	Sales (in Lakh ₹) Y_i	$x = X_i - \bar{X}$	$y = Y_i - \bar{Y}$	xy	x^2	y^2
165	167	–3	–2	6	9	4
166	168	–2	–1	2	4	1
167	165	–1	–4	4	1	16
167	168	–1	–1	1	1	1
168	172	0	3	0	0	9
169	172	1	3	3	1	9
170	169	2	0	0	4	0
172	171	4	2	8	16	4
$\Sigma X_i = 1344$	$\Sigma Y_i = 1352$	0	0	$\Sigma xy = 24$	$\Sigma x^2 = 36$	$\Sigma y^2 = 44$

Calculation:

$$\bar{X} = \frac{\Sigma X_i}{N} = \frac{1344}{8}$$

$$= 168 \text{ cm and } \sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{36}{8}}$$

$$\bar{Y} = \frac{\Sigma Y_i}{N} = \frac{1352}{8}$$

$$= 169 \text{ cm and } \sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{44}{8}}$$

Now, $$r = \frac{\Sigma xy}{N\sigma_x\sigma_y} = \frac{24}{8\sqrt{\frac{36}{8}} \times \sqrt{\frac{44}{8}}} = \frac{24}{\sqrt{36 \times 44}} = +0.6029$$

Since r is positive and 0.6. This shows that the correlation is positive and moderate (i.e. direct and reasonably good).

Example 2: From the following data compute the coefficient of correlation between X and Y.

		X	Y
No. of items	→	15	15
Arithmetic mean	→	25	18
$\Sigma(X_i - \bar{X})^2$ and $\Sigma(Y_i - \bar{Y})^2$	→	136	138
$\Sigma(X_i - \bar{X}) \cdot \Sigma(Y_i - \bar{Y})$	→	122	

Solution: Given, N = 15, $\bar{X} = 25$. $\bar{Y} = 18$

$$\Sigma(X_i - \bar{X})^2 \quad \text{i.e.} \quad \Sigma x^2 = 136$$

$$\Sigma(Y_i - \bar{Y})^2 \quad \text{i.e.} \quad \Sigma y^2 = 138$$

and $$\Sigma(X_i - \bar{X}) \cdot \Sigma(Y_i - \bar{Y}) = \Sigma xy = 122$$

Using $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \times \sqrt{\Sigma y^2}}$$

we get $$r = \frac{122}{\sqrt{136} \times \sqrt{138}} = \frac{122}{136.9} = 0.891$$

Example 3: If covariance between X and Y is 12.3 and the variance of x and y are 16.4 and 13.8 respectively. Find the coefficient of correlation between them.

Solution: Given: Covariance = cov (X, Y) = 12.3

Variance of X (σ_x^2) = 16.4

Variance of Y (σ_y^2) = 13.8

Now,

$$r = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{12.3}{\sqrt{16.4} \times \sqrt{13.8}}$$

$$= \frac{12.3}{4.05 \times 3.71} = 0.82$$

Example 4: Find the number of pair of observations from the following data.

$r = 0.25, \Sigma(X_i - X)(Y_i - Y) = 60, \sigma_y = 4, \Sigma(X_i - \bar{X})2 = 90.$

Solution: Given: $r = 0.25$

$$\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \Sigma xy = 60$$

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{N}} = \sqrt{\frac{90}{N}}$$

$$\sigma_y = 4 = \sqrt{\frac{\Sigma y^2}{N}}$$

Now, $$r = \frac{\Sigma xy}{n\sigma_x \cdot \sigma_y} = \frac{60}{N\sqrt{\frac{90}{N}} \times 4} = \frac{15}{\sqrt{90N}}$$

$\therefore$ $$0.25 = \frac{15}{\sqrt{90N}}$$

$\therefore$ $$0.25 \times \sqrt{90N} = 15$$

on squaring

$\therefore$ $0.0625 \times 90N = 225$

$\therefore$ $$90N = \frac{225}{0.0625}$$

$\therefore$ $90N = 3600$

$\therefore$ $N = 40$

Therefore, the number of pairs of observations = 40

10.7.2.1 Assumed Mean Method (Step Deviation)

If the values of X and Y are very big, the calculation becomes very tedious and if we change the variable X to $u = \frac{X_1 - A}{h}$ and Y to $v = \frac{Y_1 - B}{k}$ where A and B are the assumed means for variable X and Y respectively and h and k are common

factor of variable X and Y, As stated earlier the one of the property of correlation coefficient is that it is independent of change of origin and change of scale so

$$r_{xy} = r_{uv}$$

The formula for r can be simplified as

$$r_{xy} = r_{uv} = \frac{\Sigma uv - \left(\frac{(\Sigma u)(\Sigma v)}{N}\right)}{\sqrt{\Sigma u^2 - \frac{(\Sigma u)^2}{N}} \times \sqrt{\Sigma v^2 - \frac{(\Sigma v)^2}{N}}}$$

Example 5: The following data relates to the Cost and Sales of a Company for the past 10 months

Cost (in 000 ₹):	44	80	76	48	52	72	68	56	60	64
Sales(in 000 ₹):	48	75	54	60	63	69	72	51	57	66

Find the coefficient of correlation between the two.

Solution: Here A = 60, h = 4, B = 60 and k = 3

Table 10.4: Correlation coefficient between cost and sales

Cost (in 000 ₹)	**Sales (in 000 ₹)**	**$u = \frac{X_1 - A}{h}$**	**$v = \frac{Y_1 - B}{d}$**	**uv**	**u^2**	**v^2**
44	48	–4	–4	16	16	16
80	75	5	5	25	25	25
76	54	4	–2	–8	16	4
48	60	–3	0	0	9	0
52	63	–2	1	–2	4	1
72	69	3	3	9	9	9
68	72	2	4	8	4	16
56	51	–1	–3	3	1	9
60	57	0	–1	0	0	1
64	66	1	2	2	4	4
		$\Sigma u = 5$	$\Sigma v = 5$	$\Sigma uv = 53$	$\Sigma u^2 = 85$	$\Sigma v^2 = 85$

Calculation:

$$r_{xy} = r_{uv} = \frac{\Sigma uv - \left(\frac{(\Sigma u)(\Sigma v)}{N}\right)}{\sqrt{\Sigma u^2 - \frac{(\Sigma u)^2}{N}} \times \sqrt{\Sigma v^2 - \frac{(\Sigma v)^2}{N}}}$$

$$= \frac{53 - \left(\frac{(5)(5)}{10}\right)}{\sqrt{85 - \frac{(5)^2}{10}} \times \sqrt{85 - \frac{(5)^2}{10}}}$$

$$= \frac{53 - \left(\frac{(5)(5)}{10}\right)}{\sqrt{85 - \frac{(5)^2}{10}} \times \sqrt{85 - \frac{(5)^2}{10}}}$$

$$= \frac{53 - 2.5}{\sqrt{82.5} \times \sqrt{82.5}}$$

$$= \frac{50.5}{82.5} = 0.61$$

10.6.3 Spearman's Rank Correlation Coefficient

This method is based on the ranks of the items rather than on their actual values. The advantage of this method over the others in that it can be used even when the actual values of items are unknown. For example if you want to know the correlation between honesty and wisdom of the boys of your class, you can use this method by giving ranks to the boys. It can also be used to find the degree of agreements between the judgments of two examiners or two judges. The formula is:

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

where R = Rank correlation coefficient

D = Difference between the ranks of two items

N = the number of observations.

Note: $-1 \le R \le 1$.

(i) When $R = +1 \Rightarrow$ Perfect positive correlation or complete agreement in the same direction

(ii) When $R = -1 \Rightarrow$ Perfect negative correlation or complete agreement in the opposite direction.

(iii) When $R = 0 \Rightarrow$ No Correlation.

Computation:

(i) Give ranks to the values of items. Generally the item with the highest value is ranked 1 and then the others are given ranks 2, 3, 4 ... according to their values in the decreasing order.

(ii) Find the difference $D = R_1 - R_2$

where R_1 = Rank of X and R_2 = Rank of Y

Note that $\Sigma D = 0$ (always)

(iii) Calculate D^2 and then find ΣD^2

(iv) Apply the formula.

Note :

In some cases, there is a tie between two or more items. For example if each item have rank say 4th then they are given $\frac{4+5}{2} = 4.5$th rank. If three items are of equal rank say 4th then they are given $\frac{4+5+6}{3} = 5$th rank each. If m be the number of items of equal ranks, the factor $\frac{1}{12}(m^3 - m)$ is added to SD^2. If there is more than one of such cases then this factor added as many times as the number of such cases, then

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

Example 6 : Calculate ' R ' from the following data.

Student No. :	1	2	3	4	5	6	7	8	9	10
Rank in Maths :	1	3	7	5	4	6	2	10	9	8
Rank in Stats :	3	1	4	5	6	9	7	8	10	2

Solution:

Table 10.5: Calculation of rank correlation

Student No.	Rank in Maths (R_1)	Rank in Stats (R_2)	$D = (R_1 - R_2)$	D^2
1	1	3	–2	4
2	3	1	2	4
3	7	4	3	9
4	5	5	0	0
5	4	6	–2	4
6	6	9	–3	9
7	2	7	–5	25
8	10	8	2	4
9	9	10	–1	1
10	8	2	6	36
N = 10			$\Sigma D = 0$	$\Sigma D^2 = 96$

Calculation of R :

$$R = 1 - \frac{6\Sigma D^2}{N\left(N^2 - 1\right)} = 1 - \frac{6(96)}{10(100-1)} = 1 - \frac{6\times 96}{10\times 99} = 0.4181$$

Example 7: Calculate 'R' of 6 students from the following data.

Marks in Stats :	40	42	45	35	36	39
Marks in English :	46	43	44	39	40	43

Solution:

Table 10.6: Calculation of rank correlation

Marks in Stats	R_1	Marks in English	R_2	D	D^2
40	3	46	1	2	4
42	2	43	3.5	–1.5	2.25
45	1	44	2	–1	1
35	6	39	6	0	0
36	5	40	5	0	0
39	4	43	3.5	0.5	0.25
N = 6				$\Sigma D = 0$	$\Sigma D^2 = 750$

Here m = 2 since in series of marks in English of items of values 43 repeated twice.

$$R = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}\left(2^3 - 2\right)\right\}}{N\left(N^2 - 1\right)} = 1 - \frac{6\left\{7.5 + \frac{1}{12}(8-2)\right\}}{6(36-6)}$$

$$R = 1 - \frac{6(7.5+0.5)}{210} = 0.771$$

Example 8: The value of Spearman's rank correlation coefficient for a certain number of pairs of observations was found to be 2/3. The sum of the squares of difference between the corresponding rnarks was 55. Find the number of pairs.

Solution: We have

$$1 - \frac{6\Sigma D^2}{N\left(N^2 - 1\right)} \text{ but } R = \frac{2}{3} \text{ and } \Sigma D^2 = 55$$

$$\therefore \quad \frac{2}{3} = 1 - \frac{6\times 55}{N\left(N^2 - 1\right)}$$

$$\therefore \quad -\frac{1}{3} = -\frac{6\times 55}{N\left(N^2 - 1\right)}$$

$$\therefore \quad N(N^2 - 1)\ 6 \times 55$$

Now $N(N^2 - 1) = 990$

$\therefore$ $N(N^2 - 1) = 10 \times 99 = 10(100 - 1)$

$\therefore$ $N(N^2 - 1) = 10(102 - 1) \Rightarrow N = 10$

Therefore, there were 10 students.

INTEXT QUESTIONS 10.6

1. The marks awarded by two judges in a certain beauty contest are given below:

Judge I	56	75	45	71	61	64	58	80	76	61
Judge II	66	70	40	60	65	56	59	77	67	63

By Using Rank correlation method, Determine whether the two judges have common taste in the judgement of beauty?

WHAT YOU HAVE LEARNT

- Correlation measures the associations between variables. Correlation can be positive or negative and linear or non-linear. It is denoted by r.
- The value of r lies between -1 and +1 i.e. $-1 \le r \le +1$.
- The correlation coefficient 'r' is independent of change of origin and change of scale.
- The important methods of measuring correlation are (i) Scatter Plot (ii) Karl Pearson's coefficient of correlation; and (iii) Spearman's Rank-correlation coefficient;
- Scatter Plots are used to graphically investigate the possible relationship between two variables without calculating of any numerical value.
- The mathematical formula for computing *r* using Karl Pearson method is given:

$$r = \frac{\sum xy}{N\sigma_X\sigma_y} \qquad ...(1)$$

where $x = (X - \bar{X})$, $y = (Y - \bar{Y})$, σ_X = s.d. of X

σ_X = s.d. of Y and N = number of paris of observation

- Correlation (r) can also be calculated using actual figure of two variables X and Y as follows:

$$r = \frac{N\sum XY - [\sum X][\sum Y]}{\sqrt{N\sum X^2 - (\sum X)^2} \times \sqrt{N\sum Y^2 - (\sum Y)^2}}$$

- The covariance 'of two variables say X and Y is defined as:

$$\text{cov}(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N}$$

where N is the number of pairs of data.

If covariance is given, then $r = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$

- The correlation (r) using Assumed Mean Method is given by:

$$r_{xy} = r_{uv} = \frac{\Sigma uv - \left(\frac{(\Sigma u)(\Sigma v)}{N}\right)}{\sqrt{\Sigma u^2 - \frac{(\Sigma u)^2}{N}} \times \sqrt{\Sigma v^2 - \frac{(\Sigma v)^2}{N}}}$$

where $u = \frac{X - A}{h}$ and $v = \frac{Y - B}{k}$

A and B are the assumed means for variable X and Y respectively and h and k are common factor of variable X and Y.

- The Spearman rank correlation (R) is given by:

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

where R = Rank correlation coefficient

D = Difference between the ranks of two items

N = the number of observations.

TERMINAL EXERCISES

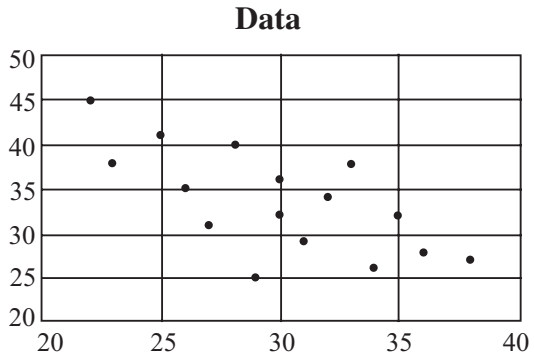
1. The data relating to variable X and Y is given below:

X	72	73	75	76	77	78	79	80	80	81	82	83	84	85	86	88
Y	45	38	41	35	31	40	25	32	36	29	34	38	26	32	28	27

(a) Sketch a scatter plot.

(b) Compute the correlation coefficient, r.

Answer 1: a.



2. Calculate and analyze the correlation coefficient between the number of study hours and the number of sleeping hours of different students.

Number of Study hours	2	4	6	8	10
Number of sleeping hours	10	9	8	7	6

3. Find the value of the correlation coefficient from the following table:

Subject	Age X	Glucose Level Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

4. The values of the same 15 students in two subjects A and B are given below; the two numbers within the brackets denoting the ranks of the same student in A and B respectively.

(1,10) (2,7) (3,2) (4,6) (5,4) (6,8) (7,3) (8,1).

(9,11) (10,15) (11,9) (12,5) (13,14) (14,12) (15,13)

Use Spearman's formula to find the rank Correlation Coefficient.

5. Calculate Karl Pearson's coefficient of correlation from the advertisement cost and sales as per the data given below:

Advertisement Cost (in'000 $)	39	65	62	90	82	75	25	98	36	78
Sales (in lakh $)	47	53	58	86	62	68	60	91	51	84

6. The following observations are given for two variables.

Y	X
5	2
8	12
18	3
20	6
22	11
30	19
10	18
7	9

(a) Compute and interpret the sample covariance for the above data.

(b) Compute and interpret the sample correlation coefficient.

7. A trainee manager wondered whether the length of time his trainees revised for an examination had any effect on the marks they scored in the examination. Before the exam, he asked a random sample of them to honestly estimate how long, to the nearest hour, they had spent revising. After the examination he investigated the relationship between the two variables.

Trainee	A	B	C	D	E	F	G	H	I	J
Revision time	4	9	10	14	4	7	12	22	1	17
Exam mark	31	58	65	73	37	44	60	91	21	84

(a) Plot the scatter diagram in order to inspect the data.

(b) Calculate the correlation coefficient.

8. Positive values of covariance indicate

(a) a positive variance of the x values

(b) a positive variance of the y values

(c) the standard deviation is positive

(d) positive relation between two variables

9. A numerical measure of linear association between two variables is the

(a) variance

(b) coefficient of variation

(c) correlation coefficient

(d) standard deviation

10. The coefficient of correlation ranges between
    (a) 0 and 1
    (b) -1 and +1
    (c) minus infinity and plus infinity
    (d) 1 and 100
11. The coefficient of correlation:
    (a) can be larger than 1
    (b) cannot be larger than 1
    (c) cannot be negative
12. If height is independent of average yearly income, what is the predicted correlation between these two variables?
    (a) 1
    (b) –1
    (c) 0
    (d) Impossible to say for sure
13. A student produces a correlation of +1.3. This is
    (a) a high positive correlation
    (b) a significant correlation
    (c) an impossible correlation
    (d) only possible if N is large
14. What sort of correlation would be expected between a company's expenditure on health and safety and the number of work related accidents..
    (a) positive
    (b) negative
    (c) none
15. If A scored the top mark in the apprentices test on computing and the correlation between that test and the test on English language was +1.0 what position did A get in the test on English language.
    (a) middle
    (b) bottom
    (c) top
    (d) cannot say from the information given

Notes

16. Which correlation is the strongest +0.65 or -0.70

    (a) –0.70

    (b) +0.65

    (c) depends on N

    (d) cannot say from the information given

17. The symbol for the Karl Pearson Correlation Co-efficient is –

    (a) Σ

    (b) σ

    (c) α

    (d) r

18. When "r" is negative, one variable increases in value, …..

    (a) the other increases

    (b) the other increases at a greater rate

    (c) the other variable decreases in value

    (d) there is no change in the other variable

    (e) all of the above

19. If two variables are absolutely independent of each other the correlation between them must be,

    (a) –1

    (b) 0

    (c) +1

    (d) +0.1

20. For a normal good, if price increases then the quantity demanded decreases. What type of correlation co-efficient would you expect in this situation?

    (a) 0

    (b) positive

    (c) 0.9

    (d) negative

    (e) unknowable

ANSWERS TO INTEXT QUESTIONS

10.1

1. A positive correlation does exst; however correlation does not imply causation
2. Positive, since in general, people grow in height increasing with age

10.2

1. (b)
2. (c)

10.3

1. (b) 2. (b) 3. (b) 4. (c) 5. (c)
6. (c) 7. (a) 8. (d) 9. (d)

10.4

1. (a) $r = 0$
2. (b) $r = 1$ or $r = -1$ these two are same as $|r| = -1$

10.5

1. (d)

10.6

1. +0.67 this indicates a strong positive relationship between the ranks given by two judges i.e. the judges have high degree of common approach towards judgement of beauty.